Few-shot Learning with Multilingual Generative Language Models

Xi Victoria Lin[*], **Todor Mihaylov**, **Mikel Artetxe**, **Tianlu Wang**, **Shuohui Chen**, **Daniel Simig**, **Myle Ott**, **Naman Goyal**, **Shruti Bhosale**, **Jingfei Du**, **Ramakanth Pasunuru**, **Sam Shleifer**, **Punit Singh Koura**, **Vishrav Chaudhary**, **Brian O'Horo**, **Jeff Wang**, **Luke Zettlemoyer**, **Zornitsa Kozareva**, **Mona Diab**, **Veselin Stoyanov**, **Xian Li**[*]
Meta AI

Abstract

Large-scale generative language models such as GPT-3 are competitive few-shot learners. While these models are known to be able to jointly represent multiple languages, their training data is dominated by English, potentially limiting their cross-lingual generalization. In this work, we train multilingual generative language models on a corpus covering a diverse set of languages, and study their few- and zero-shot learning capabilities in a wide range of tasks. Our largest model with 7.5 billion parameters sets new state of the art in few-shot learning in more than 20 representative languages, outperforming GPT-3 of comparable size in multilingual commonsense reasoning (with +7.4% absolute accuracy improvement in 0-shot settings and +9.4% in 4-shot settings) and natural language inference (+5.4% in each of 0-shot and 4-shot settings). On the FLORES-101 machine translation benchmark, our model outperforms GPT-3 counterparts on 171 out of 182 directions with 32 training examples, while surpassing the official supervised baseline in 45 directions. We conduct an in-depth analysis of different multilingual prompting approaches, showing in particular that strong in-context few-shot learning performance across languages can be achieved via cross-lingual transfer through both templates and demonstration examples.[1]

1 Introduction

Large autoregressive language models such as GPT-3 can be adapted, via few- and zero-shot learning, to a wide range of tasks with significantly less cost than full fine-tuning (Brown et al., 2020; Bommasani et al., 2021). These models have been primarily developed for English. Although the training data of GPT-3 contains a small percentage of non-English text (7%) allowing it to achieve some promising cross-lingual generalization, the model is almost exclusively deployed for use cases in English. Multilingual masked and sequence-to-sequence language models have been studied, including mBERT, XLM-R, mT5, and mBART (Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2020; Fedus et al., 2021; Goyal et al., 2021a; Liu et al., 2020). These models are typically fine-tuned on large amount of labeled data in downstream tasks. Despite notable recent work at smaller scales (Zhao and Schütze, 2021) and for domain-specific tasks (Winata et al., 2021), the multilingual few-shot learning capabilities of language models are less well understood.

In this paper, we train four multilingual generative language models (up to 7.5 billion parameters), XGLM's, and present a comprehensive study of multilingual zero- and in-context few-shot learning. We train the models using a large-scale corpus of 500B tokens that comprises 30 diverse languages, up-sampling the less-resourced languages to render a more balanced language representation. We evaluate the models on multiple multilingual natural language understanding (NLU) tasks, machine translation and a subset of English tasks demonstrated in Brown et al. (2020).

We found XGLM demonstrate strong cross-lingual capability where using English prompts together with non-English examples yields competitive zero- and few-shot learning performance. Our largest model ($\text{XGLM}_{7.5\text{B}}$) achieves strong zero- and few-shot learning performance on language completion and inference tasks (e.g. XStoryCloze: 65.4% 0-shot, 66.5% 4-shot; XNLI: 46.3% 0-shot, 47.3% 4-shot). It also establishes a new state-of-the-art on few-shot machine translation across a large number of language pairs in the FLORES-101 benchmark (Goyal et al., 2021b), significantly outperforming the GPT-3 model of comparable size

[*] Equal contribution. Correspondence to: ⟨victorialin@meta.com, xianl@meta.com⟩.

[1] Our checkpoints, code and new dataset (XStoryCloze): https://github.com/facebookresearch/fairseq/tree/main/examples/xglm.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9019–9052
December 7-11, 2022 ©2022 Association for Computational Linguistics

(6.7 billion parameters). On the other hand, multilingual pre-training causes performance drop on English. On 8 English NLU tasks, XGLM $_{7.5B}$ underperforms GPT-$3_{6.7B}$ by 10.9% on average in zero-shot learning. GPT-$3_{6.7B}$ also surpasses XGLM $_{7.5B}$ in machine translation among several high-resource language pairs, including WMT-14 *en↔fr*, WMT-16 *en↔de* and WMT-19 *en↔zh*.

We conduct an in-depth analysis of different multilingual prompting approaches by examining cross-lingual transfer through template and demonstration examples respectively. We show the non-English prompts sometimes yield unexpected low zero- and few-shot learning accuracy even if they are crafted by native speakers (§4.3). Both using the English template (§4.4) and adding demonstration examples (§4.5) provide effective remedy. However, using demonstration examples from another language often cannot further improve the zero-shot learning performance when a strong prompting language (e.g. Engilsh) is used, which indicates room for improvement in cross-lingual pre-training and in-context transfer approaches.

2 Models and Pre-training Data

2.1 Pre-training Data

Language selection and pre-processing. We extend the pipeline used for mining the CC100 corpus (Conneau et al., 2020; Wenzek et al., 2020) to generate CC100-XL, a significantly larger multilingual dataset covering 68 Common Crawl (CC) snapshots (from Summer 2013 to March/April 2020) and 134 languages. Our pretraining data include 30 languages covering 16 language families. The natural data distribution is skewed with the number of English tokens being 6 times that of the second largest language. Following previous work on multilingual pre-training (Conneau et al., 2020; Liu et al., 2020), we up-sampled the medium and low resource languages to create a more balanced language distribution (Appendix F.1).[2] Figure 1 shows the language distribution of our pre-training data before (blue) and after (green) up-sampling.

Joint sub-word vocabulary. We process all languages with a joint vocabulary of size 250k created through unigram language modeling (Kudo, 2018), using the SentencePiece library (Kudo and Richardson, 2018). We train the unigram-LM model using 10 million sentences randomly sampled from the filtered data, according to the multinomial distribution defined in Lample and Conneau (2019) with $\alpha = 0.3$.

2.2 Models and Training

We train decoder-only causal language models with the Transformer architecture similar to GPT-3 (Brown et al., 2020). This allows us to study the effect of scaling up model size along both width and depth dimensions. As a result, we compare four models with 564M, 1.7B, 2.9B and 7.5B parameters, respectively. The architecture details are summarized in Table 1. Our models match that of GPT-3 models[3] except with the additional embedding parameters from a larger vocabulary. All models are trained for up to 500B tokens, with context length of 2048 tokens. Further training details are described in Appendix A.

GPT-3			XGLM		
size	l	h	*size*	l	h
125M	12	768		—	
355M	24	1024	564M	24	1024
760M	24	1536		—	
1.3B	24	2048	1.7B	24	2048
2.7B	32	2560	2.9B	48	2048
6.7B	32	4096	7.5B	32	4096

Table 1: **Model details**. *size*: number of parameters, l: layers, h: hidden dimension. Models within the same row have comparable sizes.

3 Multilingual In-context Learning

We measure the performance of our multilingual language models on downstream tasks in different languages given the tasks and few-shot demonstrations specified via prompts without further parameter updates (Appendix B).

3.1 Multilingual and Cross-lingual Prompting

Previous work on English in-context learning has shown that performance heavily depends on

[2] We inadvertently over-sampled some of the less resourced languages which is reflected in the statistics of *ko*, *fi*, *th*, *bg*, *ca*, *hi*, *et* languages, as shown in Figure 1. We did not ablate the effect of this mistake due to the extreme computational cost. Studying optimal language balancing is an important area for future work.

[3] For XGLM 2.9B we used the optimal depth-to-width parameter allocation for GPT-3 architectures based on rank bottleneck analysis (Levine et al., 2020). This allocation is expected to have improved training efficiency. However, it did not converge for XGLM 7.5B in our experiments, and we fell back to the original GPT-3 setup.

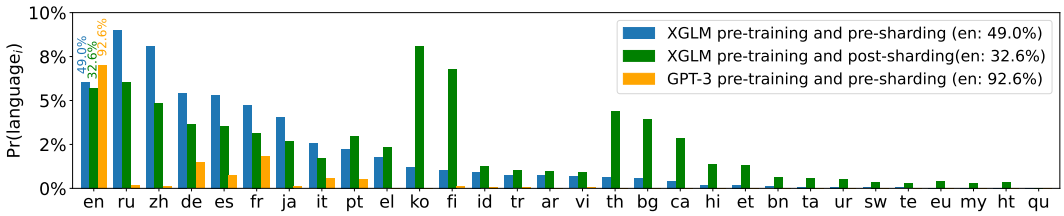


Figure 1: The % of each language l ($l = 1, 2, ..., 30$) in XGLM's pre-training data pre-upsampling (blue), post-upsampling (green), and its corresponding % in GPT-3's training data (orange). We truncate the y-axis at 10% to better visualize the tail distribution.

Task Category	Dataset	Template	Candidate Verbalizer
Reasoning	XCOPA XStoryCloze XWinograd	*cause*: `{Sentence 1}` because `[Mask]` *effect*: `{Sentence 1}` so `[Mask]` `{Context} [Mask]` `{Context}` (*with '_' replaced by* `[Mask]`)	Identity
NLI	XNLI	`{Sentence 1}`, right? `[Mask]`, `{Sentence 2}`	*Entailment*: Yes | *Neural*: Also | *Contradiction*: No
Paraphrase	PAWS-X	`{Sentence 1}`, right? `[Mask]`, `{Sentence 2}`	*True*: Yes | *False*: No
Translation	WMT, FLORES-101	`{Source sentence}` = `[Mask]`	Identity

Table 2: Handcrafted (English) prompts for multilingual natural language understanding and translation tasks.

the prompt construction, and it is challenging to find the optimal prompt for a given language model (Gao et al., 2021; Perez et al., 2021). This problem is further complicated in the multilingual setting, where we need to find the optimal prompts for examples in different languages.

In this work, we consider three approaches for obtaining the prompts for non-English tasks.

Handcrafting prompts. The first approach is to ask native speakers of the target language to handcraft the prompts. Prompts created this way are expected to have the most natural surface form. However, language expertise is expensive and we further consider two alternatives.

Translating from English prompts. We assume high-quality prompts of a task can be easily sourced in English(Sanh et al., 2021; Mishra et al., 2021). Non-verbal prompts do not contain words in any particular language (e.g. the StoryCloze and WMT prompts shown in Table 2), while verbal prompts have different realizations in different languages (Table 3). If the prompt is non-verbal, we simply apply it to the other languages. If the prompt is verbal, we translate it into the other languages using automatic translation APIs.

Cross-lingual prompting. We consider the third approach which directly applies the prompts in English (or another high-resource language) to non-English examples. We expect this approach to be competitive, as a result of the cross-lingual capability of the model after being trained on a diverse set of languages.

3.2 Learning from Cross-lingual Demonstrations

The cross-lingual nature of multilingual language models further enable the possibility of learning from a different language in context without parameter updates. To do so we simply append examples from another language as the demonstration examples in the language model context. Such capability enables cheap transfer from high-resource languages to the low-resource target languages.

4 Experiments and Results

4.1 Tasks

We evaluate the zero-shot and in-context few-shot learning capabilities (Brown et al., 2020) of XGLM on a spectrum of downstream tasks (Table 4).

Multilingual tasks. We select four multilingual tasks spanning commonsense reasoning (XCOPA), anaphora resolution (XWinograd), natural language inference (XNLI) and paraphrasing (PAWS-X). We also created a new dataset, XStoryCloze, by professionally translating the validation split[4] of

[4]We further split the translated data into train and test (20% vs. 80%, respectively) for each language, keeping the parallel sentence mapping in both splits.

Task	Lang	Template	Candidate Verbalizer		
			Entailment	Contradiction	Neutral
XNLI	en	{Sentence 1}, right? [Mask], {Sentence 2}	Yes	No	Also
	zh	{Sentence 1}[Mask], {Sentence 2}	由此可知，	所以，不可能	同时，
	es	{Sentence 1}, ¿verdad? [Mask], {Sentence 2}	Sí	No	Además
XCOPA	en	*cause*: {Sentence 1} because [Mask] | *effect*: {Sentence 1} so [Mask]	Identity		
	zh	*cause*: 因为[Mask]，所以{Sentence 1} | *effect*: 因为{Sentence 1}，所以[Mask]			

Table 3: Handcrafted multilingual prompts. English (*en*), Chinese (*zh*) and Spanish (*es*) for XNLI; English (*en*) and Chinese (*zh*) for XCOPA.

Task Category	Task	|Train|	|Dev|	|Test|	Non-En Sets	|Lang.|
Reasoning	XStoryCloze ♠	361	–	1,511	translations	11
	XCOPA ♠ (Ponti et al., 2020a)	33,410+400	100	500	translations	11
	XWinograd (Tikhonov and Ryabinin, 2021)	–	–	2,325†	translations	6
NLI	XNLI ♠ (Conneau et al., 2018)	–	2,490	5,010	translations	15
Paraphrase	PAWS-X (Yang et al., 2019)	–	2,000	2,000	translations	7

Table 4: Multilingual tasks used in our few-shot learning evaluation. All tasks use accuracy as the evaluation metrics. †: In XWinograd, each language has different number of test examples: *en*: 2,325, *jp*: 959, *ru*: 315, *pt*: 263. ‡: We use the COPA release in SuperGLUE (Wang et al., 2019). ♠: Held-out tasks.

the English StoryCloze dataset (Spring 2016 version) to 10 other typologically diverse languages (*ru*, *zh* Simplified, *es* Latin American, *ar*, *hi*, *id*, *te*, *sw*, *eu*, *my*)[5]. In addition, we evaluate our models on machine translation (§4.8) and multilingual social value tasks (Appendix E.1).

English tasks. We also evaluate our models on English commonsense reasoning and QA, a subset of benchmark tasks used by Brown et al. (2020), and compare the performance to state-of-the-art English-centric few-shot learning models. The tasks are detailed in Table A1.

