Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM

Rachel Bawden, François Yvon

Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM

Rachel Bawden
Inria, Paris, France
rachel.bawden@inria.fr

François Yvon
Université Paris-Saclay, CNRS, LISN
francois.yvon@cnrs.fr

Abstract

The NLP community recently saw the release of a new large open-access multilingual language model, BLOOM (BigScience et al., 2022) covering 46 languages. We focus on BLOOM's multilingual ability by evaluating its machine translation performance across several datasets (WMT, Flores-101 and DiaBLa) and language pairs (high- and low-resourced). Our results show that 0-shot performance suffers from overgeneration and generating in the wrong language, but this is greatly improved in the few-shot setting, with very good results for a number of language pairs. We study several aspects including prompt design, model sizes, cross-lingual transfer and the use of discursive context.

1 Introduction

Large language models (LLMs) trained at scale with simple objectives have been found to achieve results that match dedicated systems on numerous NLP tasks (Radford et al., 2019), as long as tasks are formulated as text generation though "prompting" (Liu et al., 2023). LLMs multi-task performance can even be improved with "instruction" fine-tuning (Sanh et al., 2022; Muennighoff et al., 2022), few-shot priming, and better strategies to select or learn prompts (Petroni et al., 2019; Shin et al., 2020; Schick and Schütze, 2021; Lester et al., 2021; Wei et al., 2022). In multilingual settings, their performance on machine translation (MT) tasks, as measured by automatic scores, is often close to state of the art, even when mostly trained on monolingual data (Brown et al., 2020). Moreover, prompting-based MT offers the prospect of better control of outputs, e.g. in terms of quality, style and dialect (Garcia and Firat, 2022). However, these abilities remain poorly understood, as LLM analyses primarily focus on their multitask rather than multilingual ability (see however (Vilar et al., 2022; Zhang et al., 2023; Moslem et al., 2023), which we discuss in Section 2).

In this work, we focus on the MT performance of BLOOM (BigScience et al., 2022), a (family of) open-access multilingual LLM(s), designed and trained by the collaborative BigScience project.[1] Our main aims are to (i) evaluate BLOOM's zero- and multi-shot behaviour, (ii) study the effect of prompt design, (iii) evaluate a diverse set of language pairs and (iv) assess its ability to use linguistic context. Our main conclusions, which extend those in (BigScience et al., 2022), are (i) 0-shot ability is blighted by overgeneration and generating in the wrong language, (ii) using few-shot improves both issues, with results much closer to state of the art across datasets and language pairs, (iii) there are clear transfer effects, with high scores for languages not officially seen in training, and successful transfer across language pairs via few-shot examples and (iv) although linguistic context does not lead to higher scores, there is evidence that BLOOM's translations are influenced by it. We release our code and translation outputs.[2]

2 Related work

Since the early attempts at using language models (LMs) as multi-task learners (McCann et al., 2018), MT has been a task of choice to gauge LMs' multilingual ability. Results for the zero- and few-shot ability of LMs were discussed for both GPT-2 and GPT-3 (Radford et al., 2019; Brown et al., 2020), which is especially intriguing as they were trained primarily on monolingual (English) data. These results have since been confirmed for other monolingual LMs such as T5 (Raffel et al., 2020a) and multilingual LMs such as XGLM (Lin et al., 2022), PALM (Chowdhery

[1] https://huggingface.co/bigscience/bloom

[2] https://github.com/rbawden/mt-bigscience

et al., 2022), and ALEXATM (Soltan et al., 2022). However, the focus has mainly been on global multi-task performance; often only a small part of the discussion is devoted to MT. Moreover, results are often only reported for a few well-resourced language pairs (e.g. English-French and English-German), and the scores reported (mostly BLEU), are hard to compare due to a non-systematic use of standardised evaluation protocols and metrics.[3]

There are however some in-depth analyses of MT performance of LLMs, each focusing on a specific LM's performance in a true multilingual setting with respect to prompt design and number of few-shots. For instance, Vilar et al. (2022) reevaluate the MT performance of the multilingual PALM (Chowdhery et al., 2022), focusing notably on the selection of few-shot examples. Consistent with our findings, they determine that prompt choice becomes unimportant in few-shot settings and that using few-shot examples increases performance with diminishing returns for $k > 5$ examples, using BLEURT and BLEU scores, as well as the results of a human evaluation, They find that the quality of few-shot examples has a large impact on performance. However, even with good prompts, PALM lags a couple of points behind state-of-the-art MT systems, especially when translating from English, notable due to adequacy problems. Zhang et al. (2023) focus on the evaluation of GLM-130B, a bilingual (Chinese and English) LLM (Zeng et al., 2022). Their main conclusions are also consistent with ours: (a) zero-shot performance varies greatly across different prompts; (b) increasing the number of prompts from 0 to 20 yields consistent improvements in performance, again with variance across instructions; and (c) finding the best few-shot example selection policy is difficult. It seems that having good and long examples, for instance, may help, even though none of the criteria explored in this study seem to provide any systematic improvement. A last point worth mentioning is that prompting with monolingual data hurts performance, but that using pseudo-parallel data obtained with back-translation (Bojar and Tamchyna, 2011) is an effective workaround.

Moslem et al. (2023) evaluate OpenAI's GPT-3 (Brown et al., 2020)[4] with sampling-based decoding and a prompt resembling our own `xglm-source+target` prompt. They report strong zero-shot behaviour using multiple metrics, plus clear improvements with an increased number of shots for the well-resourced languages, less so for the only low-resource language in their lot (Kinyarwanda). The main novelty of this study is to use prompting as a vehicle to perform local adaptation or to ensure terminological consistency. For this, they use fuzzy matches from a translation memory as well as MT outputs to build their prompts, yielding results that both outperform their zero-shot system, but also their initial MT engine. Additionally inserting terms and their translation in the instruction yields supplementary improvements.

Finally note the preliminary evaluation of CHATGPT in (Jiao et al., 2023), which reports interesting insights regarding the multilingual abilities of this model, as well as proposing innovative techniques to generate (artificial) prompts and to use pivoting in prompting. Similar to ours, this study considers multiple test domains such as news (WMT) and Wikipedia (Flores). A more in-depth analysis of the same model is in (Hendy et al., 2023), which confirms the strong translation abilities of CHATGPT, at least for "well-resourced"[5] language pairs. Document-level evaluations are also reported, as well as human evaluations and qualitative analyses.

Multilingual MT is also the subject of dedicated (monotask) architectures and training regimes. Originally introduced in (Dong et al., 2015; Firat et al., 2016; Luong et al., 2016) with a limited language coverage, the latest versions of these approaches are able to handle hundreds and languages, including very low-resource language pairs (Fan et al., 2021; Bapna et al., 2022; Costa-jussà et al., 2022). Although we found thatBLOOM is able to match this performance, given sufficient training material, we also see that it still lags behind for many languages pairs that are under-represented in its training data.

3 BLOOM Language Model

BLOOM is a large open-access multilingual model trained on 46 natural languages developed within

[3] See the discussion at `http://blog.benjaminmarie.com/2/comparing-uncomparable.html` of these differences, and an attempt to reconstruct consistent scores.

[4] Version: (`text-davinci-003model`.)

[5] A rather slippery concept in this context, as the content of the training data is not fully known, and seems to mostly comprise English texts.

