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),

[1] https://hf.co/bigscience/bloom
[2] https://github.com/rbawden/mt-bigscience

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). These results have since been confirmed for other monolingual LMs such as T5 (Raffel et al., 2020) and multilingual LMs such as XGLM (Lin et al., 2022), PALM (Chowdhery et al., 2022), and ALEXATM (Soltan et al., 2022). However, the focus of these studies has mainly been multi-task performance, with little analysis of MT results. Moreover, results are often only for a few well-resourced language pairs (e.g. English-French and English-German) and the scores reported (mostly BLEU) not always easy to compare.

There are however a number of recent in-depth analyses of MT performance of LLMs, each focusing, like we do, on one specific LM. Most discuss, as we do, the variation of performance with respect to prompt design and number of few-shots examples. This is the case for example of Chowdhery et al. (2022), who reanalyse PALM's translations and Zhang et al. (2023), who focus on GLM-130B, a bilingual (Chinese and English) LLM (Zeng et al., 2022). Consistent with our findings, these studies observe commandable zero-shot performance, with a great variation depending on prompt choices, which tends to diminish when more prompts are used. Using more than 5-10 examples, however, seems to bring very little return. The choice of few-shot examples does make a difference, as also observed by Moslem et al. (2023) in their evaluation of OpenAI's GPT-3 (Brown et al., 2020).[3] The study considers a single prompt resembling our `xglm-source+target` prompt, but varies the strategy used to select examples, showing that prompting can effectively serve as a vehicle to perform local adaptation and to enforce terminological consistency. Finally it is worth mentioning the preliminary evaluation of CHATGPT in (Jiao et al., 2023), and the more detailed one in (Hendy et al., 2023), which confirms the strong translation abilities of this model, at least for "well-resourced"[4] language pairs.

Overall, all these studies contribute to a better understanding of the abilities of instruction-based MT, and provide complementary angles, with variation across tasks, domains, language pairs, settings (e.g. context-aware MT or translation-memory-based MT), as well as evaluation metrics (BLEU, BLEURT, COMET) and protocols. In comparison, ours brings some additional observations related to MT performance across model sizes and for a large number of language pairs, as well as a new task (multilingual conversations).

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 limited language coverage, the latest versions of these approaches are able to handle hundreds of languages, including very low-resource language pairs (Fan et al., 2021; Bapna et al., 2022; Costa-jussà et al., 2022). Although we found that BLOOM is able to match this performance, given sufficient training data, 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 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),[5] 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.[6]

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.[7] These test sets

[3] Version: `text-davinci-003` model.

[4] A rather slippery concept in this context as the training data content, seemingly mostly English, is not fully known.

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

[6] `https://hf.co/bigscience/bloom`

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

	Prompt name	Prompt	Target
1–2	a_good_translation	Given the following source text (in L1): [source sentence], a good L2 translation is:	[target sentence]
3	version	If the original version says [source sentence] then the L2 version should say:	[target sentence]
4	gpt3	What is the L2 translation of the sentence: [source sentence]?	[target sentence]
5–6	xglm	(L1:) [source sentence] = L2:	[target sentence]
7	translate_as	[source sentence] translates into L2 as:	[target sentence]

Table 1: Seven 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 second `-source+target` version, where the source language (in red and in brackets) is explicit in the instruction.

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. We use it to test and compare BLOOM's multilinguality, including for low-resource languages. 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).[8] 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.

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[9] (Sanh et al., 2022), trained on English texts, and MT0-XXL[10] (Muennighoff et al., 2022), the multilingual version, and (ii) OPT[11] (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: (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:[12] (Brown et al., 2020) for `gpt3`, (Lin et al., 2022) for `xglm` and (Wei et al., 2022) for `translate_as`, which also resembles Raffel et al. (2020)'s prompt (*Translate English to German: "[source text]": [target sentence]*).

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).