4.2 Setup

Scoring function and calibration. We follow the guidelines suggested by Perez et al. (2021) and adopt a *cross-task generalization* setting (Triantafillou et al., 2020) to select our scoring function. We reserve three held-out tasks (XNLI, XCOPA and XStoryCloze) to perform the selection based on their development set performance, and directly apply the selected settings to the rest of the tasks. In the end, we use the averaged per-token log-probabilities ignoring the common prefix of different candidates as the scoring function for all multilingual tasks with no additional calibration or normalization. Appendix C.2 details the selection.

Few-shot learning evaluation. We focus on benchmarking the 0- and 4-shot learning performance of the models on all tasks. For cross-lingual demonstration (§4.5), scaling law (§4.9) and translation (§4.8) we also reported 1-shot and 32-shot performance. We report the average results across 5 runs, randomly sampling a different set of few-shot examples each time. Without further specification, we use few-shot examples in the same language as the target example. Appendix C.3 details our complete evaluation protocol.

4.3 Comparing Prompting Approaches

We first compare different multilingual prompting approaches proposed in §3.1 using $\text{XGLM}_{7.5\text{B}}$ on XNLI and XCOPA[6]. Native speakers among the authors handcrafted[7] the prompts for the following tasks: XNLI (*en*, *zh*, *es* and *hi*) and XCOPA (*en*, *zh*), as shown in Table 3. We compare the performance of these human-written prompts to English prompts, machine-translated (MT) prompts and human-translated (HT) prompts.

Table 5 and 6 show the performance of different

[5]For all of our multilingual NLU datasets, the non-English sections of the data are (professionally) translated from the English section. Despite being the dominant approach adopted by the community (Ruder et al., 2021), it was previously shown to introduce data artifacts that inflate the measured cross-lingual transfer of models (Artetxe et al., 2020). We leave collecting native multilingual datasets that include non-English data as future work, and strongly encourage the community to also adopt this practice.

[6]The original XCOPA release (Ponti et al., 2020b) does not contain the English section. We added the English release from SuperGLUE (Wang et al., 2019) to facilitate cross-lingual experiments.

[7]The native speakers were instructed to create a prompt that convert the task into a natural cloze-style question in their native language with no further restrictions.

Temp.	en	zh	es	hi	Avg
En (HW)	**50.8/50.6**	**48.5/47.7**	37.5/44.4	**44.0/45.5**	**45.2/47.0**
Zh (HW)	33.5/35.5	33.5/36.4	34.5/34.8	36.0/34.0	34.4/35.1
Es (HW)	39.2/49.9	44.8/45.3	**46.2/48.2**	41.5/43.5	42.9/46.7
Hi (HW)	45.0/43.5	39.5/41.0	34.2/40.5	36.2/40.5	38.8/41.4
Multi. (HW)	50.8/50.6	33.5/36.4	**46.2/48.2**	36.2/40.5	41.7/43.9
Multi. (MT)	50.8/50.6	35.8/39.5	36.5/45.0	41.0/39.9	41.0/43.8
Multi. (HT)	50.8/50.6	38.5/41.2	46.0/48.1	37.5/38.9	43.1/44.7

Table 5: 0/4-shot performance of $\text{XGLM}_{7.5\text{B}}$, evaluated on the first 400 examples of XNLI (development set in *en*, *zh*, *es* and *hi*) using different prompting approaches. Top: each target language is prompted by the language specified in column 1. Bottom: each target language is prompted by itself. HW: human-written. MT: machine-translated. HT: human-translated.

Temp.	en	zh	th	sw	Avg
En (HW)	**69.0/73.2**	63.0/66.8	**53.0/57.4**	**54.0/58.2**	**59.8**/63.9
Zh (HW)	63.0/71.0	**69.0/67.6**	50.0/57.8	47.0/54.2	57.2/62.6
Multi. (HW)	69.0/73.8	**69.0/67.6**	–	–	–
Multi. (MT)	69.0/73.8	62.0/68.4	48.0/56.6	51.0/60.2	57.5/**64.8**

Table 6: 0/4-shot performance of $\text{XGLM}_{7.5\text{B}}$, evaluated on XCOPA (development set in *en*, *zh*, *th* and *sw*).

prompting approaches[8]. English templates perform the best on average across languages for both tasks except for the 4-shot setting of XCOPA, where it slightly underperforms the machine translated templates. On the XNLI task, the English template significantly improves the performance of Chinese (*zh*) and Hindi (*hi*) over their native templates and translated templates. Similar trends are observed for Thai (*th*) and Swahili (*sw*) on XCOPA[9]. For both tasks there exist languages where the native templates strongly outperforms the English templates (Spanish (*es*) for XNLI and Chinese for XCOPA), indicating significant room for future work on language-specific prompt engineering.

4.4 Cross-lingual Transfer through Templates

We further examine if the ability of universal prompting is English specific, and in addition, what characterize a language pair for which cross-lingual prompting can work. To this end, we apply each of the human-written non-English templates to the rest of the languages. As shown in Table 5 and 6, using the Spanish prompt yields competitive 0- and 4-shot performance across all languages, with the 4-shot average performance being comparable to that of the English template. The Hindi template also achieves significantly above random performance on the XNLI tasks for most languages (especially *en*). The Chinese template, however, achieves close-to-random performance for all languages on XNLI, as well as close-to-random for Thai (0-shot) and Swahili (0-shot) on XCOPA. We hypothesize that the common sub-tokens and the amount of code-switching text in the pre-training data play a significant role in enabling cross-lingual prompting. And in general, high-resource languages with large amounts of pre-training data and vocabulary overlap with other languages act as better universal prompting languages. We leave a more systematic verification of this hypothesis to future work.

4.5 Cross-lingual Transfer through Demonstration Examples

We examine the capabilities of learning from cross-lingual demonstration examples (§3.2) of $\text{XGLM}_{7.5\text{B}}$ on XNLI. We examine two settings for each train-eval language pair: *same-language-prompting*, where the prompt templates and the example are in the same language, and *source-languauge-prompting* where the prompt templates for both the demo and test examples are in the source language. We use the human-translated prompts for *same-language-prompting*.

Table 7 shows results on a subset of language pairs of XNLI, where we evaluate transfer through demonstration examples from in-context demonstration examples from high-resource languages to lower-resourced ones, and between languages that are typologically similar. We report the difference between the 32-shot learning results and the 0-shot learning results. The non-English templates in this experiment are obtained via human-translation. While they typically underperform the in-language few-shot setting (Figure A2), most cross-lingual few-shot settings significantly improve over the 0-shot setting for the target language. Bulgarian is an exception as it does not benefit from Russian despite being in the same language family. Another language that does not work well in the cross-lingual settings is Swahili (*low* resource), for which we examined transfer from English (*high* resource) and Arabic (*medium* resource). In contrast, Thai (*medium*) and Urdu (*low* resource) significantly

[8]Appendix D.1 provides the comparison between English prompts and the MT and HT prompts on the complete dev sets of XNLI and XCOPA.

[9]The strong performance of English templates may be partially contributed to the fact that the non-English evaluation data on XNLI and XCOPA are obtained from translation. Testing how well the English templates perform on native non-English test sets is an interesting future work.

	high									medium		low
	en								ru	tr	ar	hi
	medium					low			medium	low		
prompt	bg	el	th	tr	vi	hi	sw	ur	bg	ur	sw	ur
Same-lang	**2.55**	**0.98**	**2.16**	**1.27**	**2.23**	**2.51**	-0.69	**1.21**	-2.49	-0.38	-1.64	**3.31**
Source-lang	-4.59	-2.44	**7.87**	-4.97	-1.08	**2.01**	-1.15	**7.42**	-1.43	**6.67**	-5.86	**2.31**

Table 7: Learning from cross-lingual demonstrations on XNLI, evaluated on the test set. The results are the absolute improvement over the zero-shot performance for the evaluated language using human-translated prompts. The first language group refers to the source language and the second one refers to the target language. *Same-lang* refers to a setting there the template is in the example language and *source-lang* refers to a setting where the template is only in the source language.

	Source prompt (instantiated)	Target prompt (instantiated)
Same-lang	The best thing that may be said of Podhoretz and Decter is that their biological clocks can't have many more minutes left on them, right? Yes, Decter is old.	Vâng, tôi thậm chí không nghĩ về điều đó, nhưng tôi đã rất thất vọng, và, tôi lại nói chuyện với anh ta lần nữa, đúng không? Đúng, tôi đã không nói chuyện với anh ta nữa.
Source-lang	The best thing that may be said of Podhoretz and Decter is that their biological clocks can't have many more minutes left on them, right? Yes, Decter is old.	Vâng, tôi thậm chí không nghĩ về điều đó, nhưng tôi đã rất thất vọng, và, tôi lại nói chuyện với anh ta lần nữa, right? Yes, tôi đã không nói chuyện với anh ta nữa.

Table 8: XNLI example prompts for cross-lingual transfer from *English (en)* to *Vietnamese (vi)*, with the *same-language* and *source-language* settings. The underlined text shows the verbalized part of the prompt.

benefit from cross-lingual demonstrations[10].

We also observed the benefit of cross-lingual transfer from demonstration examples is generally canceled if a better prompt (e.g. the English prompt) is used for the target language. We report the crosslingual demonstration experiments between all pairs of languages for XNLI, XCOPA and XStoryCloze and provide more discussion in Appendix D.2.

4.6 Performance on Multi-lingual Tasks

Using English as the universal prompting language, we characterize the zero- and few-shot in-context learning capabilities of $\text{XGLM}_{7.5\text{B}}$ on XNLI, XCOPA and XStoryCloze and compare them to English centric language models of comparable size.

Comparison to GPT-3. We compare $\text{XGLM}_{7.5\text{B}}$ to $\text{GPT-3}_{6.7\text{B}}$ on high, medium, low and extremely low resources languages[11]. The results are summarized in Table 9 and 10. On all three tasks, $\text{XGLM}_{7.5\text{B}}$ outperforms $\text{GPT-3}_{6.7\text{B}}$ by a large margin according to the average performance across languages, especially on medium, low and extremely low resource languages. On XNLI, $\text{GPT-3}_{6.7\text{B}}$ performs well on English and similar languages, surpassing $\text{XGLM}_{7.5\text{B}}$ on *en*, *de* (4-shot), *es* (4-shot), *fr* (0-shot). A possible explanation is that these languages have significant presence in the GPT-3 training data (*fr*: 1.8%, *de*: 1.5%, *es*: 0.8% as shown in Figure 1) and can benefit more from the lexical cognates from English.

Comparison to Translate-test Baseline. We also create a translate-test baseline, where we translate the non-English examples of the multilingual tasks to English using the Google Cloud Translation API[12] and use $\text{GPT-3}_{6.7\text{B}}$ *repl.*, an in-house replication of $\text{GPT-3}_{6.7\text{B}}$, to perform inference. We found the translate-test is a strong baseline of multilingual zero- and few-shot learning as is shown in Table 9 and 10. Across all three tasks, it significantly narrows the performance gap between English and other languages, especially on XNLI[13].

[10]Both Thai and Urdu obtained close-to-random zero-shot learning performances using the translated templates, which might make them easier to be further improved. Besides, there is inherent code switching in these languages (English presence in Thai and Urdu both lexical and morphological). Turkish and Arabic also have influence on Urdu. We hypothesize that these factors also positively impacted the cross-lingual in-context learning performance.

[11]We use GPT-3 Curie: https://blog.eleuther.ai/gpt3-model-sizes/

[12]https://cloud.google.com/translate

[13]The performance of translate-test baselines might be inflated given MT systems are often trained on backtranslations which makes it good at translating translationese (Edunov et al., 2019), which commonly exist in non-English evaluation data. Besides, the translation-test approach relies on high-quality machine translation (MT) systems trained on large amounts of parallel data.

model	# shot	high						medium							low			Avg.
		en	de	es	fr	ru	zh	ar	bg	el	th	tr	vi	hi	sw	ur	
GPT-$3_{6.7B}$	0	**55.4**	36.8	37.0	**51.2**	44.8	42.6	38.5	42.9	38.8	38.4	40.6	41.3	36.5	34.6	34.5	40.9
	4	53.0	**46.4**	**48.5**	48.3	44.3	45.8	38.2	41.7	42.1	36.8	38.7	42.3	34.3	33.7	34.5	41.9
$XGLM_{7.5B}$	0	55.3	**42.3**	**39.1**	50.8	**48.4**	**44.8**	**48.1**	**49.1**	**46.4**	**46.8**	**45.5**	**47.6**	**43.4**	**45.5**	**41.9**	**46.3**
	4	52.6	45.6	45.8	**49.4**	**48.6**	**48.8**	**46.4**	**48.9**	**48.7**	**46.6**	**45.4**	**48.5**	**46.8**	**44.5**	**43.4**	**47.3**
Translate + GPT-$3_{6.7B}$ *repl.*	0	54.6	53.7	54.5	53.9	52.0	52.6	52.0	53.4	53.5	50.6	53.3	52.6	50.7	51.3	48.7	52.5
	4	54.1	52.4	49.2	50.3	53.2	51.1	50.5	53.7	53.0	48.2	51.8	52.8	49.8	50.2	47.2	51.2

Table 9: Comparison of different models on XNLI.

		XStoryCloze												XCOPA										
model	# shot	high				medium				low			ex-low	Avg.	high	medium					low			ex-low	Avg.
		en	es	ru	zh	ar	id	hi	sw	te	eu	my			zh	id	it	th	tr	vi	et	sw	ta	ht	qu	
GPT-$3_{6.7B}$	0	73.4	62.4	56.9	55.8	48.4	56.6	50.1	49.4	52.8	51.2	49.5	55.1	55.0	60.2	**61.6**	53.6	53.4	52.8	50.8	52.2	55.0	51.8	**50.0**	54.2
	4	74.4	62.2	56.4	54.7	47.7	55.4	49.6	49.3	52.8	51.1	49.5	54.8	57.8	60.8	64.5	54.2	52.9	54.8	51.8	52.0	54.9	51.5	**49.7**	55.0
$XGLM_{7.5B}$	0	75.0	**68.1**	**71.0**	**66.6**	**58.3**	**70.1**	**60.9**	**65.0**	**61.7**	**62.3**	**60.7**	**65.4**	**62.4**	**66.6**	60.8	**56.8**	**56.8**	**61.4**	**61.6**	**57.6**	**56.2**	**57.0**	47.4	**58.6**
	4	75.9	**69.2**	**72.4**	**67.7**	**59.8**	**70.8**	**62.5**	**65.2**	**63.4**	**63.8**	**61.2**	**66.5**	**67.2**	**68.9**	**69.2**	**62.0**	**58.5**	**65.6**	**65.9**	**62.9**	**56.3**	**58.9**	47.1	**62.0**
Translate	0	81.2	75.6	75.4	72.9	71.5	71.2	70.5	70.0	66.9	70.5	72.7	72.6	75.0	73.2	76.0	53.8	72.4	72.2	72.4	63.8	67.2	65.0	-	67.4†
+ GPT-$3_{6.7B}$ *repl.*	4	82.6	75.0	75.3	73.1	71.8	72.0	71.6	71.0	68.4	72.2	72.0	73.2	78.5	75.8	80.6	57.7	73.7	76.0	73.6	67.2	69.9	67.0	-	70.0†

Table 10: Comparison of different models on XStoryCloze and XCOPA. †Google Translation API is not available for *qu*. For the averaged translate-test results we directly used the GPT-$3_{6.7B}$ *repl.* model for *qu* entry.