Prompt name	Prompt	Target
a_good_translation	Given the following source text (in L1): [source sentence], a good L2 translation is:	[target sentence]
version	If the original version says [source sentence] then the L2 version should say:	[target sentence]
gpt3	What is the L2 translation of the sentence: [source sentence]?	[target sentence]
xglm	(L1:) [source sentence] = L2:	[target sentence]
translate_as	[source sentence] translates into L2 as:	[target sentence]

Table 1: MT prompts for the WMT'14 dataset (Bojar et al., 2014). All prompts specify the target language (L2). Each prompt exists in a 'target-only' version (-target), where only the target language is specified, and two prompts also exist in a -source+target version, where the source language (in red and in brackets) is explicit in the instruction.

the BigScience project (BigScience et al., 2022). It is an auto-regressive language model designed to generate text to complete a user-entered text prefix, known as a prompt. It can be used for multiple tasks, including MT, question answering, etc. BLOOM was trained on 1.6TB of text (of which 30% English), from various sources, although 38% of the data, known as the ROOTS corpus (Laurençon et al., 2022),[6] is from Oscar web data (Ortiz Suárez et al., 2019). The model is openly released on HuggingFace in multiple sizes, ranging from 560M to 176B parameters.[7]

4 Evaluating BLOOM on the MT task

4.1 MT Datasets Used

We experiment with three datasets, chosen to test different aspects of BLOOM for MT: WMT (Bojar et al., 2014), Flores-101 (Goyal et al., 2022) and DiaBLa (Bawden et al., 2021). We use the WMT 2014 news test sets for English↔French and English↔Hindi, which we take as representative high- and lower-resource language pairs with respect to BLOOM's training data.[8] These test sets are somewhat outdated (Garcia et al., 2023), but have been used repeatedly in past LLM evaluations and are included as standard benchmarks for comparison. Flores-101 is a multi-parallel dataset in 101 languages, translated from original English sentences.[9] In fact, evaluations into English are bound to yield overly good results (e.g. (Toral et al., 2018)) and between other languages may mostly reflect their similarity with the original English. We use it to test and compare BLOOM's multilinguality, including for low-resource languages.[10] DiaBLa is a bilingual test set of spontaneous written dialogues between English and French speakers, mediated by MT. We use this as a test of MT in an informal domain and the impact of (cross-lingual) linguistic context in MT.

4.2 Experimental setup

We evaluate and compare BLOOM (and its variants) using the Language Model Evaluation Harness (Gao et al., 2021) in 0-shot and few-shot settings. For few-shot, k examples are prefixed to the prompt and separated with ### as shown in Example 1 (1-shot example is underlined).

(1) **Input**: French: je m'ennuie = English: I'm bored. ### English: Is that your dog that's just wandered in over there? = French:
Reference: Est-ce que c'est votre chien qui vient de rentrer par là ?

Results are reported on the datasets' test splits. Few-shot examples are randomly taken from the data splits according to availability (train for WMT, dev for Flores-101 and test for DiaBLa). We evaluate using BLEU (Papineni et al., 2002) as implemented in SacreBLEU (Post, 2018), using as tokenisation 13a for WMT and DiaBLa and spm for Flores-101 as recommended (Costa-jussà et al., 2022).[11] BLEU has many shortcomings, but is good enough to provide quantitative comparisons for most systems used in this study. We additionally use COMET (Rei et al., 2020) for finer grained comparisons when the scores are closer.

[6] The ROOTS corpus can now be queried using the dedicated search tool https://huggingface.co/spaces/bigscience-data/roots-search.

[7] https://huggingface.co/bigscience/bloom

[8] English, French and Hindi make up 30%, 12.9% and 0.7% of the training data respectively (Laurençon et al., 2022).

[9] Given the way the corpus was generated, it is highly biased and should be used to evaluate translation *out of English*.

[10] An extended version, Flores-200, has been recently released (Costa-jussà et al., 2022), which is larger and covers approximately twice as many languages. As this new version was released late in our evaluation process and had only been used in one paper, we decided to stick to Flores-101.

[11] BLEU+case:mixed+smooth.exp+{13a,spm}+version.2.2.1

4.2.1 Comparative models

In our cross-dataset comparison (Section 5.1), we compare BLOOM to other LLMs: (i) two task-fine-tuned models: T0[12] (Sanh et al., 2022), trained on English texts, and mT0-xxl[13] (Muennighoff et al., 2022), the multilingual version, and (ii) OPT[14] (Zhang et al., 2022), an English generative LM. We evaluate all models on the same prompt `xglm-source+target`. To evaluate multiple language pairs with Flores-101, we compare (as a topline) to the supervised 615M-parameter MT model M2M-100 (Fan et al., 2021), using the scores computed by Goyal et al. (2022).

4.2.2 Prompts

We use several prompts, designed to illustrate different sources of variation, namely (i) the inclusion (or not) of the source language name, (ii) the relative order of source and target language names, (iii) the position of the source sentence (beginning or end of the prompt) and (iv) the prompt's verbosity. These prompts, available in PromptSource (Bach et al., 2022), are shown in Table 1. The first three are inspired by previous work:[15] (Brown et al., 2020) for `gpt3`,[16] (Lin et al., 2022) for `xglm` and (Wei et al., 2022) for `translate_as`, which also resembles Raffel et al. (2020b)'s prompt (*Translate English to German: "[source text]": [target sentence]*), also used in (Wei et al., 2022; Garcia and Firat, 2022).

Considering the entries in Table 1, we can see that "prompting" in fact refers to two distinct aspects of the input: (i) the formulation of the task in natural language and (ii) the presentation of related examples (for few-shot setups) interleaved with language tags (perhaps more clearly referred to as *priming* by Pham et al. (2020)); as illustrated by the `xglm` prompt for example, the instruction part can reduced to one single word. As our results below suggest, the instruction mostly matters in 0-shot setups, but can almost be dispensed with in few-shot scenarios. The authors of (Brown et al., 2020) and (Hendy et al., 2023) also use a verbose, instruction-like prompt in their zero-shot setup, and a much more compact one for few shots experiments. Also note that InstructGPT's prompt combines both an instruction and language tags (Ouyang et al., 2022, p. 49).

[12] https://huggingface.co/bigscience/T0

[13] https://huggingface.co/bigscience/mt0-xxl

[14] https://huggingface.co/facebook/opt-66b

[15] This was not always straightforward due to incomplete documentation concerning (a) prompts tested, and (b) those actually used in each experiment (e.g. different ones for 0-shot and few-shot runs (Chowdhery et al., 2022)).

[16] Used only it seems, for zero-shot learning in the form "Q: what is the L2 translation of sentence [source sentence]. A:", where special tokens Q and A are the query and the answer texts (cf. Figure G.36, pp 59).

5 Evaluation results

Our evaluation of BLOOM starts with a comparison across the three datasets and detection of major MT errors with a focus on WMT (Section 5.1) and then we present more in-depth analyses of particular aspects: (i) using WMT, a comparative study of BLOOM model sizes (Section 5.2) and prompts (Section 5.3), (ii) using Flores-101 an evaluation of more language pairs and cross-lingual few-shot transfer (Section 5.4), and (ii) using DiaBLa, a study of the use of linguistic context (Section 5.5).