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

[9] https://hf.co/bigscience/T0

[10] https://hf.co/bigscience/mt0-xxl

[11] https://hf.co/facebook/opt-66b

[12] 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)).

5.1 Comparison across datasets

	0-shot				1-shot			
	BLOOM	T0	mT0	OPT	BLOOM	T0	mT0	OPT
WMT 2014								
en→fr	14.9	1.2	29.3	12.9	27.8	1.4	25.2	21.9
fr→en	15.5	25.8	32.9	15.5	34.6	21.0	30.0	24.6
en→hi	6.8	0.2	11.2	0.1	13.6	0.1	9.5	0.1
hi→en	12.1	0.0	26.1	0.4	25.0	0.0	20.1	0.6
DiaBLa								
en→fr	0.9	0.5	28.4	0.5	5.7	0.6	21.0	15.5
fr→en	0.8	25.5	35.0	0.8	12.1	20.6	26.9	12.1
Flores-101								
en→fr	2.8	1.9	55.5	2.8	45.0	2.1	53.5	24.4
fr→en	2.7	31.9	60.1	2.6	45.6	24.9	58.2	16.7
en→hi	1.3	0.1	67.7	0.1	27.2	0.1	54.7	0.1
hi→en	3.4	0.0	59.5	0.1	35.1	0.2	57.3	0.5

(a) Original predictions

	0-shot				1-shot			
	BLOOM	T0	mT0	OPT	BLOOM	T0	mT0	OPT
WMT 2014								
en→fr	32.2	1.2	29.2	18.9	36.3	1.4	25.2	22.3
fr→en	37.2	25.8	32.9	33.2	38.2	21.1	29.9	33.2
en→hi	12.1	0.2	11.2	0.1	15.7	0.1	9.5	0.1
hi→en	24.3	0.0	26.1	0.5	25.0	0.0	20.1	0.6
DiaBLa								
en→fr	24.2	0.5	28.4	17.4	37.6	0.6	21.9	20.7
fr→en	22.9	25.5	34.9	36.8	41.4	21.1	27.2	37.6
Flores-101								
en→fr	26.9	1.9	55.3	21.4	49.3	2.1	53.4	28.4
fr→en	40.3	31.9	60.0	39.4	47.2	25.2	58.2	39.8
en→hi	7.7	0.1	67.7	0.1	29.5	0.1	54.7	0.1
hi→en	30.2	0.0	59.5	0.2	35.1	0.2	57.3	0.5

(b) Truncated predictions

Table 2: Cross-dataset comparison of BLEU scores (spBLEU for Flores-101) 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 the following 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 scores are in Table 2a for both 0-shot and 1-shot (results with COMET are given in Appendix A). There are issues for 0-shot MT for all directions, particularly when translating into non-English languages, (BLEU scores are systematically lower than into English). Even into English, the scores remain low with respect to state of the art (e.g. 2.7 BLEU for Flores-101 fr→en BLOOM vs. 60.1 for MT0-XXL).[13] MT0-XXL's good results 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.

[13]For comparison, (Bi et al., 2020) reports state-of-the art BLEU scores for supervised MT as 45.6 and 45.4 for WMT14 en→fr and fr→en respectively.

What are the main issues? We identify two main issues (particularly for 0-shot): (i) over-generation 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 newline or the regular expression pattern = `.+?:` is discarded.

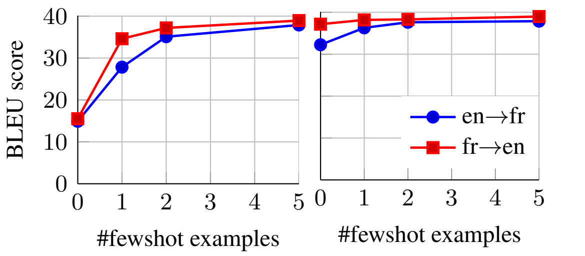
Results after truncation (Table 2b) show that for all three datasets, 0-shot and 1-shot scores are significantly improved (e.g. 1-shot DiaBLa fr→en increases from 12.05 to 41.36 and 0-shot Flores-101 hi→en increases from 3.40 to 30.19). BLOOM is capable of performing good MT but has a problem knowing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

Detecting generation in the wrong language We automatically detect the language of predictions

	en→fr		fr→en		en→hi		hi→en	
	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).

using fasttext langid[14] (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.[15,16] 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 are in the wrong language).