4.7 Performance on English Tasks

We also benchmark the performance of $XGLM_{7.5B}$ on English tasks. Figure 2 shows the comparison between $XGLM_{7.5B}$, GPT-$3_{6.7B}$ and GPT-$3_{6.7B}$ *repl.* on a subset of English tasks used by Brown et al. (2020). Our replication of GPT-$3_{6.7B}$, GPT-$3_{6.7B}$ *repl.*, performs better than or close to GPT-$3_{6.7B}$ on all tasks. While $XGLM_{7.5B}$ performs competitively on all tasks, there remains a considerable performance gap comparing to GPT-$3_{6.7B}$ and GPT-$3_{6.7B}$ *repl.*. On most tasks $XGLM_{7.5B}$ and GPT-$3_{6.7B}$ *repl.* show similar performance trend as k changes. For example, both models show a performance dip at 1-shot on HellaSwag and PIQA, and 128-shot on COPA.

There are multiple reasons why $XGLM_{7.5B}$ underperforms English centric models on the English tasks. First, only 32.6% of $XGLM_{7.5B}$'s 500B-token training data is English while both English-centric models are trained on close to 300B English tokens. Second, the model capacity of $XGLM_{7.5B}$ is shared by 30 languages, and the "curse of multilinguality" can degrade the performance across all languages (Conneau et al., 2020). Further scaling up the model capacity and training data can potentially close this gap. [14]

[14]The differences between the training corpora of the three models may have also contributed to the performance difference. While both English centric models incorporate high-quality English monolingual corpora such as BookCorpus (Zhu et al., 2019) in their training data (GPT-$3_{6.7B}$ also upsamples such high-quality data), $XGLM_{7.5B}$ is trained solely on data extracted from Common Crawl. However, we do not expect this to be the main impact factor. Scao et al. (2022) conducted a similar experiment showing that a multilingual model (1.3B parameters) pre-trained over 13 languages also significantly underperforms an English model trained from the same data source in terms of zero-shot generalization.

4.8 Performance on Machine Translation

We report machine translation results on popular WMT pairs in Table 11, and a subset of FLORES-101 (Goyal et al., 2021b) in Table 12. We use greedy decoding for both GPT-3 and our own model, and use the same 32 examples for few-shot learning in each case.

GPT-3 yields strong results on a few languages that are best represented in its training data, narrowly surpassing our model on WMT French-English, German-English, and Chinese-English, as well as a few pairs the FLORES-101 set. GPT-3 is particularly strong when English is the target language, presumably due to its strong English language modeling capability. However, it does poorly on the broader set of less-resourced languages. For instance, GPT-3 fails completely when translating into Korean, Arabic, Swahili, Hindi, Burmese and Tamil in FLORES-101, with a spBLEU score of 1.2 in the best case.

In contrast, our model obtains solid results across the board. In addition to surpassing GPT-3 in 171 out of 182 language pairs in the FLORES-101 set, our model is also competitive with the official supervised baseline for this dataset, even

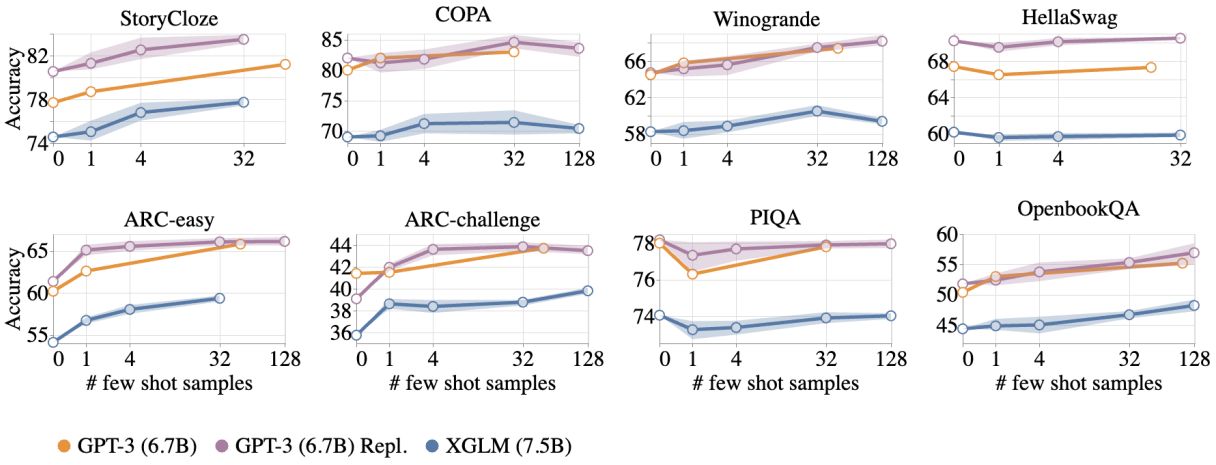


Figure 2: Performance on English tasks. For $\text{XGLM}_{7.5\text{B}}$ and $\text{GPT-3}_{6.7\text{B}}$ *repl.*, we plot the confidence interval from 5 different runs corresponding to different training sets when $k > 0$. For $\text{GPT-3}_{6.7\text{B}}$ we use the performance reported by Brown et al. (2020).

		WMT-14		WMT-16		WMT-19						Avg.	
		fr-en	en-fr	de-en	en-de	fi-en	en-fi	ru-en	en-ru	zh-en	en-zh	xx-en	en-xx
GPT-3 (API)	Ada	22.4	13.0	19.9	10.3	4.5	2.7	8.9	1.0	4.5	3.5	12.0	6.1
	Babbage	29.8	22.4	30.5	16.9	12.3	5.4	20.8	4.1	12.3	9.1	21.1	11.6
	Curie	**35.3**	**28.7**	**36.1**	**23.7**	18.4	9.9	28.6	9.8	**17.6**	**17.4**	**27.2**	17.9
$\text{XGLM}_{7.5\text{B}}$		33.2	28.5	34.6	23.5	**20.2**	**15.5**	**29.3**	**18.7**	16.7	**17.4**	26.8	**20.7**

Table 11: Machine translation results on WMT (detokenized BLEU). We use 32 examples from the previous edition for few-shot learning. BLEU scores computed using SacreBLEU with default settings (Post, 2018).

		en	de	fr	ca	fi	ru	bg	zh	ko	ar	sw	hi	my	ta	Avg.
avg out of xx	Supervised	24.0	21.0	20.4	19.1	17.5	18.6	20.2	15.5	14.9	16.1	16.6	16.2	7.2	4.8	16.6
	$\text{GPT-3}_{6.7\text{B}}$	9.9	9.1	9.4	9.3	6.4	7.0	5.5	4.9	2.4	2.9	1.7	0.5	0.2	0.3	5.0
	$\text{XGLM}_{7.5\text{B}}$	21.1	16.5	17.1	13.6	13.4	13.2	13.9	9.1	6.5	10.4	12.1	9.8	6.9	7.1	12.2
avg into xx	Supervised	26.0	20.2	26.7	20.0	16.7	18.5	24.5	14.1	13.5	11.8	16.3	19.3	2.1	2.5	16.6
	$\text{GPT-3}_{6.7\text{B}}$	18.9	9.9	14.2	9.3	4.2	4.8	2.7	4.0	0.6	0.5	0.2	0.3	0.1	0.1	5.0
	$\text{XGLM}_{7.5\text{B}}$	28.5	14.9	20.6	14.4	10.9	12.4	18.5	10.9	5.9	6.1	8.5	9.7	5.8	3.5	12.2

Table 12: Machine translation results on FLORES-101 devtest (spBLEU). $\text{GPT-3}_{6.7\text{B}}$ and $\text{XGLM}_{7.5\text{B}}$ use 32 examples from the dev set for few-shot learning. Supervised results correspond to the M2M-124 615M model from Goyal et al. (2021b). spBLEU computed using the implementation from Goyal et al. (2021b). Full results in Appendix D.3.

surpassing it in 45 language pairs. This suggests that large-scale multilingual language models have a great potential for building machine translation systems for low-resource languages, even if little or no parallel data is available.

4.9 Scaling up Model Size

Finally, we study the impact of scaling up the model parameter size on its 0- and few-shot learning capabilities. Figure 3 shows the performance ($k = 0, 4, 32, 128$) of the four XGLM models (564M, 1.7B, 2.9B, 7.5B) on the five multilingual tasks. The y-axis represents the average accuracy across languages for each task. On commonsense reasoning tasks (XStoryCloze, XCOPA, XWinograd), the performance of all models increases as k increases from 0 to 32. The performance gain from demonstration examples also gets larger as the model size increases, indicating bigger models can better leverage the in-context examples. On XNLI, the performance of all models increases as k increases from 0 to 4, but decreases for k at 32 and above. With the same number of demonstration examples, larger models do not always benefit more. PAWS-X is a task where in-context learning struggles – the performance of all models oscillates near random (50%) as k changes. A possible reason is the adversarial nature of PAWS-X, where the paraphrase and non-paraphrase pairs by design have high lexical overlap. We expect scaling to be an

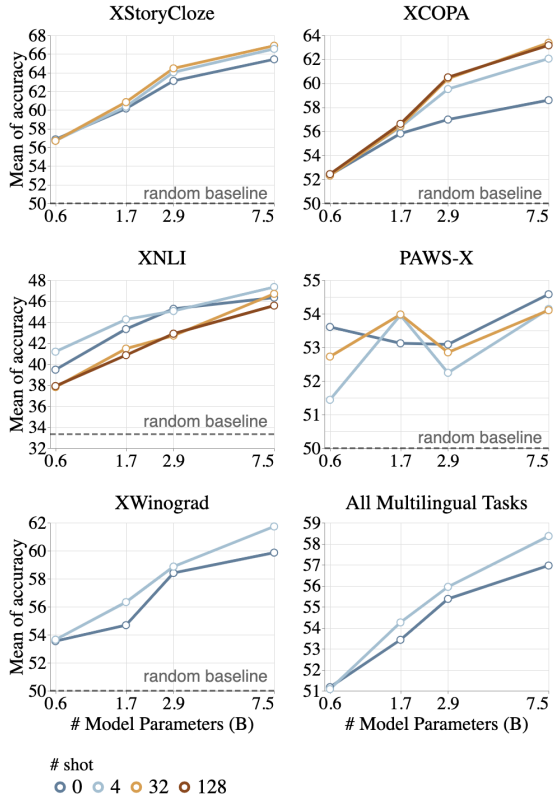


Figure 3: Zero-shot and in-language few-shot learning performance as a function of model size. The last plot shows the average performance over all five tasks in 0- and 4-shot learning.

effective recipe for building stronger multilingual language models, given the current trend.

5 Related Work

Language model prompting. Brown et al. (2020) first demonstrated in-context few-shot learning using the GPT-3 model. This method removes the need for task-specific updates to the model parameters: the few-shot examples that one would normally use for fine-tuning are provided at inference time to the same model for each task. On several high-resource Latin language pairs, GPT-3 achieves machine translation performance that is close to or better than state-of-the-art supervised models, given only a handful of demonstration examples.[15] Such change in the learning paradigm raises new questions about multilinguality, which has not been studied as extensively. Winata et al. (2021) evaluates the in-context few-shot learning abilities of several GPT-2, GPT_{NEO} and T5 on three additional languages (*de*, *es*, *fr*) using multiple NLU tasks, considering monolingual prompts as well as cross-lingual prompts, demonstrating the multilingual in-context learning skills of the English GPT and T5 models. Zhao and Schütze (2021) evaluated different fine-tuning and prompt-tuning (Liu et al., 2021) approaches on XLM-R and demonstrates the effectiveness of prompting in few-shot crosslingual transfer and in-language training of a multilingual masked language model.

Multilingual pre-training. Early multilingual pre-training work train word embeddings over multilingual corpora (Mikolov et al., 2013). The multilingual versions of contextualized embedding models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2019) and T5 (Raffel et al., 2020) were also developed: mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mBART (Liu et al., 2020), and mT5 (Xue et al., 2020). Such models were trained on a single, multilingual text corpus such as mC4 (Xue et al., 2020) or CC25 (Liu et al., 2020).

Several approaches have been developed to facilitate cross-lingual transfer, including sub-word tokenizers which enabled efficient, shared vocabulary learning across languages (Kudo and Richardson, 2018), joint training for efficient knowledge transfer across languages (Pires et al., 2019; Jiang et al., 2020; Kassner et al., 2021), etc. A notable concurrent work is BLOOM [16], which scales multilingual pre-training to 46 languages and 175 billion parameters.

6 Conclusion

We introduce four multilingual generative language models (XGLMs) at different scales, and study their in-context few- and zero-shot learning capabilities. We show that the few-shot learning capability of XGLM steadily improves as it scales. Our largest model (7.5B parameters) sets a new state of the art for few-shot learning in more than 20 languages (including mid- and low-resource languages) on commonsense reasoning, NLI and machine translation tasks. An in-depth analysis shows the models are highly cross-lingual, which leads to strong few-shot learning performance in non-English languages.

[15] Study shows that language contamination in pre-training data can effectively boost the cross-lingual capability of English-centric language models (Blevins and Zettlemoyer, 2022). With a heavier tail of deliberately introduced multilingual data, PALM-540B (Chowdhery et al., 2022) later achieves even stronger few-shot machine translation performance.

[16] https://bigscience.huggingface.co/blog/bloom

Limitations

Although the multilingual language model is an important step towards building inclusive general-purpose foundation models, our current models have the following limitations.