5.1 Comparison across datasets

	0-shot				1-shot			
	BLOOM	T0	mT0-xxl	OPT	BLOOM	T0	mT0-xxl	OPT
WMT 2014								
en→fr	14.91	1.21	29.27	12.95	27.83	1.41	25.24	21.92
fr→en	15.52	25.79	32.88	15.54	34.61	21.01	30.03	24.55
en→hi	6.80	0.16	11.20	0.14	13.62	0.12	9.50	0.08
hi→en	12.05	0.00	26.13	0.42	25.04	0.01	20.15	0.58
DiaBLa								
en→fr	0.88	0.52	28.44	0.53	5.70	0.61	21.03	15.52
fr→en	0.85	25.51	34.96	0.83	12.05	20.57	26.88	12.05
Flores-101								
en→fr	2.77	1.86	55.45	2.76	44.99	2.13	53.53	24.36
fr→en	2.73	31.90	60.10	2.59	45.59	24.86	58.22	16.74
en→hi	1.29	0.15	67.69	0.07	27.25	0.06	54.66	0.12
hi→en	3.40	0.00	59.55	0.10	35.06	0.19	57.32	0.45

(a) Original predictions

	0-shot				1-shot			
	BLOOM	T0	mT0-xxl	OPT	BLOOM	T0	mT0-xxl	OPT
WMT 2014								
en→fr	32.25	1.21	29.24	18.86	36.29	1.41	25.19	22.31
fr→en	37.16	25.80	32.87	33.18	38.18	21.07	29.95	33.25
en→hi	12.10	0.16	11.20	0.11	15.73	0.12	9.50	0.08
hi→en	24.29	0.00	26.06	0.51	25.04	0.01	20.06	0.61
DiaBLa								
en→fr	24.23	0.52	28.44	17.42	37.57	0.61	21.89	20.71
fr→en	22.94	25.51	34.92	36.80	41.36	21.09	27.20	37.63
Flores-101								
en→fr	26.91	1.85	55.34	21.40	49.32	2.13	53.40	28.41
fr→en	40.28	31.90	60.01	39.41	47.24	25.20	58.24	39.82
en→hi	7.74	0.15	67.69	0.12	29.52	0.06	54.66	0.12
hi→en	30.19	0.00	59.55	0.23	35.06	0.19	57.27	0.50

(b) Truncated predictions

Table 2: Comparison of BLEU scores (spBLEU for Flores-101) across the three datasets using the `xglm-source+target` prompt.

We first prompt BLOOM and the comparative models using the same prompt across datasets, restricting the directions tested to en↔fr and to en↔hi. We choose to systematically use

the `xglm-source+target` prompt (Table 1), which corresponds to this template:

(2) `L1: [source sentence] = L2:`

where `L1` and `L2` refer to the source and target languages respectively (e.g. English and French for en→fr) and `[source sentence]` is replaced by a given source sentence.

BLEU score results are in Table 2a for both 0-shot and 1-shot. There are issues for 0-shot MT for all directions, particularly when translating into a non-English language, (BLEU scores are systematically poorer than into English). Even into English, the scores remain low with respect to state of the art (e.g. 2.73 BLEU for Flores-101 fr→en BLOOM vs. 60.10 for MT0-XXL). MT0-XXL's good results however illustrate the benefits of prompt fine-tuning. Results are considerably improved in the 1-shot setting (except for T0 and MT0-XXL, both fine-tuned) and are comparable to other LLM MT results reported in the literature; they are markedly better for BLOOM than for OPT, which was mostly trained on English data. These general observations apply for WMT and Flores-101, while issues remain for DiaBLa.

What are the main issues? We identify two main issues (particularly for 0-shot): (i) overgeneration and (ii) generating in the wrong language (an obvious problem for MT). Examples 3 and 4 illustrate some instances of these problems. The first example shows a very common problem, with repetition of the prompt in multiple languages until the maximum generation length is reached.

(3) **Input:** French: Il dit avoir conçu une sonnette de porte Wi-Fi. = English:
BLOOM: He says he conceived a Wi-Fi doorbell. = Spanish: Dice que concibió un timbre de puerta Wi-Fi. = German: Er sagt, er hat einen WLAN-Türglocken gegründet. = Italian: Dice di aver concepito un chiuso Wi-Fi [. . .]

(4) **Input:** French: Le tigre fait partie de la même famille (genre Panthera) que les lions, les léopards et les jaguars. Ces quatre félins sont les seuls capables de rugir. = English:
BLOOM: The tiger is a member of the same family (genus Panthera) as lions, leopards and jaguars. These four cats are the only ones capable of roaring. Le tigre est un grand félin qui vit en Asie. Il est le plus grand des félins. Il est aussi le plus gros carnivore terrestre. Il est très rapide et peut courir à plus de 60 km/h. [. . .]

Separating MT quality from overgeneration Overgeneration as seen in Example 3 is a separate issue from BLOOM's capacity to translate into another language. We therefore devise a custom truncating method for this type of overgeneration such that only the first translation in a prediction is kept, i.e. anything after a first newline or the regular expression pattern `= .+?:` is discarded.

Results after truncation (Table 2b) show that for all three datasets, 0-shot and 1-shot scores are significantly improved (e.g. 1-shot DiaBLa fr→en increases from 12.05 to 41.36 and 0-shot Flores-101 hi→en increases from 3.40 to 30.19). BLOOM is capable of performing good MT but has a problem knowing when to stop generating. We use the same truncation elsewhere too and indicate when we show results for original or truncated outputs.

	en→fr		fr→en		en→hi		hi→en	
Lang. / Shot #	0	1	0	1	0	1	0	1
Target	2814	2959	2954	2979	1998	2431	2469	2499
Source	181	32	47	22	476	48	29	2
Other	8	12	2	2	33	28	9	6
Total	3003	3003	3003	3003	2507	2507	2507	2507

Table 3: The number of outputs (after truncation) classified as being in the (correct) target language, the source language, or another language for 0-shot and 1-shot setups (for WMT).

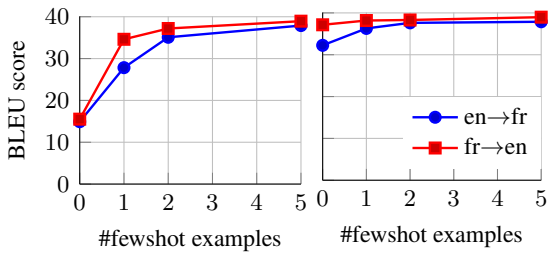
Detecting generation in the wrong language We automatically detect the language of predictions using fasttext langid[17] (Joulin et al., 2017). Table 3 shows the number of translations identified as being in the correct target language, or alternatively in the source or another language for 0-shot and 1-shot setups after truncation.[18,19] The number of sentences in the correct target language increases from 0- to 1-shot, particularly for the two non-English target languages. When translating into Hindi (0-shot), 1/5 (509) of predictions are not detected as Hindi; the 1-shot largely mitigates the issue (only 76 outputs in the wrong language).

Increasing the number of few-shot examples Both problems improve significantly in the 1-shot setup, a trend that continues as the number of few-shot examples increases, resulting in higher BLEU scores, as can be seen in Figure 1 for WMT en↔fr. However, we see diminishing returns, particularly

[17] `https://fasttext.cc/docs/en/language-identification.html`, we use the compressed version `lid.176.ftz`

[18] Raw tables can be found in Tables 13 and 14 in Appendix B.

[19] These numbers are better than the initial ones reported in (BigScience et al., 2022), as we use a different prompt and truncation. See below for a detailed analysis per prompt.



(a) Original outputs. (b) Truncated outputs.

Figure 1: BLEU scores for WMT 2014 en↔fr and the `xglm` prompt, with an increasing number of few-shot examples.

visible between 2 to 5 examples, suggesting that gains beyond 5-shot would be more marginal.