(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.

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 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 (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 see 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 that (i) truncation improves scores for all model sizes and prompts and (ii) the choice of a bad prompt can result in catastrophic MT performance as compared to a good one.

Model	en→fr		fr→en	
BLOOM	11.2	3.0–22.0	15.4	10.3–26.8
BLOOM-7b1	6.5	1.5–12.1	12.8	4.8–25.1
BLOOM-3b	3.6	1.2–9.6	10.6	2.8–19.3
BLOOM-1b1	1.7	0.5–3.9	7.1	0.7–11.4
BLOOM-560m	0.6	0.4–0.9	3.7	1.4–5.4

(a) 0-shot

Model	en→fr		fr→en	
BLOOM	32.6	27.8 –36.4	34.9	33.1–36.6
BLOOM-7b1	25.9	20.8–29.9	29.1	25.4–32.5
BLOOM-3b	21.6	16.7–26.8	25.7	18.6–29.6
BLOOM-1b1	10.1	6.3–13.2	16.1	12.2–19.9
BLOOM-560m	3.6	2.2–4.4	8.6	5.8–12.1

(b) 1-shot

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

5.3 Per-prompt analysis

Looking at average WMT results computed with respect to prompt choice (using the 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 Section 5.2.[17] 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),

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

[15] See the raw results in Tables 12 and 13 in Appendix B.

[16] 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.

[17] 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	6.7 0.6–15.4	18.7 4.1–36.4	11.0 5.4–14.2	25.8 11.6–36.6
a_good_translation-target	3.1 0.4–10.1	20.3 3.2–35.5	12.1 5.1–16.8	**25.9** 12.1–36.2
gpt3-target	2.5 0.5– 7.9	16.6 2.2–32.5	4.5 0.7 –12.7	19.3 5.8–33.1
translate_as-target	3.3 0.4– 5.0	17.1 3.2–32.7	6.9 2.1 –11.3	21.6 7.6–35.1
version-target	7.5 0.6–22.0	**21.4** 4.3–34.2	**17.1** 3.9–26.8	24.9 7.8–35.4
xglm-source+target	**8.3** 0.9–14.9	17.5 3.3– 27.8	11.8 5.0–15.5	22.1 7.8–34.6
xglm-target	1.6 0.7– 3.0	16.7 4.4–29.0	6.2 2.6–10.3	20.7 7.5–33.3

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	0.7 0.1–1.9	5.8 0.3–14.5	4.8 0.9–10.2	13.1 2.8–24.6
a_good_translation-target	0.2 0.1–0.8	5.5 0.3–14.1	6.3 1.1–13.0	13.2 2.8–24.8
gpt3-target	0.1 0.0–0.3	1.4 0.0–6.5	0.2 0.0–0.7	2.2 0.0–10.0
version-target	0.7 0.1–2.0	5.6 0.2–14.0	**6.8** 1.7–11.5	**13.3** 2.4–25.8
xglm-source+target	**2.1** 0.1–6.8	**6.9** 0.3–13.6	4.4 0.6–12.1	11.9 1.7–25.0
xglm-target	0.2 0.0–0.6	5.1 0.1–14.6	1.6 0.2–4.1	6.6 0.5–13.2

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

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 `gpt3` 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.

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.

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 language directions are generally good, surpassing M2M scores for high-resource, except for es→fr.[18] 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

[18]French and Spanish, although related and comparably represented in ROOTS, have very different scores. Our preliminary analysis suggests that this is due to the Spanish references being less literal than the French. See Appendix E for some examples.