Training Data. Our models are trained on a static multilingual corpus extracted from CommonCrawl, with English text comprising 32.6% of the total number of tokens corresponding to 163B tokens. The English data portion of the corpus corresponds to roughly 54% only of GPT-3's training data. We applied several data filtering strategies as proxies for data quality assurance (see a comprehensive list in the Data Card in Appendix F), such as removing duplicated documents and paragraphs by URLs, filtering out paragraphs with high ratio of digits and punctuation, removing paragraphs with profanity, filtering by max number of URLs and minimum length, etc. Such filtering may potentially result in bias of the remaining data used in pretraining, which would need further analysis to understand. Furthermore, the raw data were taken from static CommonCrawl snapshots, which may not include entities and events beyond the time span of the snapshots (till March 2020), such as COVID-19, etc. As such we also note the potential difference in genres between CommonCrawl and the genres used in GPT-3 comprising in addition to CommonCrawl, corpora such as BookCorpus and Wikipedia.

Moreover, GPT-3 is trained on 118 languages despite the fact that 93% of the data is English.[17] In contrast our models are trained on 30 languages after rigorous language identification and filtering.

Performance on English tasks. As is shown in Section 4.7 and Figure 2, our model underperforms English-centric models on eight tasks ranging from commonsense reasoning to QA. There are several factors which could be contributing to this gap, such as

- Difference in training data quality (XGLM is trained on filtered CommonCrawl data only, while the English-centric models are trained on data including both CommonCrawl as well as high-quality corpora such as BookCorpus and Wikipedia) and quantity (as is described in the previous paragraph, the multilingual model was trained on 54% of the English data used in English-centric models);
- Curse of multilinguality. Previous work in multilingual training has shown that increasing the number of languages in model with shared parameters hurts performance on all training languages, e.g. English (Conneau et al., 2020).

Additional experiments controlling for these factors would shed more light on the observed gap.

Model architecture and training objective. In this work, we only experimented with causal language models with a decoder-only architecture, which had previously demonstrated promising few-shot learning capabilities (Brown et al., 2020). However, such architecture and pretraining objective do not leverage bidirectional context such as those used by masked language models (MLM), or sequence-to-sequence architectures with denoising autoencoder pretraining objectives.

Model evaluation via in-context learning. We compare our language models to the baselines primarily in the in-context learning paradigm, using the same prompts for all language models in the comparison unless explicitly specified. Despite minimal effort engineering the prompts for any model, it is possible that the prompts work better with some models than the others, which introduces bias to the evaluation. However, we expect this factor to have small impact and the relative strengths of the models can be reliably measured given the volume of tasks they were evaluated on.

Evaluation on social value tasks for more languages. We evaluate and analyze the models' performance on hate speech detection and gender bias for professional occupations. These studies are limited by the available evaluation datasets. We are limited in our study as we only investigate this problem space for six languages (English, French, Spanish, Italian, Portuguese, and Polish) where a majority of them (5) pertain to the Romance language family. It would be pertinent to investigate the impact of multilingual models on social value tasks across a wider and more diversified set of languages before drawing solid conclusions. Moreover, we contend that studies on other tasks such as stereotype (Nangia et al., 2020; Nadeem et al., 2021), ethics (Hendrycks et al., 2020) would pro-

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

vide a more comprehensive view of model behavior for social value tasks.

Ethical Considerations

Devising multilingual pre-trained language models can serve as a powerful tool in the NLP arsenal for multiple reasons.

Energy and maintenance efficiency. From an engineering perspective, XGLM pertains to a family of models that represent single unified models catering to many languages which have wide application across many applications. Such a unified single model saves on carbon footprint as well as energy consumption (comparing to the alternative: separate models for different languages) leading to more energy efficiency. A single model, despite having the risk of being a single point of failure, has the powerful incentive of being easier to maintain, access, distribute, and track.

Diversity and inclusion. Models such as XGLM represent a paradigm shift from the Anglo-centric view of the world of NLP to being able to cater to all languages on an equal footing. Paying attention to the design of such models is critical to ensure equitability and inclusion, exemplified here by attempting to balance language representation. The further power of XGLM specifically is its ability to perform comparably to Anglo-centric models in zero to few shot settings. Possessing powerful multilingual models that can perform well in such settings especially for medium to extremely low resource languages helps alleviate the burden of creating supervised data for such languages especially for economically challenged languages (medium to low digital presence typically goes hand in hand with economic disparities). Moreover, having such models catering to scarcer languages spurs scientific research in such languages leading to more diversified NLP, and more diversified science in the broader sense.

Social values. We further investigate the impact of our models on social valued problems such as hate speech detection and bias (Appendix §E). Despite inconclusive results overall (bordering on negative), we note that for the relatively scarcer data setting (Polish) the multilingual models outperform the Anglo-centric models indicating that XGLM will be performant for less resourced languages. This is especially significant for social value tasks where obtaining training data is quite problematic due to the inherent expense of obtaining high quality annotated data.

Transparency and Accountability. In the spirit of transparency and accountability for large-scale language modeling we include detailed model card and data card with the model and paper release.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7674–7684. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination explains the cross-lingual capabilities of english pretrained models. *CoRR*, abs/2204.08110.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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.

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. *CoRR*, abs/2204.02311.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

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

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. *CoRR*, abs/1809.05053.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2019. On the evaluation of machine translation systems trained with back-translation. *arXiv preprint arXiv:1908.05204*.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. `http://web.archive.org/save/http://Skylion007.github.io/OpenWebTextCorpus`.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, pages 394–398. The Association for Computer Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021a. Larger-scale transformers for multilingual masked language modeling. *CoRR*, abs/2105.00572.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021b. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1440–1448.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models. pages 5943–5959.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3250–3258. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pages 66–71.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. 2020. Limits to depth efficiencies of self-attention. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. pages 7871–7880.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. Technical report, AI21 Labs.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *CoRR*, abs/2103.10385.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. *CoRR*, abs/1809.02789.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *CoRR*, abs/1604.01696.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Association for Computational Linguistics (ACL)*.

Sebastian Nagel. 2016. Cc-news. `http://web.archive.org/save/http://commoncrawl.org/2016/10/news-dataset-available`.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. FAIRSEQ: A fast, extensible toolkit for sequence modeling. In *North American Association for Computational Linguistics (NAACL): System Demonstrations*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *CoRR*, abs/2105.11447.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

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

Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020a. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2362–2376. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020b. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 2362–2376, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research (JMLR)*, 21:1–67.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10215–10245. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022. What language model to train if you have one million GPU hours? In *Challenges & Perspectives in Creating Large Language Models*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3534–3546. Association for Computational Linguistics.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. 2020. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *CoRR*, abs/2109.07684.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. pages 483–498.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. *CoRR*, abs/1908.11828.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable prompt makes pre-trained language models better few-shot learners. *CoRR*, abs/2108.13161.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. *CoRR*, abs/2109.03630.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2019. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*.

A Pretraining Details

XGLM. All models are trained with the Fairseq library (Ott et al., 2019). We use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1e-8$. We adjust the learning rate based on model size, e.g. $1.5e-3$ for the 564M and 1.7B model, $7.5e-4$ for the 2.9B model, and $1.2e-4$ for the 7.5B models. Learning rates were adjusted with a 2000 warm-up updates followed by a polynomial decay schedule. All models are trained with data parallel and an effective batch size of 4M tokens. The XGLM 7.5B model was trained on 256 A100 GPUs for about 3 weeks, at a speed of 311.6k words per second[18].

GPT-$3_{6.7B}$ *repl.*. We replicate the GPT-$3_{6.7B}$ architecture and optimization hyperparameters to the best of our knowledge for training this model. The most significant difference between this model and GPT-$3_{6.7B}$ is in the training data. The training data used by GPT-$3_{6.7B}$ *repl.* is a combination of six English-language datasets, totaling 453GB and 112B tokens (which we up-sampled to 300B tokens):

- *BookCorpus* (Zhu et al., 2019), a dataset consisting of more than 10K unpublished books (4GB);
- *English Wikipedia*, excluding lists, tables and headers (12GB);
- *CC-News* (Nagel, 2016), a dataset containing 63 millions English news articles crawled between September 2016 and February 2019 (76GB);
- *OpenWebText* (Gokaslan and Cohen, 2019), an open source recreation of the WebText dataset used to train GPT-2 (38GB);
- *CC-Stories* (Trinh and Le, 2018), a dataset containing a subset of CommonCrawl data filtered to match the story-like style of Winograd schemas (31GB);
- *English CC100* (Wenzek et al., 2020), a dataset extracted from CommonCrawl snapshots between January 2018 and December 2018, filtered to match the style of Wikipedia (292GB).

The data are encoded using the same Byte-Pair Encoding (BPE) as GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019) with a vocabulary of 50K subword units.

A.1 Validation Perplexity

We use in-domain validation perplexity to validate the convergence status of the models. Figure A1 shows the average perplexity of the four models evaluated using a validation dataset sampled from CC100-XL. The validation data contains 30k sentences for each language that do not overlap with the pre-training data. We group the results by resource level.

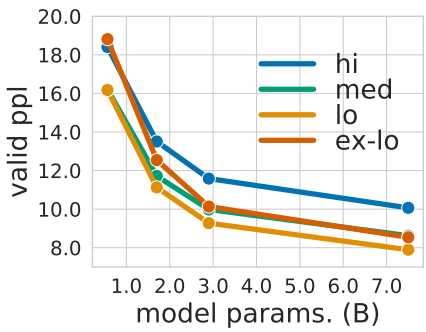


Figure A1: XGLM perplexity on CC100_XL validation set as a function of model size.

[18] On 256 A100 GPUs, the inference speed can reach 1.47 million words per second. Besides, inference can be done with significantly less resources. For example, using 8 v100 GPUs, it took 6 hrs to evaluate XGLM 7.5B on XStoryCloze.

B Multilingual In-context Learning Formulation

We extend the in-context learning framework proposed by Brown et al. (2020) to the multilingual setting. Let $\mathcal{M}$ be a causal language model and $\mathcal{D}$ be a task. $\mathcal{D} = (\mathcal{P}, \mathcal{E})$ consists of a task description $\mathcal{P}$ and a few demonstration examples in one or more languages

$$\mathcal{E} = \bigcup_{l=1}^{|\mathcal{L}|} \mathcal{E}^l.$$

We consider the setting where the task description comes in the form of a prompt $\mathcal{P} = (\mathcal{T}, v)$. $\mathcal{T}$ is a cloze-style template that converts an example input x into a string $\mathcal{T}(x)$ that contains a `[Mask]` symbol.[19] For classification and multiple-choice problems, $v : \mathcal{Y} \rightarrow \mathcal{V}^*$ is a verbalizer that maps each candidate label or choice $y \in \mathcal{Y}$ into a string $v(y)$. Both $\mathcal{T}(x)$ and $v(y)$ can be tokenized into a sequence of one or more tokens in the language model vocabulary $\mathcal{V}$. An instantiated prompt $\mathcal{P}(x, y)$ is obtained by substituting the `[Mask]` symbol in $\mathcal{T}(x)$ with $v(y)$. Table 2 shows the prompts used by all tasks in our main experiments.

Zero-shot learning. Given a test example $\tilde{x}^t$ in any target language t, the zero-shot prediction is $\hat{y}$ which maximizes a language model based scoring function (§C.2).

$$\hat{y} = \arg\max_y \sigma(\mathcal{M}, \mathcal{P}(\tilde{x}^t, y)). \tag{1}$$

This general formulation can cover most NLP tasks. For classification problems, v is a mapping from classes to strings; for multiple-choice problems, v is an identity function that maps each candidate choice to itself. For text generation problems, v is identity and we decode free-form text from `[Mask]`, which in this case is positioned at the end of $\mathcal{T}(x)$.

Few-shot learning. Suppose we have k demonstration examples available in a source language:

$$\mathcal{E}^s = \{(x_i^s, y_i)\}_{i=1}^k.$$

In this case, we concatenate the instantiated prompts of the demonstration examples $\{\mathcal{P}(x_i^s, y_i)\}_{i=1}^k$ and make it the prefix of the input string used in the zero-shot learning setting to form the objective:

$$\hat{y} = \arg\max_y \sigma(\mathcal{P}(x_1^s, y_1) \texttt{ [Sep] } \ldots \mathcal{P}(x_k^s, y_k) \texttt{ [Sep] } \mathcal{P}(\tilde{x}^t, y)), \tag{2}$$

where `[Sep]` is a separator symbol chosen empirically.

- When $s = t$, we have the *in-language* few-shot learning setup.
- When $s \neq t$, we have the *cross-lingual* few-shot learning setup.

C Evaluation Details

C.1 English Evaluation Tasks

Table A1 shows all the English tasks used in our evaluation.

[19] We relaxed the prompt format of GPT-3 by allowing the `[Mask]` symbol to appear anywhere in $\mathcal{T}(x)$ instead of just in the end. Having this additional flexibility leads to better performance on some tasks. This is inspired by the masked language modeling prompts constructed by recent work (Schick and Schütze, 2021; Zhang et al., 2021).

Reasoning	StoryCloze (Mostafazadeh et al., 2016)	–	1,871	1,871	N/A	1
	COPA‡ (Gordon et al., 2012)	400	100	499	N/A	1
	WinoGrande (Sakaguchi et al., 2020)	40,398	1,267	1,767	N/A	1
	HellaSwag (Zellers et al., 2019)	39,905	10,042	10,003	N/A	1
QA	ARC-easy (Clark et al., 2018)	2,251	570	2,376	N/A	1
	ARC-cha. (Clark et al., 2018)	1,119	299	1,172	N/A	1
	PIQA (Bisk et al., 2020)	16,113	1,838	3,084	N/A	1
	OpenbookQA (Mihaylov et al., 2018)	4,957	500	500	N/A	1

Table A1: English tasks used in our few-shot learning evaluation. All tasks use accuracy as the evaluation metrics.

C.2 Scoring Functions

We considered the following functions for scoring an instantiated prompt using a language model:

(1) sum of per-token log probabilities;

(2) average of per-token log probabilities;

(3) average of per-token log probabilities, ignoring the common prefix of different candidates.

We also considered the calibration approach proposed by Zhao et al. (2021) and character normalization proposed by Lieber et al. (2021).