5.2 BLOOM model size

Several versions of BLOOM exist, with differing numbers of parameters. To test how size impacts performance, we report average scores and ranges for WMT across the seven prompts. Table 4 shows that as the size decreases (shown by the successive versions from 176B to 560M parameters), the performance also decreases significantly. We see substantial gains for all models when moving from 0-shot to 1-shot, the smaller models (e.g. BLOOM-7b1, BLOOM-3b) slightly closing the gap with the largest one. As the ranges in Table 4 are computed across prompts, we observe that different prompts yield markedly different BLEU scores in the 0-shot setup; for 1-shot, we still see variations of 6-8 BLEU points between the best and the worst prompt. Similar analyses performed with post-processing and also for English↔Hindi (Appendix C) confirm (a) that truncation greatly improves scores for all model sizes and prompts; (b) that the choice of a bad prompt can result in catastrophic MT performance as compared to a good one.

5.3 Per-prompt analysis

Looking at average results for WMT computed with respect to prompt choice (using the seven prompts in Table 1) allows us to further investigate cross-prompt variability.

Which prompt works best? This variability is illustrated in Tables 5 and 6 report performance across prompts for en↔{fr,hi}, averaged over the five BLOOM models from

Model	en→fr		fr→en	
BLOOM	11.17	2.97–21.96	15.38	10.33–26.79
BLOOM-7b1	6.51	1.52–12.12	12.82	4.77–25.10
BLOOM-3b	3.60	1.21–9.60	10.61	2.84–19.34
BLOOM-1b1	1.66	1.44–3.89	7.10	0.73–11.44
BLOOM-560m	0.57	0.39–0.86	3.66	1.42–5.43

(a) 0-shot

Model	en→fr		fr→en	
BLOOM	32.60	27.83 –36.39	34.91	33.14–36.56
BLOOM-7b1	25.92	20.78–29.94	29.12	25.44–32.51
BLOOM-3b	21.62	16.69–26.83	25.67	18.64–29.56
BLOOM-1b1	10.07	6.31–13.24	16.12	12.17–19.89
BLOOM-560m	3.62	2.16–4.41	8.63	5.83–12.14

(b) 1-shot

Table 4: Average BLEU scores and ranges across the seven prompts for decreasing sizes of BLOOM (original outputs).

Section 5.2.[20] The corresponding tables for truncated outputs are in Appendix D. `version` and `a_good_translation (source+target)` get the highest average (and maximum) scores. Both prompts are more verbose (instruction-like), but the performance gap in the 1-shot setting between these prompts and the simpler, 'priming-style' prompts (e.g. `xglm` narrows). The worst results are seen for `gpt3`. With this prompt, translating into French after a text that only contains English seems particularly difficult: half of the 0-shot translations for `gpt` are classified as non-French by langid (most of them are English). When translating into Hindi, only 10 outputs are detected as being in Hindi.

Does it help to specify the source language in the prompt? We compare the two versions (`-target` and `-source+target`) of `a_good_translation` and `xglm`. Results in Tables 5 and 6 are inconclusive. For these language directions and prompts, we see small differences for 1-shot, which may be due to variance between runs. For 0-shot, it clearly helps `xglm` to indicate the source language, but for the more verbose `a_good_translation`, it helps one direction and hurts the other. This question would need to be further explored to draw more solid conclusions, including with non-English prompts.

[20] For a given prompt, the range mainly reflects the performance of the different sizes of BLOOM model.

Prompt / Few-shot #	en→fr 0	en→fr 1	fr→en 0	fr→en 1
a_good_translation-source+target	6.66 0.57 – 15.38	18.68 4.12 – 36.39	10.96 5.43 – 14.15	25.76 11.64 – 36.56
a_good_translation-target	3.15 0.40 – 10.14	20.32 3.23 – 35.47	12.05 5.11 – 16.81	**25.86** 12.14 – 36.24
gpt3-target	2.48 0.46 – 7.90	16.64 2.16 – 32.55	4.50 0.73 – 12.73	19.28 5.83 – 33.14
translate_as-target	3.30 0.39 – 5.04	17.15 3.23 – 32.74	6.85 2.15 – 11.35	21.63 7.60 – 35.12
version-target	7.48 0.58 – 21.96	**21.38** 4.33 – 34.22	**17.06** 3.87 – 26.79	24.88 7.85 – 35.42
xglm-source+target	**8.28** 0.86 – 14.91	17.52 3.34– 27.83	11.76 5.01 – 15.52	22.09 7.85 – 34.61
xglm-target	1.57 0.70 – 2.97	16.67 4.41 – 28.99	6.23 2.65 – 10.33	20.73 7.48 – 33.30

Table 5: Average, min and max BLEU scores by prompt for en↔fr (original outputs). Best average result per setting in bold.

Prompt / Few-shot #	en→hi 0	en→hi 1	hi→en 0	hi→en 1
a_good_translation-source+target	0.72 0.07–1.90	4.84 0.88 – 10.19	5.80 0.34 – 14.49	13.06 2.78 – 24.60
a_good_translation-target	0.22 0.06–0.77	6.29 1.13 – 13.04	5.54 0.33 – 14.14	13.23 2.84 – 24.84
gpt3-target	0.09 0.03–0.26	0.20 0.00 – 0.66	1.42 0.02 – 6.51	2.23 0.00 – 9.98
version-target	0.74 0.08–1.96	**6.82** 1.74 – 11.48	5.58 0.17 – 13.95	**13.26** 2.42 – 25.80
xglm-source+target	**2.05** 0.06–6.80	4.40 0.63 – 12.05	**6.93** 0.32 – 13.62	11.94 1.66 – 25.04
xglm-target	0.19 0.02–0.63	1.57 0.22 – 4.10	5.13 0.08 – 14.55	6.58 0.47 – 13.22

Table 6: Average, min and max BLEU scores per prompt for en↔hi (original outputs). Best average result per setting in bold.

5.4 Evaluating more language directions

We further explore more language directions in the 1-shot setting using Flores-101. As in Section 5.1, we use the `xglm-source+target` prompt.[21]

5.4.1 Per-language results

To optimise computational resources, instead of running all language combinations, we concentrate on: (i) high-resource language pairs, (ii) high→mid-resource language pairs, (iii) low-resource language pairs and (iv) related languages (specifically Romance languages). Results are shown in Tables 7 and 8 for original outputs, given that overgeneration is less problematic for 1-shot.

High-resource and high→mid-resource The results for high-resource and high→mid-resource languages are generally good, surpassing M2M scores for high-resource, except for es→fr.[22] This suggests that BLOOM a has good multilingual capacity, even across scripts (between (extended) Latin, Chinese, Arabic and Devanagari scripts).

Low-resource For low-resource languages, the results are more variable; some language directions see better results than M2M, notably most into-English directions, but others are less good (e.g. into Hindi and Swahili). Results for the lowest-resourced languages tested (sw↔yo and en↔yo) are particularly disappointing, because the scores indicate that the resulting translations are meaningless, even though Yoruba and Swahili are present (although under-represented) in BLOOM's training data (<50k tokens each).

Romance languages This contrasts with the results between Romance languages, where results are good across-the-board, including from and into Italian (it) and Galician (gl), which are not officially in the training data. Note that Galician shares many similarities with the other Romance languages, in particular with Portuguese (pt). These contrasted results show the performance of an LLM not only depends on the amount of training data, but also largely on the similarity with seen languages. To be complete, these analyses should also take into account the possibility of mislabellings in the training data,[23] which have been found to explain a great deal of cross-

[21] It behaved well on average in the previous experiments and is one of the least verbose, making it more suitable in a multilingual setting.