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,[19] which have been found to explain a great deal of cross-lingual abilities of LLMs (Blevins and Zettlemoyer, 2022).

Src ↓	Trg →	ar	en	es	fr	zh
ar	BLOOM	–	40.3	23.3	33.1	17.7
	M2M	–	25.5	16.7	25.7	13.1
en	BLOOM	28.2	–	29.4	45.0	26.7
	M2M	17.9	–	25.6	42.0	19.3
es	BLOOM	18.8	32.7	–	24.8	20.9
	M2M	12.1	25.1	–	29.3	14.9
fr	BLOOM	23.4	45.6	27.5	–	23.2
	M2M	15.4	37.2	25.6	–	17.6
zh	BLOOM	15.0	30.5	20.5	26.0	–
	M2M	11.6	20.9	16.9	24.3	–

(a) High-resource language pairs.

Src ↓	Trg →	en	fr	hi	id	vi
en	BLOOM	–	45.0	27.2	39.0	28.5
	M2M	–	42.0	28.1	37.3	35.1
fr	BLOOM	45.6	–	18.5	31.4	32.8
	M2M	37.2	–	22.9	29.1	30.3
hi	BLOOM	35.1	27.6	–	–	–
	M2M	27.9	25.9	–	–	–
id	BLOOM	43.2	30.4	–	–	–
	M2M	33.7	30.8	–	–	–
vi	BLOOM	38.7	26.8	–	–	–
	M2M	29.5	25.8	–	–	–

(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,[20] 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.

[19]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%).

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

Src↓	Trg→	en	bn	hi	sw	yo
en	BLOOM	–	24.6	27.2	20.5	2.6
	M2M	–	23.0	28.1	26.9	2.2
bn	BLOOM	29.9	–	16.3	–	–
	M2M	22.9	–	21.8	–	–
hi	BLOOM	35.1	23.8	–	–	–
	M2M	27.9	21.8	–	–	–
sw	BLOOM	37.4	–	–	–	1.3
	M2M	30.4	–	–	–	1.3
yo	BLOOM	4.1	–	–	0.9	–
	M2M	4.2	–	–	1.9	–

(a) Low-resource languages

Src↓	Trg→	ca	es	fr	gl	it	pt
ca	BLOOM	–	28.9	33.8	19.2	19.8	33.0
	M2M	–	25.2	35.1	33.4	25.5	35.2
es	BLOOM	31.2	–	24.8	23.3	16.5	29.1
	M2M	23.1	–	29.3	27.5	23.9	28.1
fr	BLOOM	37.2	27.5	–	24.9	24.0	38.9
	M2M	28.7	25.6	–	32.8	28.6	37.8
gl	BLOOM	37.5	27.1	33.8	–	18.3	32.2
	M2M	30.1	27.6	37.1	–	26.9	34.8
it	BLOOM	31.0	25.4	31.4	20.2	–	29.2
	M2M	25.2	29.2	34.4	29.2	–	31.5
pt	BLOOM	39.6	28.1	40.3	27.1	20.1	–
	M2M	30.7	26.9	40.2	33.8	28.1	–

(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	29.9	0.444	29.9	0.444
Opposite	en→bn	21.8	0.313	29.4	0.414
Related src	hi→en	30.1	0.449	30.5	0.460
Related src (WMT)	hi→en	29.1	0.422	29.1	0.427
HR unrelated src	fr→en	17.2	0.315	29.7	0.396
HR unrelated src	fr→ar	8.4	-0.102	28.0	0.322

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

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.