In the end, we use the average of per-token log-probabilities ignoring the common prefix of different candidates as the scoring function for all multilingual tasks. This is selected based on the development set performance of StoryCloze and XNLI.

For English tasks, we use the same modeling choices as Brown et al. (2020). Specifically, we use the task prompts as detailed in Appendix G of Brown et al. (2020), and a single newline as the separator for few-shot learning. For WinoGrande, we take the log-likelihood of the common suffix of the different candidates as the scoring function. For ARC-easy, ARC-challenge and OpenBookQA, we normalize by the unconditional probability of each candidate by taking $\frac{p(\texttt{completion}|\texttt{context})}{p(\texttt{completion}|\texttt{answer_context})}$, where we use the string *"Answer: "* as answer_context. For all the other tasks, we take the average of per-token log-probabilities, ignoring the common prefix of the different candidates.

C.3 Evaluation Protocol

All few-shot learning results are obtained with the *in-language* setting (both the training and test examples are in the same language) unless otherwise specified. We report results on the test set for all multilingual tasks (including the held-out tasks). For English tasks, we report results on the test set for ARC-easy, ARC-challenge, OpenBookQA and StoryCloze, and on the development set for the rest, following Brown et al. (2020). For few-shot learning, we report the average results across 5 runs, randomly sampling a different set of few-shot examples each time. For tasks with a training set, we sample the few-shot examples from the training set; for tasks with no training set, we sample from the dev set and report evaluation results on the test set; for dev-set examples on XNLI and XCOPA, we sample few-shot examples from the test set, since these two tasks do not have the training sets for all languages. While Brown et al. (2020) tuned the few-shot value k as a hyperparameter on the dev set, we pre-selected a few k values (0, 1, 4, 32, 128) and report the corresponding results.

C.3.1 Example Truncation

Following Brown et al. (2020), we truncate the input such that they fit the maximum context length of XGLM ($n_{\text{ctx}} = 2048$) and preserve only the complete demonstration examples after truncation. For each task, we report results up to the k's corresponding to the maximum fit.[20] Table A2 shows the average number of demonstration examples that fit the maximum context length of XGLM ($n_{\text{ctx}} = 2048$) for each task in our experiments.

[20]XWinograd has only a test split, and we sampled few-shot examples directly from it, following the practice used by Brown et al. (2020) for evaluating GPT-3 on Winograd. As a result we only report 0-, 1- and 4-shot results for XWinograd to minimize inflating the few-shot performance by training and testing on the same examples.

XStoryCloze	en	zh	ru	es	id	ar	hi	sw	te	eu	my				
	32.0	31.6	30.6	31.5	32.0	27.6	24.7	29.5	25.2	25.6	18.8				
XCOPA	en	zh	it	id	th	vi	tr	et	ta	sw	ht	qu			
	100.0	100.0	98.0	100.0	100.0	99.4	100.0	100.0	75.2	97.9	95.4	84.9			
XWinograd	en	ru	ja	pt											
	93.7	63.7	62.4	83.1											
XNLI	en	zh	ru	es	de	fr	el	th	vi	tr	ar	bg	hi	ur	sw
	48.3	47.6	43.4	44.7	43.1	39.3	37.8	44.8	39.8	46.8	42.3	41.4	37.4	38.4	42.9
PAWS-X	en	zh	es	ja	de	fr	ko								
	34.5	27.2	31.3	23.1	32.0	29.1	28.0								
	en														
COPA	124.3														
Winogrande	84.6														
HellaSwag	21.4														
ARC-easy	62.9														
ARC-challenge	53.2														
PIQA	59.6														
OpenbookQA	106.4														

Table A2: Average # of few-shot examples that fit the maximum context length of XGLM ($n_{\text{ctx}} = 2048$) in our few-shot evaluation benchmark. The languages are sorted according to the amount of pre-training data (high to low).

Representation Bias. We observe that the language model tend to fit more examples in a high-resource language in context compared to those in a low-resource language.[21] English, as the highest resourced language (Table A10), always fit the most examples. This reflects the unequal representation of different languages in our joint multilingual BPE vocabulary (§2.1). With this vocabulary induction scheme (Sennrich et al., 2015), the underrepresented languages tend to have smaller sub-word units and higher fertility (defined as number of subwords per linguistic word), making it more challenging to learn word- and higher-level semantics for such languages. Other factors can also impact the tokenization granularity. For example, sharing sub-strings with other high resource languages can boost the granularity of a language; and some languages have smaller tokenization granularity as a result of their alphabet system (e.g. Chinese has an average sub-word length of 1.4, indicating the dominance of single-character tokens, despite being the third largest language in our pre-training data according to disk size).

D Additional Results

D.1 Comparing Multilingual Prompting Approaches on XNLI and XCOPA

We compare the performance of English prompts and `MT` and `HT` prompts on two of our held-out tasks, XNLI and XCOPA, using their development sets. For `MT` prompts, we translate the English prompts into the target languages using the Google Cloud Translation API. We use the exact prompts as shown in Table 2 as the input of the translation API and manually recover the placeholders in the API output based on brackets markers (e.g. "`{Sentence 1}` because `[Mask]`" is translated to "`{Sentence 1}`因为`[Mask]`"). When the candidate set is closed, we replace `[Mask]` with each verbalized label and translate them separately. For example, "`{Sentence 1}`, right? Yes, `{Sentence 2}`" is translated to "`{Sentence 1}`，对吗？是的，`{Sentence 2}`". On XNLI, we also compared to prompts manually translated from English to eliminate the impact of translation noise on the comparison.[22]

As shown in Table A3 and Table A4, the in-context learning performance is sensitive to the prompting choices across all languages. For both XNLI and XCOPA, using the English prompts on average yield significantly better performance than using the machine-translated prompts. For XNLI, human translated (HT) prompts significantly improve over machine translated (MT) prompts for most languages. Surprisingly, the performance of human translated prompts lags behind that of the English prompts in the

[21] XStoryCloze, XCOPA, XNLI and PAWS-X all contain parallel examples, which allows us to compare the maximum fit of the same set of examples across different languages.

[22] We ask native speakers to translate the English template into *zh*, *es*, *fr*, *el*, *hi*, *vi*, *ar* and *bg*. For the rest of the languages, one of the authors verified and corrected the machine translated templates using bilingual dictionaries.

		hi	medium											low			
# shot	prompt	en	zh	ru	es	de	fr	el	th	vi	tr	ar	bg	hi	ur	sw	Avg.
0	En	54.5	**45.0**	47.2	38.2	42.4	**50.7**	**47.1**	**46.1**	**47.5**	44.6	**47.5**	**50.0**	**43.4**	**42.7**	**46.2**	**46.2**
	MT	54.5	34.6	40.7	37.5	37.6	47.8	45.4	33.3	35.4	37.1	46.5	49.3	38.8	33.5	44.4	41.1
	HT	54.5	34.1	**47.4**	**50.0**	**49.5**	50.4	**47.1**	34.9	45.2	**46.3**	46.0	49.3	37.5	33.7	44.4	44.7
4	En	51.8	**48.0**	48.2	45.1	45.2	49.2	**48.4**	**46.3**	**48.2**	45.1	**46.4**	**49.1**	**46.6**	**43.5**	**44.9**	**47.1**
	MT	51.8	39.8	45.1	45.8	43.6	**49.4**	44.3	33.7	41.9	35.0	45.5	48.6	39.9	35.4	43.5	42.9
	HT	51.8	39.8	**50.0**	**49.9**	**49.8**	45.7	44.0	37.3	41.9	**47.0**	45.7	48.6	40.2	35.0	43.5	44.7
32	En	53.4	**50.4**	40.0	46.4	46.2	46.7	47.3	**46.8**	**48.6**	44.3	43.3	42.8	**45.6**	**45.6**	**46.5**	**46.3**
	MT	53.4	40.7	38.0	47.0	43.1	**49.1**	48.5	35.0	44.5	35.2	41.9	**51.9**	37.2	34.8	43.8	42.9
	HT	53.4	43.1	**51.8**	**51.7**	**49.6**	46.2	**48.9**	38.0	47.0	**47.5**	**43.9**	**51.9**	40.9	34.6	43.8	46.2

Table A3: Comparison between English prompts, MT (machine-translated) prompts and HT (human-translated) prompts for 0, 4 and 32-shot learning on XNLI dev set using $\text{XGLM}_{7.5\text{B}}$.

		hi	medium						low			ex-low		
# shot	prompt	zh	ru	it	id	th	vi	tr	et	ta	sw	ht	qu	Avg.
0	En	**63.0**	**64.0**	59.0	**54.0**	**53.0**	63.0	**65.0**	**63.0**	**59.0**	**54.0**	57.0	**70.0**	**60.3**
	MT	62.0	59.0	**69.0**	51.0	48.0	**64.0**	64.0	60.0	**59.0**	51.0	**60.0**	69.0	59.7
4	En	66.8	**73.8**	66.2	**54.6**	**57.4**	**68.8**	69.8	**69.2**	60.6	58.2	**62.6**	**62.2**	**64.2**
	MT	**68.4**	65.4	**68.8**	54.0	56.6	67.2	67.4	66.6	**62.2**	**60.2**	60.8	61.8	63.3

Table A4: Comparison bewtween English prompts and MT (machine-translated) prompts for 0-shot and 4-shot learning on XCOPA dev set using $\text{XGLM}_{7.5\text{B}}$.

0-shot and 4-shot settings.

Further examination of the per-language performance reveals that the relative strengths of different prompting approaches vary across languages. For *es* and *de*, HT prompts offer large gains compared to the MT prompts and the English prompts. However, for *zh* and *ur*, using translated prompts (either HT or MT) significantly hurts the performance. For *zh*, *fr*, *vi*, *ar* and *hi*, using native-speaker translated prompts still yields significantly lower performance compared to using the English prompts in at least one setting, suggesting that translation error is not the sole cause of the performance drop.

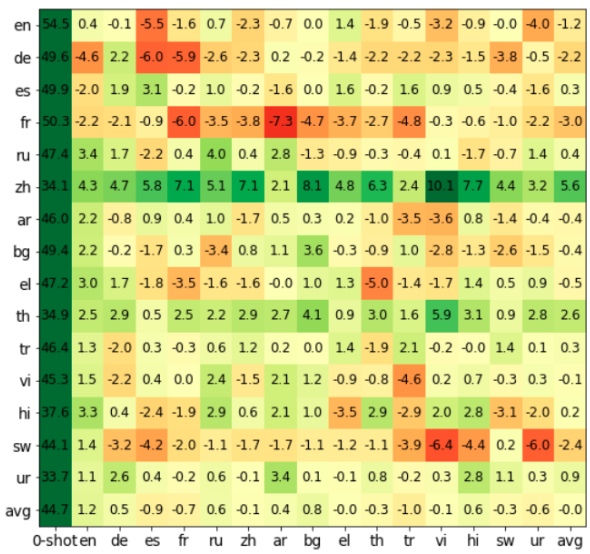
D.2 Full Results on Learning from Cross-lingual Demonstrations

We evaluated $\text{XGLM}_{7.5\text{B}}$ on XNLI in the learning from cross-lingual demonstration setting, using both the *same-language-prompting* and *English-prompting* setups. In *same-language-prompting*, the prompt fields and the examples are always in the same language. And in *English-prompting*, English prompts is used for all examples. All few-shot performances in this section are obtained using the k-shot *per label* setting as described in §D.4.

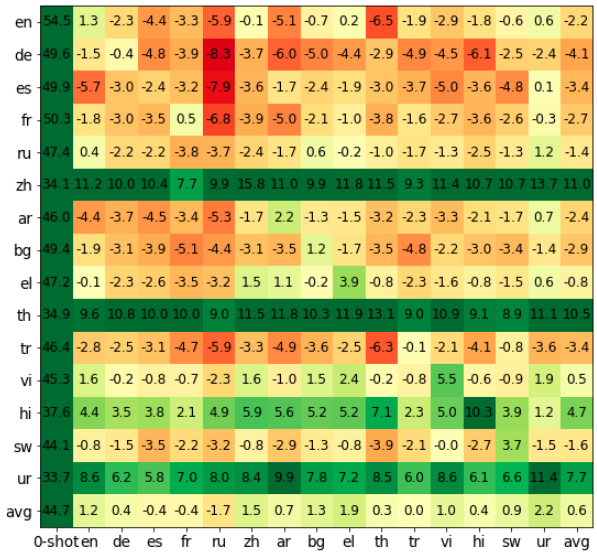
As shown in Figure A2, for many language pairs transferring from source language demonstration can significantly improve over the zero-shot performance in the target language when human-translated templates is used. The improvement is especially significant for languages such as Chinese (*zh*), Thai (*th*) and Urdu (*ur*), whose zero-shot performance is close to random with human translated templates. However, we found that the effect of cross-lingual transfer from template and cross-lingual transfer from demonstration examples typically do not add up. As shown in Figure A3, using the English template significantly improves the zero-shot performance of most languages, including Chinese, Thai and Urdu. In this case, the demonstration examples in general do not help unless they are in the same language as the target example (diagonals).

Figure A4 shows the results on XCOPA.

Figure A5 shows the results on XStoryCloze, where we observed almost no improvement for any language pair. Possible reasons for the poor transfer results on XStoryCloze is that it requires reasoning about implicit relations between multiple sentences which is much harder to do especially in a cross-lingual setting.