[22] French and Spanish are related and represent comparable shares of ROOTS, but their scores differ greatly. Our preliminary analysis suggests that this is due to the Spanish references being less literal than the French and structurally more different from the original English. See examples in Appendix E.ff

[23] In a personal communication, N. Muennighoff estimates that Italian accounts for ∼0.33% of the ROOTS corpus, slightly below the proportion of Hindi texts (0.47%). See https://huggingface.slack.com/archives/C03J0PCH895/p1663602387266089.

lingual abilities of LLMs (Blevins and Zettlemoyer, 2022).

Src ↓	Trg →	ar	en	es	fr	zh
ar	BLOOM	–	40.28	23.32	33.12	17.68
	M2M	–	25.50	16.74	25.69	13.10
en	BLOOM	28.21	–	29.42	44.99	26.69
	M2M	17.92	–	25.57	41.99	19.33
es	BLOOM	18.76	32.70	–	24.80	20.92
	M2M	12.11	25.09	–	29.33	14.86
fr	BLOOM	23.44	45.59	27.51	–	23.15
	M2M	15.36	37.17	25.60	–	17.61
zh	BLOOM	15.05	30.50	20.54	26.01	–
	M2M	11.55	20.91	16.92	24.32	–

(a) High-resource language pairs.

Src ↓	Trg →	en	fr	hi	id	vi
en	BLOOM	–	44.99	27.25	39.00	28.54
	M2M	–	41.99	28.15	37.26	35.10
fr	BLOOM	45.59	–	18.47	31.44	32.76
	M2M	37.17	–	22.91	29.14	30.26
hi	BLOOM	35.06	27.62	–	–	–
	M2M	27.89	25.88	–	–	–
id	BLOOM	43.25	30.35	–	–	–
	M2M	33.74	30.81	–	–	–
vi	BLOOM	38.71	26.85	–	–	–
	M2M	29.51	25.82	–	–	–

(b) High→mid-resource language pairs.

Table 7: 1-shot MT results (spBLEU) on the FLORES-101 devtest set (original outputs).

5.4.2 Cross-lingual transfer

1-shot results are positive for many of the language directions tested (including low-resource), provided they are sufficiently represented in the ROOTS corpus. To better understand how cross-lingual BLOOM is and how the 1-shot mechanism functions, we vary the language direction of the few-shot examples, taking Bengali→English (bn→en) translation as our case study. Taking random 1-shot dev set examples,[24] we compare the use of 1-shot examples from (i) the same direction (bn→en), (ii) the opposite direction (en→bn), (iii) a language direction whereby the source languages are related (hi→en), (iv) the same related direction but from a different dataset (the WMT dev set) (v) a high-resource direction into the same target language (fr→en) and (vi) a high-resource unrelated language direction (fr→ar).

The results (Table 9) show that cross-lingual transfer is possible, but using a different language direction can impact overgeneration and translation quality. The unrelated direction fr→ar gives the worst results, with most overgeneration (see the score difference between original and truncated), but also the worst quality after truncation, suggesting that language relatedness does play a role. Overgeneration is still a problem (although less so) when using the opposite direction (en→bn) or the same target language (fr→en). Using a related (higher-resource) source language (hi→en) reduces overgeneration and also gives the best MT results. However, better results are seen when using Flores-101 rather than WMT examples, suggesting that in-domain examples are best.

[24] The random seed is kept the same for all runs.

Src↓	Trg→	en	bn	hi	sw	yo
en	BLOOM	–	24.65	27.25	20.51	2.60
	M2M	–	23.04	28.15	26.95	2.17
bn	BLOOM	29.91	–	16.34	–	–
	M2M	22.86	–	21.76	–	–
hi	BLOOM	35.06	23.77	–	–	–
	M2M	27.89	21.77	–	–	–
sw	BLOOM	37.40	–	–	–	1.31
	M2M	30.43	–	–	–	1.29
yo	BLOOM	4.08	–	–	0.89	–
	M2M	4.18	–	–	1.93	–

(a) Low-resource languages

Src↓	Trg→	ca	es	fr	gl	it	pt
ca	BLOOM	–	28.92	33.79	19.24	19.85	33.05
	M2M	–	25.17	35.08	33.42	25.50	35.17
es	BLOOM	31.16	–	24.80	23.28	16.49	29.11
	M2M	23.12	–	29.33	27.54	23.87	28.10
fr	BLOOM	37.16	27.51	–	24.92	23.97	38.94
	M2M	28.74	25.60	–	32.82	28.56	37.84
gl	BLOOM	37.49	27.09	33.77	–	18.26	32.16
	M2M	30.07	27.65	37.06	–	26.87	34.81
it	BLOOM	31.00	25.40	31.36	20.16	–	29.15
	M2M	25.20	29.23	34.39	29.23	–	31.47
pt	BLOOM	39.56	28.07	40.34	27.10	20.06	–
	M2M	30.69	26.88	40.17	33.77	28.09	–

(b) Romance languages

Table 8: 1-shot MT results (spBLEU) on the Flores-101 devtest set (original outputs).

		Original		Truncated	
1-shot example direction type		spBLEU	COMET	spBLEU	COMET
Same	bn→en	29.91	0.4440	29.91	0.4440
Opposite	en→bn	21.81	0.3132	29.42	0.4143
Related src	hi→en	30.14	0.4492	30.54	0.4603
Related src (WMT)	hi→en	29.06	0.4216	29.07	0.4274
HR unrelated src	fr→en	17.19	0.3147	29.68	0.3960
HR unrelated src	fr→ar	8.44	-0.1025	27.99	0.3218

Table 9: 1-shot results for Flores-101 bn→en when varying the origin of 1-shot examples. HR=high-resource.

5.5 Use of Linguistic Context

There has been a considerable amount of research on linguistic context in MT, e.g. to disambiguate lexically ambiguous texts or when additional in-

1-shot example			en→fr		fr→en	
Origin	Direction	Truncate	BLEU	COMET	BLEU	COMET
Rand.	rand.	×	5.70	0.3421	12.05	0.6138
		✓	37.57	0.6343	41.36	0.7576
Prev.	rand.	×	6.10	0.3280	12.34	0.6166
		✓	38.51	0.6139	41.57	0.7513
Prev.	same	×	19.32	0.5965	20.71	0.7190
		✓	**38.95**	**0.6325**	**42.10**	**0.7607**
Prev.	opp.	×	3.64	0.0635	8.56	0.5184
		✓	37.76	0.5898	41.20	0.7423

Table 10: DiaBLa results when using the previous or a random sentence as the 1-shot example. In bold the best results for each direction.

formation is necessary for the output to be well-formed (e.g. translating anaphoric pronouns into a language that requires agreement with a coreferent) (Hardmeier, 2012; Libovický and Helcl, 2017; Bawden et al., 2018; Voita et al., 2018; Lopes et al., 2020; Nayak et al., 2022).

We test the usefulness of linguistic context in DiaBLa in the 1-shot setting (again using `xglm-source+target`) by changing the origin of 1-shot examples: (i) a random example vs. (ii) the previous dialogue utterance. If linguistic context is useful, we would expect there to be an improvement for (ii). We also vary the language direction of the 1-shot example. By default, given that the dataset is bilingual, the direction of 1-shot examples is en→fr or fr→en, independent of the current example's direction. Given the results in Section 5.4.2 and the poor 0-shot results in Table 2a, it is important to account for this to provide a fair comparison. We therefore compare each type of context (random/previous) with (i) the same random directions, and (ii-iii) the same (and opposite) language directions as the current example. We show results for original and truncated outputs.