5.5 Use of Linguistic Context

1-shot example			en→fr		fr→en	
Origin	Dir.	Trunc.	BLEU	COMET	BLEU	COMET
Rand.	rand.	×	5.7	0.342	12.1	0.614
		✓	37.6	0.634	41.4	0.758
Prev.	rand.	×	6.1	0.328	12.3	0.617
		✓	38.5	0.614	41.6	0.751
Prev.	same	×	19.3	0.597	20.7	0.719
		✓	**39.0**	**0.632**	**42.1**	**0.761**
Prev.	opp.	×	3.6	0.064	8.6	0.518
		✓	37.8	0.590	41.2	0.742

Table 10: Comparison of 1-shot results (BLEU) for DiaBLa when using the previous/random sentence for the 1-shot example (using the `xglm-source+target` prompt). In bold are the best results for each language direction.

There has been a considerable amount of research on linguistic context in MT, e.g. to disambiguate lexically ambiguous texts or when additional information 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 (38.5 vs. 37.6), whereas COMET is lower (0.328 vs. 0.342). 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. For evidence of this, see Table 19 in Appendix F, which 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 was made possible with the collective efforts of the BigScience community, who designed, developed and prepared the tools and datasets used to train BLOOM. Special mention to evaluation working group members and especially to Niklas Muenninghoff and Pawan Sasanka Ammanamanchi for producing some of our results.

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 (GENCI). R. 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 by her Emergence project, DadaNMT, funded by Sorbonne Université.

References

Bach, Stephen, Victor Sanh, Zheng Xin Yong, , [. . .], and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proc. of the ACL: System Demonstrations*.

Bapna, Ankur, Isaac Caswell, Julia Kreutzer, [. . .], and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *CoRR*, abs/2205.03983.

Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proc. of NAACL-HLT*.

Bawden, Rachel, 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).

Bi, Bin, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020. PALM: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. In *Proc. of EMNLP*.

BigScience, Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, [. . .], and Thomas Wolf. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Blevins, Terra and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proc. of EMNLP*.

Bojar, Ondřej, Christian Buck, Christian Federmann, [. . .], and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proc. of WMT*.

Brown, Tom, Benjamin Mann, Nick Ryder, [. . .], and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*, volume 33.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, [. . .], and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *CoRR*, abs/2204.02311.

Costa-jussà, Marta R., James Cross, Onur Çelebi, [. . .], and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Dong, Daxiang, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proc. of ACL-IJCNLP*.

Fan, Angela, Shruti Bhosale, Holger Schwenk, [. . .], and Armand Joulin. 2021. Beyond English-Centric multilingual machine translation. *Journal of Mach. Learn. Research*, 22(107).

Firat, Orhan, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proc. of NAACL-HLT*.

Gao, Leo, Jonathan Tow, Stella Biderman, [. . .], and Andy Zou. 2021. A framework for few-shot language model evaluation.

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

Garcia, Xavier, 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.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, [. . .], and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. of the ACL*, 10.

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

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, [. . .], and Hany Hassan. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *CoRR*, abs/2302.09210.

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a Good Translator? A Preliminary Study. *CoRR*, abs/2301.08745.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proc. of the EACL*.

Laurençon, Hugo, Lucile Saulnier, Thomas Wang, [. . .], and Yacine Jernite. 2022. The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Proc. of NeurIPS: Datasets and Benchmarks Track*.

Lester, Brian, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proc. of EMNLP*.

Libovický, Jindřich and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proc. of the ACL*.

Lin, Xi Victoria, Todor Mihaylov, Mikel Artetxe, [. . .], and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proc. of EMNLP*.

Liu, Pengfei, 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).

Lopes, António, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proc. of the ACL*.

Luong, Minh-Thang, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proc. of ICLR*.

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

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

Muennighoff, Niklas, Thomas Wang, Lintang Sutawika, [. . .], and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *CoRR*, abs/2211.01786.

Nayak, Prashanth, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Investigating Contextual Influence in Document-Level Translation. *Information*, 13(5).

Ortiz Suárez, Pedro Javier, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proc. of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the ACL*.

Petroni, Fabio, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proc. of EMNLP-IJCNLP*.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proc. of WMT*.

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

Raffel, Colin, 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. *Journal of Mach. Learn. Research*, 21(140).