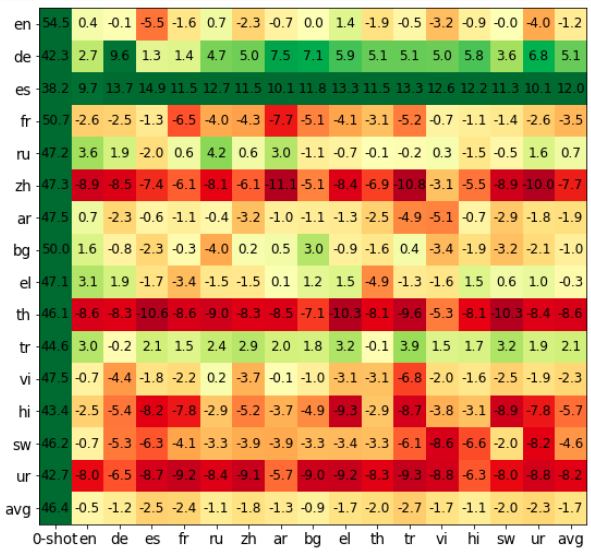


(a) Same-language prompting with human-translated templates

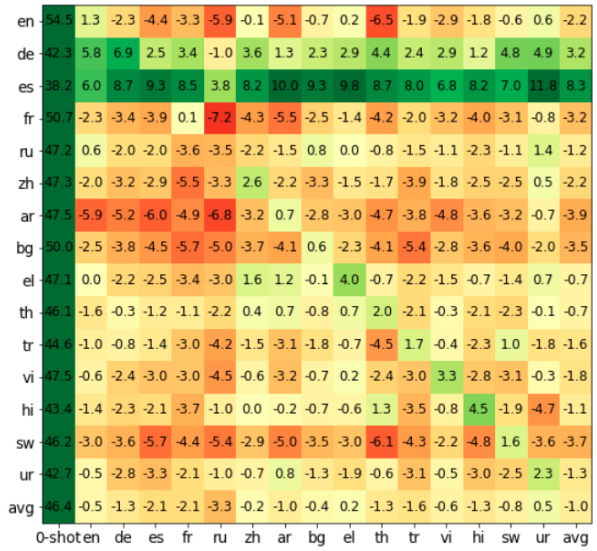


(b) English prompting

Figure A2: Cross-lingual few-shot in-context learning on XNLI development set. The leftmost column shows the 0-shot performance (*with human-translated templates*) of each language. The rest of the matrix shows the difference between 4-shot (per label) and 0-shot (*with human-translated templates*) performance. Row: target language. Column: source language.

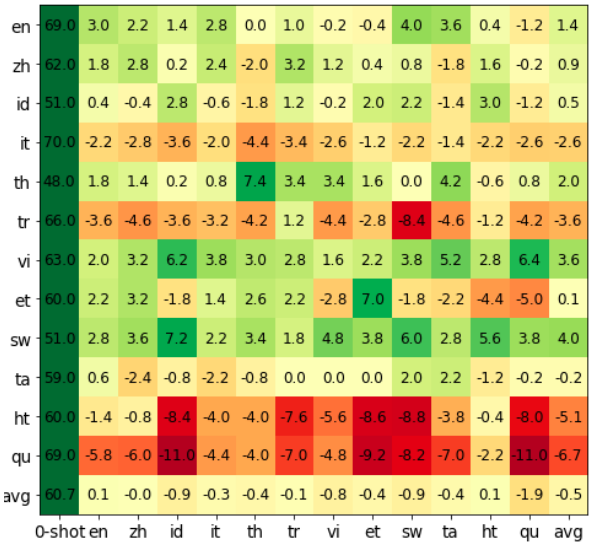


(a) Same-language prompting with human-translated templates

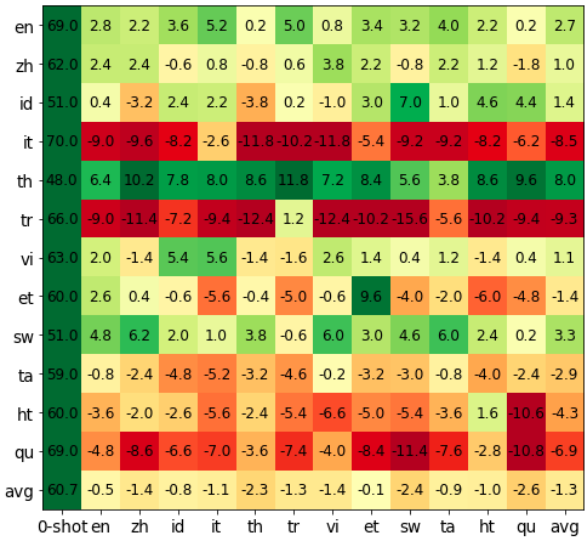


(b) English prompting

Figure A3: Cross-lingual few-shot in-context learning on XNLI development set. The leftmost column shows the 0-shot performance (*with English templates*) of each language. The rest of the matrix shows the difference between 4-shot (per label) and 0-shot (*with English templates*) performance. Row: target language. Column: source language.



(a) Same language prompting with human-translated templates



(b) English prompting

Figure A4: Cross-lingual few-shot in-context learning on XCOPA development set. The leftmost column shows the 0-shot (*with machine translated templates*) performance. The rest of the matrix shows the difference between 4-shot (per label) and 0-shot (*with machine translated templates*) performance. Row: target language. Column: source language.

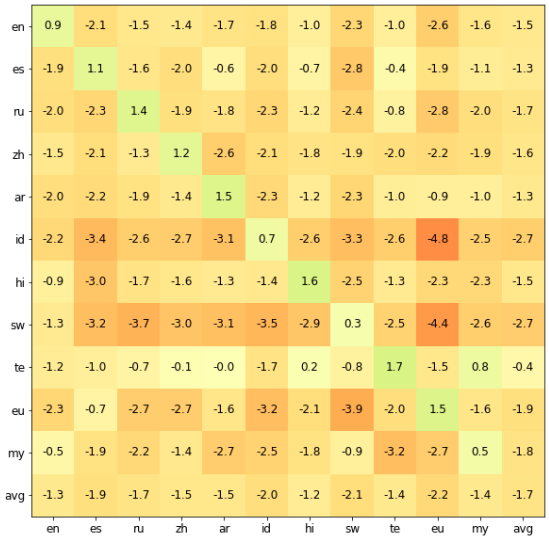


Figure A5: Cross-lingual few-shot in-context learning on XStoryCloze test set. The matrix shows the difference between 4-shot (per label) and 0-shot performance. For XStoryCloze, there is no difference between same-language prompting and English prompting since the task does not use a verbalized template. Row: target language. Column: source language.

D.3 Full Results in FLORES-101

Table A5 reports our full results in FLORES-101.

		en	de	fr	ca	fi	ru	bg	zh	ko	ar	sw	hi	my	ta	avg
en	Supervised	–	32.6	42.0	31.2	24.2	27.1	37.4	19.3	18.5	17.9	26.9	28.1	3.5	3.4	24.0
	GPT-$3_{6.7B}$	–	25.9	**36.1**	23.8	10.2	11.2	5.9	12.5	1.2	1.1	0.5	0.3	0.1	0.0	9.9
	$XGLM_{7.5B}$	–	**27.6**	36.0	**34.0**	**23.3**	**24.2**	**33.1**	**15.6**	**12.0**	**11.5**	**18.0**	**19.9**	**11.0**	**8.5**	**21.1**
de	Supervised	35.8	–	35.5	25.8	22.6	24.6	31.5	17.2	16.6	14.8	21.0	23.4	2.3	2.3	21.0
	GPT-$3_{6.7B}$	**40.4**	–	26.2	17.2	8.1	9.3	4.8	9.0	1.0	0.9	0.5	0.3	0.1	0.1	9.1
	$XGLM_{7.5B}$	38.8	–	**27.9**	**19.1**	**20.5**	**19.7**	**25.8**	**12.3**	**3.4**	**6.6**	**11.7**	**14.3**	**9.9**	**4.8**	**16.5**
fr	Supervised	37.2	28.5	–	28.7	21.9	24.5	32.2	17.6	16.7	15.4	17.2	22.9	2.1	0.8	20.4
	GPT-$3_{6.7B}$	**42.8**	**20.9**	–	23.7	8.0	9.7	4.6	9.1	1.0	1.0	0.4	0.3	0.1	0.0	9.4
	$XGLM_{7.5B}$	40.4	20.4	–	**32.1**	**19.4**	**19.8**	**26.3**	**10.6**	**2.4**	**5.9**	**14.5**	**13.7**	**9.7**	**6.6**	**17.1**
ca	Supervised	33.4	24.8	35.1	–	19.0	21.1	28.6	15.1	13.9	13.4	18.7	20.5	2.1	2.6	19.1
	GPT-$3_{6.7B}$	40.2	18.6	31.4	–	7.0	**9.3**	4.3	8.0	0.9	**0.9**	0.3	0.4	0.1	0.1	9.3
	$XGLM_{7.5B}$	**41.1**	**18.9**	**33.8**	–	**11.3**	3.3	**23.9**	**10.8**	**1.3**	0.8	**13.8**	**6.1**	**7.9**	**3.1**	**13.6**
fi	Supervised	27.2	23.0	29.3	21.6	–	20.6	26.4	16.0	14.8	12.4	14.2	19.8	1.7	0.9	17.5
	GPT-$3_{6.7B}$	25.3	13.5	17.1	10.0	–	6.4	2.8	5.7	0.7	0.7	0.3	0.3	0.1	0.0	6.4
	$XGLM_{7.5B}$	**29.2**	**17.4**	**22.2**	**17.0**	–	**16.5**	**17.5**	**12.4**	**7.5**	**7.6**	**8.0**	**10.1**	**6.2**	**2.0**	**13.4**
ru	Supervised	27.5	23.5	30.1	22.0	19.4	–	31.0	16.5	15.3	13.5	18.1	20.9	2.2	2.3	18.6
	GPT-$3_{6.7B}$	28.1	14.8	20.4	13.1	5.4	–	7.4	1.2	0.2	0.2	0.1	0.2	0.1	0.1	7.0
	$XGLM_{7.5B}$	**30.4**	**17.9**	**24.0**	**14.6**	**8.0**	–	**26.3**	**11.6**	**5.5**	**7.4**	**7.1**	**9.1**	**7.3**	**3.1**	**13.2**
bg	Supervised	33.0	26.1	33.7	24.9	20.8	26.5	–	17.5	16.4	14.5	20.9	23.1	2.3	2.4	20.2
	GPT-$3_{6.7B}$	21.6	11.4	16.0	9.7	4.3	6.5	–	1.2	0.2	0.2	0.1	0.2	0.1	0.1	5.5
	$XGLM_{7.5B}$	**35.5**	**19.2**	**26.3**	**12.9**	**14.2**	**22.9**	–	**11.9**	**6.8**	**9.2**	**9.4**	**7.5**	**3.2**	**1.0**	**13.9**
zh	Supervised	20.9	17.6	24.3	17.4	16.0	17.2	22.1	–	15.9	11.6	15.5	18.5	1.9	2.5	15.5
	GPT-$3_{6.7B}$	**21.1**	**9.5**	**14.3**	8.2	4.3	3.6	1.3	–	1.1	0.4	0.2	0.2	0.1	0.0	4.9
	$XGLM_{7.5B}$	20.7	8.3	8.5	**10.5**	**4.4**	**4.8**	**14.8**	–	**9.3**	**4.2**	**5.6**	**12.0**	**8.6**	**6.2**	**9.1**
ko	Supervised	20.9	16.7	22.1	16.5	14.9	15.5	21.1	15.7	–	10.6	15.1	18.7	1.9	4.0	14.9
	GPT-$3_{6.7B}$	8.3	4.6	6.4	4.4	**2.1**	**1.7**	0.8	2.5	–	0.2	0.1	0.1	0.1	0.1	2.4
	$XGLM_{7.5B}$	**19.9**	**10.3**	**13.7**	**5.3**	1.4	1.2	**10.9**	**11.9**	–	**2.7**	**3.2**	**1.0**	**2.2**	**1.4**	**6.5**
ar	Supervised	25.5	18.7	25.7	18.9	15.6	17.8	23.8	13.1	13.3	–	15.4	19.4	1.8	0.9	16.1
	GPT-$3_{6.7B}$	10.5	5.3	9.6	6.0	2.2	2.2	0.9	0.9	0.1	–	0.1	0.1	0.2	0.0	2.9
	$XGLM_{7.5B}$	**27.7**	**12.2**	**17.9**	**8.8**	**8.5**	**9.1**	**18.4**	**8.9**	**0.8**	–	**7.7**	**7.8**	**3.4**	**3.7**	**10.4**
sw	Supervised	30.4	19.4	26.7	20.1	15.6	17.6	23.8	13.2	12.2	12.0	–	19.2	2.1	4.0	16.6
	GPT-$3_{6.7B}$	5.0	2.9	3.9	2.8	1.7	1.8	1.3	1.3	0.5	0.5	–	0.4	0.1	0.1	1.7
	$XGLM_{7.5B}$	**31.6**	**13.4**	**21.8**	**15.4**	**10.2**	**13.1**	**15.2**	**9.5**	**6.0**	**8.9**	–	**7.6**	**3.4**	**1.0**	**12.1**
hi	Supervised	27.9	19.4	25.9	18.9	15.7	16.9	23.9	13.5	13.9	12.2	16.8	–	2.5	3.8	16.2
	GPT-$3_{6.7B}$	1.2	0.9	1.4	0.8	0.4	0.4	0.3	0.2	0.1	0.1	0.1	–	0.1	0.2	0.5
	$XGLM_{7.5B}$	**25.2**	**12.3**	**15.4**	**8.8**	**9.8**	**11.5**	**11.3**	**10.8**	**8.5**	**6.1**	**4.7**	–	**1.5**	**1.9**	**9.8**
my	Supervised	10.0	6.9	10.4	8.5	6.0	6.7	9.5	5.7	6.1	4.6	7.2	9.1	–	2.5	7.2
	GPT-$3_{6.7B}$	0.5	0.3	0.4	0.4	0.2	0.1	0.2	0.0	0.0	0.0	0.1	0.2	–	0.1	0.2
	$XGLM_{7.5B}$	**14.1**	**7.6**	**10.1**	**3.8**	**5.7**	**7.1**	**8.9**	**7.1**	**6.9**	**3.6**	**3.5**	**8.9**	–	**2.6**	**6.9**
ta	Supervised	8.3	4.9	6.8	5.8	5.0	4.7	7.0	2.5	2.3	1.1	5.2	6.9	1.2	–	4.8
	GPT-$3_{6.7B}$	1.0	0.5	0.8	0.5	0.2	0.3	0.3	0.1	0.2	0.1	0.1	0.2	0.0	–	0.3
	$XGLM_{7.5B}$	**16.3**	**8.4**	**10.3**	**5.1**	**5.2**	**8.1**	**7.6**	**8.1**	**6.2**	**5.4**	**2.8**	**7.2**	**0.9**	–	**7.1**
avg	Supervised	26.0	20.2	26.7	20.0	16.7	18.5	24.5	14.1	13.5	11.8	16.3	19.3	2.1	2.5	16.6
	GPT-$3_{6.7B}$	18.9	9.9	14.2	9.3	4.2	4.8	2.7	4.0	0.6	0.5	0.2	0.3	0.1	0.1	5.0
	$XGLM_{7.5B}$	**28.5**	**14.9**	**20.6**	**14.4**	**10.9**	**12.4**	**18.5**	**10.9**	**5.9**	**6.1**	**8.5**	**9.7**	**5.8**	**3.5**	**12.2**

Table A5: Machine translation results on FLORES-101 devtest (spBLEU). Source language in rows, target language in columns. GPT-$3_{6.7B}$ and $XGLM_{7.5B}$ use 32 examples from the dev set for few-shot learning. Supervised results correspond to the M2M-124 615M model from Goyal et al. (2021b). Underline denotes better than supervised, bold denotes best of GPT-3 and XGLM. spBLEU computed using the implementation from Goyal et al. (2021b).