Results are shown in Table 10. Truncation helps considerably; even for 1-shot, BLOOM struggles not to overgenerate and this is considerably reduced when the same rather than the opposite language direction is used for the 1-shot example. It is unclear whether using previous rather than random context helps: BLEU is higher (38.51 vs. 37.57), whereas COMET is lower (0.3280 vs. 0.3421). These differences could be the result of randomness in 1-shot example selection, and different results could be obtained with a different random seed. Despite these inconclusive results, it is clear that using previous context influences the translation, for better or worse. Table 11 provides three such examples: (i) an unlucky negative influence on the translation of an ambiguous word *glace* 'ice cream or mirror' from the previous context, resulting in the wrong sense being chosen, (ii) the use of a coreferent *instrument* 'instrument' from the previous sentence and (iii) the correct gender agreement of the pronoun *they* into French (*elles* 'they (fem.)' as opposed to *ils* 'they (masc.)') to correspond to the feminine coreferent *filles* 'girls'.

6 Conclusion

We have evaluated BLOOM's MT performance across three datasets and multiple language pairs. While there remain problems of overgeneration and generating in the wrong language (particularly for 0-shot MT), MT quality is significantly improved in few-shot settings, closer to state-of-the-art results. Low-resource MT remains challenging for some language pairs, despite the languages being in the training data, questioning what it means to be a BLOOM language. However, we see evidence for cross-lingual transfer, for non-BLOOM languages and when using few-shot examples from other language pairs. Finally, although using linguistic context does not give improvements with automatic metrics, there is evidence that discursive phenomena are taken into account.

Acknowledgements

This work would not have been possible without the collective efforts of all the members of the BigScience collaboration, who specified designed, developed and prepared the tools and datasets that were needed to train BLOOM models. Special mention to the all members of the evaluation working group for the discussions and the developments of the `eval_harness` framework and resources that were used in this study, and among them to Niklas Muennighoff and Pawan Sasanka Ammanamanchi for producing some of the experimental results discussed in our work.

This work was granted access to the HPC resources of Institut du développement et des ressources en informatique scientifique (IDRIS) du Centre national de la recherche scientifique (CNRS) under the allocations 2021-AD011011717R1, AD011012254R2, 2021-A0101012475 and 2022-AD010614012 made by Grand équipement national de calcul intensif

1-shot origin	Context	Reference	Prediction
Prev. same	French: Pensez vous en permanence à la glace qui se mange ? = English: Do you constantly think about the ice-cream that's eaten? ### French: ou bien à une glace pour se regarder ? = English:		*or to an ice-cream to look at?
Prev. opposite	English: Do you constantly think about the ice-cream that's eaten? = French: Pensez vous en permanence à la glace qui se mange ? ### French: ou bien à une glace pour se regarder ? = English:	Or a mirror to look into?	or to a mirror to look at?
Random	French: N'empêche, on vit une époque folle, folle! = English: Still, what a crazy, crazy time we're living in! ### French: ou bien à une glace pour se regarder ? = English:		or to a mirror to look at yourself?
Prev. same	English: What kind of instrument were you thinking of? = French: Tu penses à quelle sorte d'instrument ? ### English: A wooden one I suppose... = French:		Un instrument en bois, je suppose...
Prev. opposite	French: Tu penses à quelle sorte d'instrument ? = English: What kind of instrument were you thinking of? ### English: A wooden one I suppose... = French:	Un instrument en bois, je suppose.	Un instrument en bois, je suppose...
Random	French: Ils vont vous changer les idées après votre dure journée ! = English: They'll help you take your mind off things after your hard day! ### English: A wooden one I suppose... = French:		Un en bois, je suppose...
Prev. same	English: He showed me how it works, but if I get stuck the girls in here will always help me. = French: Il m'a montré comment cela marchait, mais si je n'y arrive pas, les filles ici m'aideront sans problème. ### English: They are very kind. = French:		Elles sont très gentilles.
Prev. opposite	French: Il m'a montré comment cela marchait, mais si je n'y arrive pas, les filles ici m'aideront sans problème. = English: He showed me how it works, but if I get stuck the girls in here will always help me. ### English: They are very kind. = French:	Elles sont très gentilles.	Elles sont très gentilles.
Random	English: I don't know about loans. = French: Je ne sais pas pour les prêts. ### English: They are very kind. = French:		*Ils sont très gentils.

Table 11: Ambiguous DiaBLa examples with different 1-shot contexts. Words that are relevant to the ambiguity are underlined and incorrect translations are marked with an asterisk.

(GENCI).

Rachel Bawden's participation was partly funded by her chair position in the PRAIRIE institute, funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001, and her Emergence project, DadaNMT, funded by Sorbonne Université.

References

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *CoRR*, abs/2205.03983.

Rachel Bawden, Eric Bilinski, Thomas Lavergne, and Sophie Rosset. 2021. DiaBLa: a corpus of bilingual spontaneous written dialogues for machine translation. *Language Resources and Evaluation*, 55(3):635–660.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Workshop BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo

Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

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

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*,

pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

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

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.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Xavier Garcia, Yamini Bansal, Colin Cherry, George F. Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *CoRR*, abs/2302.01398.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *CoRR*, abs/2202.11822.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours - Revue de linguistique, psycholinguistique et informatique*, (11).

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *CoRR*, abs/2302.09210.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

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

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre–train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9).

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *CoRR*, abs/2301.13294.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *CoRR*, abs/2211.01786.

Prashanth Nayak, Rejwanul Haque, John D Kelleher, and Andy Way. 2022. Investigating contextual influence in Document-Level translation. *Information*, 13(5):249.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia,

Pennsylvania, USA. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Minh Quang Pham, Jitao Xu, Josep Crego, François Yvon, and Jean Senellart. 2020. Priming neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 516–527, 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, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical report*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, 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, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *Proceedings of the International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, Chandana Satya Prakash, Mukund Sridhar, Fabian Triefenbach, Apurv Verma, Gökhan Tür, and Prem Natarajan. 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *CoRR*, abs/2208.01448.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *CoRR*, abs/2211.09102.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*, New Orleans, Louisiana, USA.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. GLM-130B: an open bilingual pre-trained model. *CoRR*, abs/2210.02414.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *CoRR*, abs/2301.07069.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models.

A COMET Results for Main Comparison

Table 12 shows the COMET scores for the cross-dataset and model comparison. The conclusions drawn for the Table 2 with BLEU scores hold here.

	0-shot				1-shot			
	BLOOM	T0	mT0-xxl	OPT	BLOOM	T0	mT0-xxl	OPT
WMT 2014								
en→fr	-0.9851	-0.6996	0.4531	-0.9193	0.0852	-1.0354	-0.0152	-0.1650
fr→en	-0.6755	0.3369	0.5668	-0.4934	0.4484	-0.0874	0.2501	0.0394
en→hi	-0.4817	-1.8185	0.4838	-1.5246	0.2885	-1.7329	0.0258	-1.4600
hi→en	-0.3866	-1.3462	0.5137	-1.2004	0.3779	-1.6239	-0.0188	-1.2901
DiaBLa								
en→fr	-1.5725	-0.5280	0.3801	-1.7616	0.3421	-0.5845	-0.0184	0.1231
fr→en	-1.5807	0.2281	0.5344	-1.5074	0.6138	-0.0321	0.3655	0.3894
Flores-101								
en→fr	-1.4689	-0.6821	0.7972	-1.4381	0.6022	-0.9834	0.6052	0.1301
fr→en	-1.1425	0.4989	0.8325	-1.0078	0.6873	-0.0805	0.7061	0.4040
en→hi	-0.9716	-1.8484	1.0252	-1.6989	0.4535	-1.7954	0.7184	-1.6218
hi→en	-0.3387	-1.3913	0.7967	-1.4932	0.5380	-1.2642	0.6673	-1.2633