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proc. of EMNLP*.

Sanh, Victor, Albert Webson, Colin Raffel, [. . .], and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *Proc. of ICLR*.

Schick, Timo and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proc. of NAACL-HLT*.

Shin, Taylor, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proc. of EMNLP*.

Soltan, Saleh, Shankar Ananthakrishnan, Jack FitzGerald, [. . .], and Prem Natarajan. 2022. AlexaTM 20B: few-shot learning using a large-scale multilingual seq2seq model. *CoRR*, abs/2208.01448.

Vilar, David, 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.

Voita, Elena, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proc. of the ACL*.

Wei, Jason, 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 *Proc. of NeurIPS*.

Zeng, Aohan, Xiao Liu, Zhengxiao Du, [. . .], and Jie Tang. 2022. GLM-130B: an open bilingual pre-trained model. *CoRR*, abs/2210.02414.

Zhang, Susan, Stephen Roller, Naman Goyal, [. . .], and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *CoRR*, abs/2205.01068.

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

A COMET Results for Main Comparison

Table 11 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	OPT	BLOOM	T0	mT0	OPT
WMT 2014								
en→fr	-0.985	-0.700	0.453	-0.919	0.085	-1.035	-0.015	-0.165
fr→en	-0.675	0.337	0.567	-0.493	0.448	-0.087	0.250	0.039
en→hi	-0.482	-1.819	0.484	-1.525	0.288	-1.733	0.026	-1.460
hi→en	-0.387	-1.346	0.514	-1.200	0.378	-1.624	-0.019	-1.290
DiaBLa								
en→fr	-1.573	-0.528	0.380	-1.762	0.342	-0.585	-0.018	0.123
fr→en	-1.581	0.228	0.534	-1.507	0.614	-0.032	0.365	0.389
Flores-101								
en→fr	-1.469	-0.682	0.797	-1.438	0.602	-0.983	0.605	0.130
fr→en	-1.143	0.499	0.833	-1.008	0.687	-0.081	0.706	0.404
en→hi	-0.972	-1.848	1.025	-1.699	0.454	-1.795	0.718	-1.622
hi→en	-0.339	-1.391	0.797	-1.493	0.538	-1.264	0.667	-1.263

(a) Original predictions

	0-shot				1-shot			
	BLOOM	T0	mT0	OPT	BLOOM	T0	mT0	OPT
WMT 2014								
en→fr	0.434	-0.700	0.452	0.034	0.424	-1.035	-0.017	-0.000
fr→en	0.604	0.336	0.566	0.534	0.532	-0.090	0.247	0.449
en→hi	0.053	-1.819	0.483	-1.491	0.448	-1.733	0.026	-1.460
hi→en	0.445	-1.346	0.511	-1.113	0.386	-1.624	-0.022	-1.274
DiaBLa								
en→fr	0.433	-0.528	0.380	-0.002	0.634	-0.585	-0.023	0.192
fr→en	0.567	0.228	0.534	0.554	0.758	-0.039	0.356	0.639
Flores-101								
en→fr	0.182	-0.683	0.793	0.027	0.622	-0.984	0.601	0.180
fr→en	0.697	0.499	0.831	0.689	0.690	-0.086	0.702	0.594
en→hi	-0.608	-1.849	1.025	-1.638	0.461	-1.795	0.718	-1.622
hi→en	0.509	-1.391	0.797	-1.166	0.538	-1.264	0.666	-1.251

(b) Truncated predictions

Table 11: Comparison of COMET scores across the three datasets using the `xglm-source+target` prompt.

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 12 and 13 give the full analysis including raw figures for language identification for WMT14 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).

	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 12: 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).

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 14 and 15, 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 (looking at the range of scores) in the 1-shot setting for all models.

	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 13: 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).

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 16 and 17. 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 18). 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”.