D.4 Majority Label Bias

In the main paper, we define k-shot learning as learning from k unique examples randomly drawn from the entire training population. This setting may lead to skewed few-shot training sets, especially when k is small. As shown in Table A6, the XNLI task is a three-way classification problem where the model needs to judge whether the relationship between a pair of sentences is "entailment", "neurtral" or "contradiction". While the original XNLI dev set has a uniform class distribution, the few-shot training sets randomly

	24-shot			48-shot		
Seed	E	N	C	E	N	C
0	9	7	8	22	11	15
1	6	11	7	12	16	20
2	9	7	8	20	11	17
3	8	11	5	16	16	16
4	6	4	14	14	16	18

Table A6: Distribution of XNLI few-shot training sets obtained by randomly sampling from the original dev set. `E`: entailment, `N`: neutral, `C`: contradition.

# shots		high						medium						low			Avg	Std
		en	zh	ru	es	de	fr	el	th	vi	tr	ar	bg	hi	ur	sw		
24	rand.	54.0 (±2.3)	50.0 (±1.9)	41.5 (±6.1)	47.1 (±4.0)	46.1 (±2.1)	47.7 (±4.1)	48.8 (±2.2)	47.8 (±2.7)	48.8 (±3.5)	44.8 (±2.0)	45.4 (±3.7)	44.4 (±4.8)	45.8 (±2.4)	45.6 (±1.4)	46.4 (±3.0)	47.0	2.9
	unif.	**56.0** (±1.4)	**52.1** (±0.8)	**42.9** (±1.2)	**47.6** (±0.8)	**49.2** (±1.0)	**49.2** (±1.6)	**50.9** (±1.6)	**48.0** (±1.2)	**50.2** (±1.5)	**47.5** (±1.0)	**47.3** (±0.8)	**47.0** (±2.2)	**49.4** (±1.2)	**46.6** (±0.9)	**47.4** (±0.8)	**48.7**	2.9
Max	rand.	54.2 (±2.2)	51.7 (±1.1)	37.4 (±1.3)	45.4 (±3.2)	44.7 (±3.2)	45.0 (±3.0)	46.9 (±2.6)	45.8 (±2.3)	47.9 (±2.7)	42.0 (±1.4)	42.3 (±1.8)	40.4 (±1.6)	45.6 (±2.2)	46.2 (±0.7)	46.5 (±2.2)	45.5	4.1
	unif.	**54.3** (±0.5)	**52.2** (±0.4)	**39.1** (±1.0)	**45.8** (±0.9)	**46.6** (±1.2)	**46.4** (±1.8)	**48.3** (±0.8)	**47.3** (±0.9)	**48.0** (±1.2)	**45.0** (±2.3)	**44.1** (±1.2)	**42.9** (±2.1)	**46.7** (±0.0)	**46.3** (±0.0)	**47.2** (±0.7)	**46.7**	**3.5**

Table A7: $\text{XGLM}_{7.5\text{B}}$ in-language few-shot learning performance of on XNLI dev set using training sets of uniform class distribution and randomly sampled class distribution. We report the mean and standard deviation (in parentheses) of 5 different training sets sampled via different random seeds for each sampling strategy.

sampled[23] from it often has a much more skewed class distribution.

For a $|\mathcal{Y}|$-way classification task, a skewed training set distribution can cause the model to score the majority class as disproportionately more likely than the other classes. This was shown by Zhao et al. (2021) as the *majority label bias* problem. As a result, previous work such as Zhao and Schütze (2021) adopts a k-shot *per class* setting, where k unique examples are randomly drawn from each class to form a training set of size $k \times |\mathcal{Y}|$.

We compare learning from a uniform class distribution (randomly sampling k examples per class) to learning from a more skewed distribution (randomly sampling $k \times |\mathcal{L}|$ examples from the total population) on the XNLI task. We use the 24-shot and maximum fit (truncated 48-shot) settings. As shown in Table A7, for both settings, learning from a uniform class distribution leads to significantly higher accuracy in all languages compared to learning from the skewed distributions. *de*, *tr*, *bg*, *hi* suffer the most learning from the skewed distributions (> 2 absolute accuracy gap in the 24-shot setting), while *es* suffers the least. Moreover, the variances among few-shot trials using different random seeds shrink considerably when the training set class distribution is uniform. These results highlight the severeness of the *majority label bias* issue in the multilingual in-context learning framework.

D.5 Knowledge Probing

We evaluate to what extent our multilingual language model can effectively store factual knowledge in different languages. To this end, we evaluate knowledge triplet completion using the mLAMA dataset (Kassner et al., 2021), which was translated from the English benchmark LAMA (Petroni et al., 2019) using Google Translate. The data is from TREx (Elsahar et al., 2018) with triples of the format ⟨object, relation, subject⟩. Following the convention of LAMA, triples are converted to templates for querying the language model. For example, a triple like ⟨Paris, capital-of, France⟩ is converted to template "Paris is the capital of `[MASK]`". While each query in the original mLAMA dataset contains hundreds of candidates on average, we restrict it to three candidates one of which is the ground truth candidate and the other two candidates are randomly sampled to ensure fast inference and save API cost. Following the evaluation protocol of mLAMA, we report precision @1 averaged over all relations per language.

[23] We implement our random sampling procedure using the `numpy.random.choice` function: https://numpy.org/doc/stable/reference/random/generated/numpy.random.choice.html.

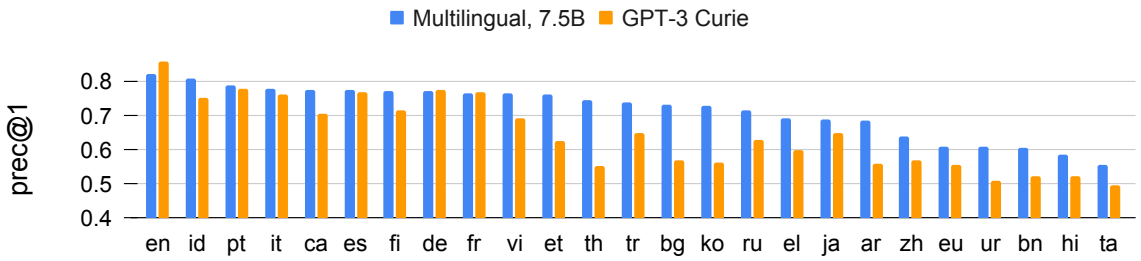


Figure A6: Knowledge probing on 25 languages. The performance of a random baseline is 0.33 since we down-sampled the candidate set of each query to contain three candidates.

We evaluate on the 25 languages covered in XGLM's pre-training data. We compare to the GPT-$3_{6.7B}$ model. As shown in Figure A6, both our multilingual model and GPT-3 Curie perform well on English. For non-English languages, our multilingual model maintains performance (above 0.6) while GPT-3 Curie drops drastically especially for medium and low resource languages. Overall, compared to an English-centric language model, our multilingual language model are better at retaining factual knowledge on a wider range of languages with $+7.1$ points on average.

E Safety and Bias Analysis

Given the centrality of large scale Language models, it is important to ensure such powerful models are used responsibly. Accordingly, we further examine XGLM's behavior on two tasks:

- **Hate speech detection:** A safety task to test language models' ability to identify hateful and offensive text;
- **Occupation Identification:** A bias task to study language models' performance disparity between different gender groups on the task of occupation identification.

Through extensive experiments, we have following findings: First, hate speech detection in an in-context learning setting is quite challenging. Moreover, language models are not effectively leveraging few-shot examples to improve the performance. Second, although language models have relatively good performance on the occupation identification task, they run the risk of exhibiting strong gender bias for certain occupations.

E.1 Hate Speech Detection

E.1.1 Setup

Datasets. We adopt datasets introduced by Huang et al. (2020) that include hate speech data from Twitter in five languages: English, Italian, Portuguese, Polish and Spanish. All hyperlinks, usernames and hashtags are replaced with generic symbols (`URL`, `USER`, `HASHTAG`) to anonymize user information. We remove tweets containing more than 2 generic symbols to encourage more informative examples. We further filter out tweets of length less than 5 tokens or more than 30 tokens. In the spirit of creating balanced data, we randomly sample 500 each positive (hate speech) negative (not hate speech) examples for each language. For further comparison, we translate non-English data into English by using Google Translate and then evaluate English models performance on the task.

Prompts. We evaluate two approaches to prompting, similar to Section **??**. For English prompts, we *prefix* "The sentence is <candidate>" to the input sentence to form a prompt. We consider 10 verbalization *candidates* including 5 negative (*normal., common., ok., usual., acceptable.*) corresponding to classification of *not hate speech* and 5 positive (*sexist., racist., offensive., abusive., hateful.*) representing classification of *hate speech*. For code-switched prompt, we translate the English prefix and candidates into the corresponding target language by using Google Translate. For example, "The sentence is normal" is translated into "Questa frase è normale." for Spanish. For few-shot learning, we randomly draw examples from the training data and report the average performance across 5 runs.

Model	language condition	# shot	English	Italian	Portuguese	Polish	Spanish
			Accuracy				
GPT-$3_{6.7B}$ *repl.*	code switched	0	54.5	54.3	57.0	51.3	52.5
		4	**55.5***	53.4	48.5	52.5	52.2
GPT-$3_{6.7B}$	code switched	0	**59.2**	**61.8**	55.8	58.7	**55.6**
		4	53.6	**54.8**	53.2	**54.5**	**53.7**
$XGLM_{7.5B}$	code switched	0	56.0	60.4	50.8	47.0	53.1
		4	50.4	50.7	**56.8**	51.0	50.7
GPT-$3_{6.7B}$ *repl.*	same language	0	-	50.5	**60.2**	50.1	52.4
		4	-	51.0	47.2	50.9	50.7
XGLM $_{7.5B}$	same language	0	-	57.5	41.8	50.0	53.1
		4	-	53.2	56.5	51.1	52.7
GPT-$3_{6.7B}$ *repl.*	English	0	-	55.8	57.5	**62.8***	55.0
		4	-	52.5	49.2	53.9	52.8
			Recall				
GPT-$3_{6.7B}$ *repl.*	code switched	0	57.0	90.2	67.0	21.6	77.0
		4	**73.0***	**76.7***	**72.5**	65.4	**77.1***
GPT-$3_{6.7B}$	code switched	0	62.4	88.6	51.5	49.0	76.4
		4	65.7	69.2	59.6	58.5	61.1
$XGLM_{7.5B}$	code switched	0	**77.2***	**95.4***	**80.4**	53.8	**95.8***
		4	14.9	18.8	18.8	15.8	19.5
GPT-$3_{6.7B}$ *repl.*	same language	0	-	1.4	9.1	0.2	11.6
		4	-	50.0	66.1	**77.8**	33.4
$XGLM_{7.5B}$	same language	0	-	87.4	39.5	0.0	79.4
		4	-	39.0	24.9	55.8	44.6
GPT-$3_{6.7B}$ *repl.*	English	0	-	93.8	77.8	**74.8***	77.6
		4	-	72.4	71.2	75.7	73.7

Table A8: Accuracy and recall scores of our multilingual model and other English models on the Hate Speech Detection task. We evaluate five target languages. For each target language, we bold the highest number for zero-shot and four-shot respectively. * indicates the number is significantly higher than others. For language condition, we consider three cases: "code switched" means the prefix, candidates are in English and tweets are in the target language; "same language" means prefix, candidates and tweets are in the target language; "English" means prefix, candidates and tweets are in English, i.e. note that, "same language" and "English" reduce to the same experimental condition when the target language is English.

Metrics. We compute precision, recall and accuracy for all experimental conditions. Since the test data is balanced, the accuracy of a random baseline is 50%.

E.1.2 Results

We show accuracy and recall scores in Table A8. Bolded results are the highest in the table and those with an (*) are statistically significantly better than other comparable conditions. Hate speech detection is a challenging task for all models. We observe that across the five languages, in-context learning results are only slightly better than random (50%). The results are also unstable and sensitive to prompt changing. Overall, the $XGLM_{7.5B}$ model has better recall compared to the English-centric models. For example, the $XGLM_{6.7B}$ En-only model has very low recall score in the zero-shot setting with the language condition set as "same language", indicating that it blindly predicts almost everything as negative (not hate speech). Another interesting observation is that most few-shot results are worse than zero-shot, which indicates that with the prefix described above, language models are not able to utilize examples. Interestingly, we also find that in one-shot experiments models tend to copy the label of the given example instead of predicting based on the input tweet. This further demonstrates that language models are struggling with learning from few-shot examples in this task.

Model	English		Spanish		French	
	Avg.↑	|Diff|↓	Avg.↑	|Diff|↓	Avg.↑	|Diff|↓
XGLM $_{6.7B}$ *En*-only	**90.73**	3.19	**91.23**	**2.65**	83.46	4.85
GPT-3$_{6.7B}$	90.42	3.53	86.91	5.18	**90.85**	**2.30**
XGLM $_{7.5B}$	86.49	**2.83**	82.93	4.28	76.55	2.95
XGLM $_{6.7B}$ *En*-only + translate	-	3.19	91.18	3.75	90.07	2.35

Table A9: Accuracy and bias scores of our multilingual model and other English models on the occupation identification task. "|Diff|" stands for the average absolute accuracy gap between male and female groups aggregated across all occupations. We bold the highest accuracy score for each language.