(a) Original predictions

	0-shot				1-shot			
	BLOOM	T0	mT0-xxl	OPT	BLOOM	T0	mT0-xxl	OPT
WMT 2014								
en→fr	0.4338	-0.6999	0.4517	0.0340	0.4237	-1.0352	-0.0166	-0.0001
fr→en	0.6042	0.3356	0.5661	0.5340	0.5317	-0.0901	0.2474	0.4485
en→hi	0.0527	-1.8185	0.4829	-1.4908	0.4480	-1.7328	0.0258	-1.4599
hi→en	0.4454	-1.3462	0.5111	-1.1131	0.3864	-1.6240	-0.0221	-1.2739
DiaBLa								
en→fr	0.4332	-0.5280	0.3801	-0.0021	0.6343	-0.5847	-0.0232	0.1916
fr→en	0.5672	0.2281	0.5337	0.5544	0.7576	-0.0391	0.3561	0.6395
Flores-101								
en→fr	0.1817	-0.6832	0.7932	0.0266	0.6217	-0.9840	0.6008	0.1797
fr→en	0.6968	0.4989	0.8306	0.6887	0.6901	-0.0859	0.7018	0.5936
en→hi	-0.6081	-1.8486	1.0252	-1.6384	0.4614	-1.7954	0.7184	-1.6218
hi→en	0.5088	-1.3913	0.7967	-1.1656	0.5380	-1.2641	0.6662	-1.2512

(b) Truncated predictions

Table 12: Comparison of COMET scores across the three datasets using the `xglm-source+target` prompt.

	0-shot		1-shot		2-shot		5-shot	
	N	Δ	N	Δ	N	Δ	N	Δ
cs	1	408	-	-	-	-	-	-
de	1	3	2	146	2	-12.5	1	2
en	181	16	32	57	10	73.8	8	92.2
es	1	12	3	89.3	-	-	-	-
fr	2814	7.9	2959	2.1	2989	1.5	2992	1.6
ht	1	57	1	89	-	-	-	-
it	2	4.5	3	13.3	-	-	-	-
nl	1	131	-	-	-	-	-	-
pt	1	146	-	-	-	-	-	-
ms	-	-	1	28	-	-	-	-
ru	-	-	1	16	-	-	-	-
zh	-	-	1	10	-	-	-	-
ca	-	-	-	-	1	198	1	18
uk	-	-	-	-	1	3	1	3

(a) en→fr

	0-shot		1-shot		2-shot		5-shot	
	N	Δ	N	Δ	N	Δ	N	Δ
en	2954	1	2979	0.8	2988	1	2987	1.3
fr	47	-23.4	22	-1.4	13	1.3	13	-2.2
it	1	3	-	-	2	6	3	5.3
tr	1	-1	1	-1	-	-	-	-
es	-	-	1	1	-	-	-	-

(b) fr→en

Table 13: Raw figures for language identification and length differences of outputs compared to the reference translation for WMT2014 en→fr using the `xglm-source+target` prompt. For 0-5 few-shot examples, N is the number of sentences identified as being in each language (the target language's row (correct) is indicated in green and the source language's row (one of the many incorrect options) in red) and Δ is the length difference in number of characters (N.B. it is negative when the prediction is longer than the reference).

B Wrong language prediction and over-generation

As described in Section 5.1, one problem identified with BLOOM, particularly for 0-shot translation, is generating in the wrong language. Tables 13 and 14 give the full analysis including raw figures for language identification for WMT2014 fr↔en and hi↔en translation directions. For 0-5 few-shot examples, we indicate the number of truncated outputs identified as being from each language (indicated by the rows), the correct language (the target) being indicated in green, and the source language (therefore incorrect) being indicated in red. We also provide the average length difference (Δ) between BLOOM's outputs and the reference translations (negative numbers indicate that the prediction is longer than the reference).

For 0-shot translation, a significant number of examples are classed as being in the source language for en→fr, and even more so for en→hi (almost one fifth of the outputs are in the wrong language). As we increase the number of few-shot examples used, both of these problems are significantly reduced, and almost disappear for all language pairs and directions with 5 examples.

C Analysis per model

In this section, we complete the results of Section 5.2 with Tables 15 and 16, respectively for French↔English and Hindi↔English, reporting results without truncation. As expected, the systems are ranked according to their size. For French–English we see that decent performance can already be obtained with the second largest model BLOOM-7b1, using 1-shot. Using this model, or even a model half this size can provide good indication of the performance of prompts, and be reliably used as test beds. We obtain less satisfactory results with English↔Hindi, even with the large BLOOM; for this language pair, we even observe a large variation across prompts

	0-shot		1-shot		2-shot		5-shot	
	N	Δ	N	Δ	N	Δ	N	Δ
ceb	1	-150	-	-	-	-	-	-
en	476	10.5	48	12.4	71	13.9	26	18.8
eo	1	-134	-	-	-	-	-	-
fi	1	19	-	-	-	-	-	-
fr	2	94.5	-	-	-	-	-	-
gom	2	6.5	1	4	-	-	1	0
hi	1998	9.3	2431	6	2403	5.5	2457	5.5
hsb	1	98	-	-	-	-	-	-
ht	2	147	6	257.5	11	135.3	1	158
hu	1	71	-	-	-	-	-	-
lv	3	63.3	-	-	-	-	-	-
mr	5	64.4	11	14.6	17	11.7	19	6
ne	5	7.6	9	28.2	4	16.8	3	8.3
nl	2	-13.5	-	-	-	-	-	-
pt	1	24	-	-	-	-	-	-
sa	1	-25	-	-	-	-	-	-
sw	1	12	-	-	-	-	-	-
tl	1	24	-	-	-	-	-	-
war	3	3	-	-	-	-	-	-
vec	-	-	1	-38	-	-	-	-
new	-	-	-	-	1	25	-	-

(a) en→hi

	0-shot		1-shot		2-shot		5-shot	
	N	Δ	N	Δ	N	Δ	N	Δ
en	2469	4	2499	5.1	2503	3.8	2498	3
fr	1	151	1	-5	-	-	1	8
hi	29	3.3	2	0	-	-	-	-
ht	6	199.8	-	-	-	-	-	-
it	1	139	-	-	1	-18	3	4.3
nl	1	9	-	-	-	-	2	-3
id	-	-	1	-6	-	-	-	-
nds	-	-	1	16	-	-	-	-
pl	-	-	1	-14	-	-	-	-
tr	-	-	1	-15	-	-	-	-
war	-	-	1	344	-	-	-	-
de	-	-	-	-	1	-15	1	188
es	-	-	-	-	1	2	-	-
la	-	-	-	-	1	17	-	-
fi	-	-	-	-	-	-	1	-1
pt	-	-	-	-	-	-	1	1

(b) hi→en

Table 14: Raw figures for language identification and length differences of outputs compared to the reference translation for WMT2014 en→hi using the `xglm-source+target` prompt. For 0-5 few-shot examples, N is the number of sentences identified as being in each language (the target language's row (correct) is indicated in green and the source language's row (one of the many incorrect options) in red) and Δ is the length difference in number of characters (N.B. it is negative when the prediction is longer than the reference).

(looking at the range of scores) in the 1-shot setting for all models.