F DiaBLa context-use examples

Table 19 contains examples where the preceding context in 1-shot examples has a positive, negative or neutral influence on the current prediction, showing that the choice of the 1-shot example is important and is taken into account by the model. Some details of these experiments are found in the accompanying Section 5.5 in the main text.

Model / Direction	0-shot en→fr	0-shot fr→en	1-shot en→fr	1-shot fr→en
BLOOM	**11.2** 3.0 – 22.0	**15.4** 10.3 – 26.8	**32.6** 27.8 – 36.4	**34.9** 33.1 – 36.6
BLOOM-7b1	6.5 1.5 – 12.1	12.8 4.8 – 25.1	25.9 20.8 – 29.9	29.1 25.4 – 32.5
BLOOM-3b	3.6 1.2 – 9.6	10.6 2.8 – 19.3	21.6 16.7 – 26.8	25.7 18.6 – 29.6
BLOOM-1b1	1.7 0.5– 3.9	7.1 0.7 – 11.4	10.1 6.3 – 13.2	16.1 12.2 – 19.9
BLOOM-560m	0.6 0.4 – 0.9	3.7 1.4 – 5.4	3.6 2.2 – 4.4	8.6 5.8 – 12.1

Table 14: 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	**2.1** 0.3 – 6.8	**8.3** 0.7 – 13.0	**12.9** 6.5 – 14.6	**19.8** 10.0 – 25.8
BLOOM-7b1	0.1 0.1 – 3.0	5.7 0.3 – 9.5	5.9 0.3 – 10.4	12.4 1.0 – 17.5
BLOOM-3b	0.2 0.0 – 0.5	3.6 0.0 – 7.0	4.9 0.2 – 7.2	8.9 0.1 – 13.5
BLOOM-1b1	0.1 0.0 – 0.1	1.5 0.0 – 4.5	1.4 0.1 – 3.1	4.6 0.00 – 8.2
BLOOM-560m	0.1 0.0 – 0.1	0.8 0.0 – 1.7	0.2 0.0 – 0.3	1.5 0.1 – 2.8

Table 15: 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`	8.5 0.7–17.0	19.1 4.32–37.12	16.4 7.5–22.2	26.0 12.0–37.0
`a_good_translation-target`	4.6 0.6–13.9	20.9 3.4–36.8	21.7 6.6–35.2	26.31 12.5–36.9
`gpt3-target`	4.0 0.7–14.0	18.7 3.0–36.4	8.3 1.3–25.7	21.6 7.2–37.2
`translate_as-target`	6.4 0.6–10.1	18.1 3.5–33.1	11.5 2.3–20.4	22.9 8.2–35.7
`version-target`	9.7 0.7–30.3	21.9 4.4–36.7	22.2 4.7–35.2	25.3 8.0–37.2
`xglm-source+target`	**17.2** 1.33–32.2	**23.2** 5.0–36.3	**25.6** 8.3–37.2	**26.7** 11.1–38.2
`xglm-target`	2.5 1.1–4.6	20.1 6.8–33.1	11.0 4.5–17.6	23.1 10.4–36.4

Table 16: 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`	1.2 0.1–3.3	5.8 0.3–14.5	6.2 1.0–12.7	13.0 2.6–24.4
`a_good_translation-target`	0.4 0.1–1.3	5.5 0.3–14.1	10.8 1.1–25.4	13.2 2.7–24.7
`gpt3-target`	0.0 0.0–0.1	1.6 0.0–7.6	0.0 0.0–0.0	2.5 0.0–11.4
`version-target`	1.0 0.1–3.0	5.5 0.2–13.9	**11.3** 2.4–21.4	**13.5** 2.7–25.7
`xglm-source+target`	**3.9** 0.1–12.1	**7.3** 0.2–15.8	8.8 0.9–24.3	12.4 1.2–25.0
`xglm-target`	0.3 0.0–1.0	5.1 0.0–14.5	2.1 0.3–5.8	6.5 0.1–13.0

Table 17: 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 18: 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.

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 19: 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.