E.2 Gender Bias in Occupation Identification

E.2.1 Setup

Datasets We use the English bio dataset introduced in (De-Arteaga et al., 2019) to study gender bias based on identifying a person's occupation from their bios. For multilingual bio datasets we use those created by (Zhao et al., 2020). Originally there are 28 occupations in English, 69 occupations in Spanish and 27 occupations in French. To ensure we have plenty of test data for each occupation, we only keep occupations with at least 1000 male examples and 1000 female examples. This leads to 16 occupations in English, 6 occupations in Spanish and 4 occupations in French. We follow the setup in (Zhao et al., 2020) where people's names and pronouns are removed from the bios. We then prefix "The occupation of this person is <candidate>" to the input bio to form a prompt. The candidate set consists of five occupations, including the ground truth one and four other randomly sampled male and female occupations (two male and two female). Male (female) occupations refer to ones having predominantly more male (female) samples.

Metrics Similar to the metric for Hate Speech detection, we first obtain the scores for 5 candidates and consider a prediction correct if the ground truth candidate yields the highest score among five candidates. We then compute the bias score as the absolute gap between the accuracy scores on the male and female samples,[24] averaged across all occupations. A lower bias score indicates that a model has less divergence in identifying occupations for men and women.

E.2.2 Results

We present the overall accuracy scores and the bias scores (|Diff|) in Table A9. Results indicate that the XGLM $_{6.7B}$ En-only model achieves the best performance on English and Spanish, while the GPT-3$_{6.7B}$ model achieves the best performance on French. XGLM $_{7.5B}$ model, instead, falls behind on all three languages, especially for Spanish and French. We think this is potentially due to that all pronouns and people's names are removed from the test data but not training data. The training data for XGLM $_{7.5B}$ contains more Spanish and French compared to the other two models. Thus, XGLM $_{7.5B}$ may have more severe morphological mismatch on Spanish and English. Regarding the bias score, the GPT-3$_{6.7B}$ model is the most biased model on both English and Spanish but least biased on French. XGLM $_{6.7B}$ En-only and XGLM $_{7.5B}$ exhibit the least bias on Spanish and English, respectively.

F Data Card

We follow the recommendations of Gebru et al. (2018) and provide a datacard for the dataset used to train XGLM, which is a subset of CC100-XL, a larger multilingual dataset we curated.

F.1 Data Sources

Following the recent success of multilingual self-supervised pre-training (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020; Xue et al., 2020; Goyal et al., 2021a; Liu et al., 2020), we train our language models on a mixture of monolingual text of different languages. We extend the pipeline used for mining the CC100 corpus (Conneau et al., 2020; Wenzek et al., 2020) to generate CC100-XL,

[24] We only consider gaps that are statistically significant.

a significantly larger multilingual dataset covering 68 Common Crawl (CC) snapshots (from Summer 2013 to March/April 2020) and 134 languages. As the first step to balance the language distribution, we sampled 30% of the data from the languages that contain more than 15 billion tokens and more than 20 million documents. This resulted in a 8.4 TB multilingual corpus with 1.9 trillion tokens.

F.2 Motivation

- **For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.** The CC100-XL dataset was collected to create a high quality monolingual dataset for at least 100 languages. It was mainly used for training foundation multilingual language models which may be applied to a broad list of language tasks, including neural machine translation, speech translation, question answering, etc. CC100-XL involves sentence level filtering, preserves context, improves the filtering mechanism, and paves a way for mining 200+ languages.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** Meta AI.

- **Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.** Meta AI.

- **Any other comments?** No.

F.3 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.** The instances are textual documents sampled from Commoncrawl snapshots.

- **How many instances are there in total (of each type, if appropriate)?** The training dataset of XGLM contains 1.74 billion documents in total.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).** The dataset is a subset of CC100-XL. For each language, the data is either a full set or a random subset of CC100-XL data. Especially, the medium- and low-resource languages are upsampled. In terms of language representation, the CC100-XL dataset contains 134 languages extracted using fasttext[25] from Common Crawl snapshots. We further selected a subset of 30 languages to train XGLM, taking geo-location, language family and typology diversity of the languages into account.

- **What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.** Each instance consists of raw text data.

- **Is there a label or target associated with each instance? If so, please provide a description.** No.

- **Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.** No.

[25] https://fasttext.cc/docs/en/language-identification.html

ISO code	Language	Tokens (M)	Size (GiB)	ISO code	Language	Tokens (M)	Size (GiB)
en	English	803,527	3,324.45	-	Arabic Romanized	685	1.65
ru	Russian	147,792	1,007.38	mn	Mongolian	681	4.26
zh	Chinese	132,770	485.32	la	Latin	635	2.20
de	German	89,224	369.30	ne	Nepali	600	5.32
es	Spanish	87,303	363.83	si	Sinhalese	524	3.96
fr	French	77,420	303.76	mr	Marathi	458	3.59
ja	Japanese	66,054	293.39	kn	Kannada	446	3.41
it	Italian	41,930	170.76	so	Somali	436	1.56
pt	Portuguese	36,586	147.12	cy	Welsh	398	1.27
el	Greek	28,762	180.37	jv	Javanese	389	1.23
ro	Romanian	24,176	93.63	ps	Pashto	387	1.97
uk	Ukrainian	23,723	156.68	uz	Uzbek	332	1.64
hu	Hungarian	22,718	89.87	gu	Gujarati	327	2.10
ko	Korean	20,002	79.08	km	Khmer	272	2.14
pl	Polish	19,293	73.59	-	Urdu Romanized	245	0.73
no	Norwegian	17,600	70.89	am	Amharic	169	0.85
nl	Dutch	17,163	68.36	-	Bengali Romanized	166	0.48
fi	Finnish	16,804	67.28	pa	Punjabi	153	0.93
da	Danish	16,274	64.74	gl	Galician	137	0.50
id	Indonesian	15,424	67.51	ha	Hausa	124	0.42
hr	Croatian	14,455	54.27	mg	Malagasy	116	0.38
tr	Turkish	12,413	51.51	sa	Sanskrit	107	0.42
ar	Arabic	12,249	64.34	eu	Basque	105	0.35
vi	Vietnamese	11,199	50.45	my	Burmese	101	0.74
th	Thai	10,842	99.86	su	Sundanese	91	0.30
bg	Bulgarian	9,704	61.10	or	Oriya	91	0.62
fa	Persian	9,355	57.46	ht	Haitian	87	0.28
sv	Swedish	9,169	36.54	lo	Lao	84	0.59
ms	Malay	9,106	38.57	ky	Kyrgyz	70	0.34
he	Hebrew	8,637	42.13	br	Breton	57	0.16
cs	Czech	8,616	32.46	ga	Irish	49	0.15
sk	Slovak	8,251	30.70	yo	Yoruba	48	0.14
ca	Catalan	7,076	26.90	eo	Esperanto	47	0.14
lt	Lithuanian	4,847	18.38	-	Tamil Romanized	40	0.13
sl	Slovene	4,029	14.97	zu	Zulu	40	0.14
hi	Hindi	3,448	26.63	ti	Tigrinya	40	0.19
et	Estonian	3,287	12.18	-	Telugu Romanized	37	0.11
lv	Latvian	2,815	10.67	ku	Kurdish	36	0.10
tl	Tagalog	2,389	8.13	om	Oromo	27	0.09
sq	Albanian	2,382	8.76	xh	Xhosa	26	0.09
sr	Serbian	2,101	12.68	gd	Scottish Gaelic	19	0.05
-	Hindi Romanized	2,045	6.60	ig	Igbo	18	0.06
az	Azerbaijani	1,904	8.41	as	Assamese	17	0.10
bn	Bengali	1,627	11.19	lg	Ganda	15	0.05
ta	Tamil	1,477	12.36	wo	Wolof	14	0.03
ur	Urdu	1,352	7.77	fy	Western Frisian	12	0.04
kk	Kazakh	1,278	8.40	tn	Tswana	11	0.03
hy	Armenian	1,261	7.16	ff	Fula	11	0.03
ka	Georgian	1,261	10.48	gn	Guaraní	10	0.03
is	Icelandic	1,163	4.21	sd	Sindhi	8	0.04
be	Belarusian	1,004	5.81	ln	Lingala	7	0.02
bs	Bosnian	950	4.00	bm	Bambara	6	0.02
ml	Malayalam	935	8.08	iu	Inuktitut	6	0.03
mk	Macedonian	927	6.05	kg	Kongo	4	0.01
sw	Swahili	908	3.19	qu	Quechua	3	0.01
af	Afrikaans	819	3.04	ss	Swati	2	0.01
te	Telugu	689	5.28	-	Unassigned	503	2.30

Table A10: Languages and statistics of the training data set selected from CC100 XL corpus.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.** A small percentage of document instances (<2%) are duplicated. Other than that, there are no relationships between individual instances.

- **Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.** This dataset is split into training and validation only. For each high resource language, at least 5,000 randomly selected documents and 30,000 lines were split into validation set, and the rest documents are for training; for low-resource languages, at least 100 randomly selected documents and 1,000 lines (a couple of very low resource languages contain 80 documents) were split into valid set and leave the rest for training. There are 3.5 million lines of text in total for the validation set. This split is mainly to ensure a good size of validation data with the coverage and balance over all languages, meanwhile, the validation size is not too large to affect the overall training speed.

- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** 10% of Russian sample were lost during internal data transferring. Therefore, we ended up taking 26.7% random subset of the whole Russian data from CC100-XL.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** It's self-contained.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.** CC100-XL is exclusively extracted from Common Crawl; and the information in it is not considered confidential.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.** CC100-XL is a subset of public Common Crawl data, which could contain sentences that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety.

- **Does the dataset relate to people? If not, you may skip the remaining questions in this section.** Some documents of this data relate to people, such as news articles, Wikipedia descriptions, etc.

- **Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.** No.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how** Other than the individuals who are celebrities, politicians, etc, and have their Wikipedia pages; it is possible to identify other individuals by their names, twitter account names, etc. But we built personally identifiable information (PII) identification tools following the guidelines of General Data Protection Regulation (GDPR) and National Institute of Standards and Technology (NIST) and run against this dataset, we did not found highly sensitive PII information, such as U.S. social security number, login credentials, etc.

- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.** We use a curated special word list of 100 languages which covers profanities, hate speech, bulling language, common slangs and profane multi-word expressions (MWEs) to tag paragraphs and remove the docs containing them. Given the size of this data, it could still contain

such sensitive information (as the above lists may not be exhaustive) but should be a very small percent of instances.

- **Any other comments?** No

F.4 Collection Process

- **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/ derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.** Please refer to the main document for details.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?** Please refer to the main document for details.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** Please refer to the main document for details.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** This data is mined, filtered and sampled by machines.
- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.** The data was collected from 68 Common Crawl (CC) snapshots (from Summer 2013 to March/April 2020). Therefore, it does not contain a lot of information about recent events such as COVID-19.
- **Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.** No.
- **Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.** No.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?** N/A
- **Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.** N/A
- **Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.** N/A
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).** N/A
- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.** Some responsible AI related evaluations were performed. Please refer to the main document.

- **Any other comments?** No

F.5 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.** Yes, the detailed steps are as below:
  - *Downloading and Sharding Commoncrawl Snapshots* We downloaded 68 Commoncrawl snapshots and divided the data in 240 shards based on web-domain. At this stage, textual data gets extracted from the WET files provided by Common Crawl which involves cleaning excessive tabs and newlines.
  - *Language Identification (LID) at Document Level* For this stage, we used the fastText language identification (LID) model on the entire document which helped further divide the data by language. In addition to the original languages supported by fastText, we also added support for 28 romanized languages. In total, the data for each language contains 240 shards.
  - *Deduplicating Documents based on URL* We aggregated the data based on URL which yields 60% reduction in volume. In case two documents had the same URL, we selected the document having more recent text content.
  - *Document Splitting and LID at Paragraph Level* We segmented the documents based on newline and also stored the information about the order in which the paragraphs were appearing in the original document (i.e. seq_num). Next, we performed LID at the paragraph level again in order to divide the original documents into clusters of paragraphs where each cluster represents sentences belonging to a particular language.
  - *Deduplicating Paragraphs* Data extracted from Commoncrawl snapshots still have a lot of duplicate text even if the document is different. In order to tackle this, we applied the normalization function from CCNet (Wenzek et al., 2020) and then computed a SHA-1 hash of the normalized text. This helped in reducing the content by 88%. Choosing which <paragraph, url> combination to keep can be tricky as it can lead to a lot of fragmented documents. So we devised a strategy to choose documents based on sorted <url, seq_num> order which would help in preventing fragmentation as much as possible.
  - *Language Model Scores* We scored every paragraph using a Language Model trained on data collected from OPUS (Tiedemann, 2012) (monolingual data collected from the availble bitexts) using a 4-gram KenLM (Heafield, 2011). Note that since the LMs were not trained on data belonging to a specific domain, this feature helped in eliminating general non-fluent sentences.
  - *Heuristic based approaches* We use the following techniques to further refine the filtering step (especially useful for Low resource languages having no or poor quality LM)
    - Ratio of digit+punctuation to total characters (current threshold <0.25)
    - Maximum number of URLs per sentence (current value 1)
    - Type-token ratio (current threshold >0.6 + removing bottom 1% per language)
    - Minimum number of tokens per sentence (current value 3; not applied for agglutinative languages)
  - *Tagging profane words and removing instances containing such words*
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.** The "raw" data is publiclly available in in https://commoncrawl.org/the-data.
- **Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.** The software is proprietary to Meta Platforms and currently unavailable publicly.
- **Any other comments?** No

F.6 Uses

- **Has the dataset been used for any tasks already? If so, please provide a description.** Yes, this dataset and its precursor CC100 data have been used to train machine translations and multilingual language models, which are foundation to many downstream language tasks.
- **Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.** No.
- **What (other) tasks could the dataset be used for?** This data can be used to pretrain multilingual language models, which are foundation to many current and future language tasks.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?** The pipeline for creating this dataset paves a way for building a scalable infrastructure for mining datasets to be be used for training large-scale models.
- **Are there tasks for which the dataset should not be used? If so, please provide a description.** No.
- **Any other comments?** No.

F.7 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.** No.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?** N/A
- **When will the dataset be distributed?** No.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.** No.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.** No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.** N/A
- **Any other comments?** No.

F.8 Maintenance

- **Who is supporting/hosting/maintaining the dataset?** Meta AI.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** Refer to the main document.

- **Is there an erratum? If so, please provide a link or other access point.** Currently no.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?** No plan for updating.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.** N/A
- **Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.** N/A
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/ distributing these contributions to other users? If so, please provide a description.** No.
- **Any other comments?** No.