D Analysis per prompt

In this section, we replicate the analysis of Section 5.3 and report results per prompt with truncated outputs in Tables 17 and 18. The conclusions are overall consistent with what we report for non-truncated outputs in the main text. We note that after truncating the outputs, `xglm-source+target` yields very good results across the board, outperforming its closest contenders `a_good_translation-source+target` and `version-target` in almost all configurations. However, the choice of the prompt seems to matter more (a) in the zero-shot setting, (b) when translating out of English. Conversely our more stable results are for fr-en, 1-shot.

E Translation divergences in Flores 101

A striking observation reported in the main text (Section 5.4.1) is the difference between French and Spanish for the Flores-101 experiments. This is unexpected, as both languages are well represented in the training data. Yet, when translating from and into English the difference in spBLEU score is huge; and there is a clear gap with the other Romance languages as well. A related question is the poor translation between French and Spanish, not much better than for French→Arabic. Looking at some sample outputs, this seems to be due to the peculiarities of the Spanish translations, which appear to be less literal than their French counterparts, but which yield equally good translations into English. This can be seen when we compare translations back into English for these languages (see a random subset in Table 19). The last example illustrates this very clearly: we see "34 percent" in both the original English and in the translation from French, while translation from Spanish starts with "one third".

Model / Direction	0-shot en→fr	0-shot fr→en	1-shot en→fr	1-shot fr→en
BLOOM	**17.45** 4.60 – 32.25	**26.87** 16.86 – 37.16	**35.64** 33.08 – 37.12	**36.94** 35.72 – 38.18
BLOOM-7b1	10.51 2.08 – 25.15	21.81 7.21 – 33.28	28.07 24.73 – 30.39	30.77 27.97 – 32.78
BLOOM-3b	6.25 1.89 – 19.24	17.92 5.08 – 28.85	23.81 19.26 – 26.98	27.18 22.02 – 29.68
BLOOM-1b1	2.82 0.69– 7.92	11.67 1.29 – 20.42	11.56 7.32 – 17.22	17.93 13.43 – 21.68
BLOOM-560m	0.80 0.55 – 1.33	5.16 2.24 – 8.33	4.36 3.03 – 6.85	9.91 7.25 – 12.50

Table 15: Average, min and max BLEU scores per model of increasing size, for WMT14 en↔fr (original outputs). Best average result per setting in bold.

Model / Direction	0-shot en→hi	0-shot hi→en	1-shot en→hi	1-shot hi→en
BLOOM	**2.05** 0.26 – 6.80	**8.59** 0.66 – 13.04	**12.88** 6.51 – 14.55	**20.58** 9.98 – 25.80
BLOOM-7b1	0.95 0.05 – 2.96	5.67 0.29 – 9.50	5.92 0.30 – 10.38	13.03 1.03 – 17.53
BLOOM-3b	0.22 0.03 – 0.54	3.86 0.02 – 6.96	4.89 0.21 – 7.25	9.76 0.08 – 13.86
BLOOM-1b1	0.07 0.02 – 0.11	1.40 0.00 – 4.54	1.44 0.05 – 3.08	5.18 0.00 – 8.15
BLOOM-560m	0.06 0.03 – 0.08	0.79 0.01 – 1.74	0.21 0.02 – 0.34	1.70 0.06 – 2.84

Table 16: Average, min and max BLEU scores per model of decreasing size, for WMT14 en↔hi (original outputs). Best average result per setting in bold.

Prompt / Few-shot #	en→fr 0	en→fr 1	fr→en 0	fr→en 1
a_good_translation-source+target	8.54 0.69–17.04	16.42 7.49–22.19	19.08 4.32–37.12	25.99 11.96–36.96
a_good_translation-target	4.62 0.55–13.89	21.72 6.60–35.20	20.88 3.37–36.81	26.31 12.50–36.90
gpt3-target	4.03 0.67–13.96	8.31 1.29–25.71	18.71 3.03–36.42	21.57 7.25–37.18
translate_as-target	6.45 0.60–10.12	11.54 2.29–20.35	18.07 3.55–33.08	22.88 8.15–35.72
version-target	9.66 0.68–30.31	22.20 4.71–35.17	21.88 4.37–36.71	25.31 7.99–37.24
xglm-source+target	**17.18** 1.33–32.25	**25.61** 8.33–37.16	**23.18** 5.02–36.29	**26.65** 11.10–38.18
xglm-target	2.50 1.08–4.60	11.00 4.45–17.61	20.08 6.85–33.08	23.10 10.45–36.42

Table 17: Average, min and max BLEU scores per prompt for WMT14 en↔fr (truncated outputs). Best average result per setting in bold.

Prompt / Few-shot #	en→hi 0	en→hi 1	hi→en 0	hi→en 1
a_good_translation-source+target	1.18 0.08–3.29	6.25 1.02–12.65	5.77 0.32–14.46	13.00 2.61–24.44
a_good_translation-target	0.36 0.07–1.34	10.83 1.08–25.40	5.50 0.31–14.11	13.21 2.72–24.74
gpt3-target	0.04 0.02–0.05	0.01 0.00–0.03	1.64 0.03–7.59	2.50 0.00–11.44
version-target	1.05 0.10–2.98	**11.34** 2.39–21.39	5.56 0.15–13.90	**13.52** 2.68–25.75
xglm-source+target	**3.94** 0.07–12.10	8.80 0.86–24.29	**7.32** 0.23–15.73	12.40 1.21–25.04
xglm-target	0.29 0.02–1.03	2.14 0.27–5.75	5.11 0.01–14.46	6.54 0.15–13.01

Table 18: Average, min and max BLEU scores per prompt for WMT14 en↔hi (truncated outputs). Best average result per setting in bold.

en	They are cooler than the surrounding surface in the day and warmer at night.
fr→en	“They are cooler than the surrounding surface during the day and warmer at night ”.
es→en	During the day, its temperature is lower than that of the surrounding surface, and at night, higher.
en	“This is not going to be goodbye. This is the closing of one chapter and the opening of a new one.”
fr→en	“It’s not goodbye. It’s a page that is turning, and another that is opening.”
es→en	”This will not be a farewell; it is just the end of one chapter and the beginning of another”.
en	“We now have 4-month-old mice that are non-diabetic that used to be diabetic,” he added.
fr→en	”We now have mice that are four months old and are not diabetic, whereas they were before”, he added.
es→en	“Currently, we have mice that are four months old and used to be diabetic, but they are no longer diabetic”, he added.
en	“We will endeavour to cut carbon dioxide emissions per unit of GDP by a notable margin by 2020 from the 2005 level,” Hu said.
fr→en	“We will strive to significantly reduce carbon dioxide emissions per unit of GDP by 2020 compared to the 2005 level,” said Mr. Hu.
es→en	Hu said, “We will work hard to reduce the level of carbon dioxide emitted per unit of GDP by 2020, so that the difference is significant compared to 2005.”
en	Scientists say this animal’s plumage was chestnut-brown on top with a pale or carotenoid-colored underside.
fr→en	Scientists say that the plumage of this animal was chestnut brown on top and pale or carotenoid on the underside.
es→en	According to the experts, this animal has a brown plumage on the upper part and a pale or carotenoid color on the lower part.
en	34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia’s last monarch.
fr→en	34 % of the people surveyed share this view, and want Queen Elizabeth II to be the last monarch to rule Australia.
es-en	One third of the respondents share this view and want the last queen to be Queen Elizabeth II.

Table 19: A random subset of Flores-101 examples translated using BLOOM into English from French and Spanish (N.B. English was the original language of the sentences). Each block of three sentences contains the original English and the automatic French→English and Spanish→English translations.