Task	Template	Candidate Verbalizer
XCOPA	*cause*: `{Sentence 1}` because `[Mask]` *effect*: `{Sentence 1}` therefore `[Mask]`	Identity
XStoryCloze	`{Context}` `[Mask]`	Identity
XNLI	`{Sentence 1}`, right? `[Mask]`, `{Sentence 2}`	*Entailment*: Yes | *Neutral*: Also | *Contradiction*: No
PAWS-X	`{Sentence 1}`, right? `[Mask]`, `{Sentence 2}`	*True*: Yes | *False*: No
MGSM	Question: `{Question}` Step-by-Step Answer:	None

Table 2: **Handcrafted English prompts for multilingual tasks.** The identity function maps each candidate choice to itself. In the case of MGSM there is no verbalizer, because the model generates an answer that is extracted with a regular expression.

B Additional results

In this section, we report additional results that cover direct vs. self-translate, self-translate vs. MT, results by language and translation metrics.

B.1 Direct vs. self-translate

We include additional direct vs. self-translate results for **BLOOM** (Scao et al., 2023), **LLaMA 2** (Touvron et al., 2023b), **OpenLLaMA** (Geng and Liu, 2023), **OpenLLaMA V2** (Geng and Liu, 2023), **Redpajama** (Computer, 2023) and **PolyLM** (Wei et al., 2023). Similar to XGLM, BLOOM has a multilingual focus and covers many languages. The rest of the models are similar to LLaMA, which is primarily trained on English and is much stronger in this language, while also showing some multilingual capabilities. Table 3 shows the results as accuracy of the **direct** and **self-translate** methods in all tasks for different models and sizes. Results resemble the ones obtained by XGLM and LLaMA in the main results, so we can conclude that self-translate is consistent across different models.

B.2 Self-translate vs. MT

We include additional self-translate vs. MT results for **XGLM** (Lin et al., 2022) and **LLaMA** (Touvron et al., 2023a). Table 4 shows task accuracy for different sizes of these models, using **self-translate** inference and **MT**. The last column shows the average accuracy over all tasks.

B.3 Results by language

We include additional language results for **XGLM** (Lin et al., 2022) and **LLaMA** (Touvron et al., 2023a). Tables 5 to 9 show the results by language in different tasks, using different model sizes and the **direct** inference, **self-translate**, and **MT** methods. The last column shows the average accuracy over all languages except English.

B.4 Translation metrics

We obtain similar results with BLEU (Papineni et al., 2002) and COMET (Rei et al., 2022) metrics. We report the average COMET and BLEU scores across all languages for NLLB, XGLM, BLOOM and LLaMA in Tables 10 and 11.

B.5 Translation metrics by language

We report NLLB, XGLM, BLOOM and LLaMA COMET metrics for each language and task in Tables 12 to 16, and BLEU metrics in Tables 17 to 21.

Model	Size	Method	XStoryC	XCOPA	XNLI	PAWS-X	MGSM	Avg
BLOOM	0.6B	Direct	**52.9**	**54.0**	36.6	**49.3**	**1.7**	38.9
		Self-translate	**52.9**	51.0	**41.4**	48.4	1.5	**39.0**
	1.7B	Direct	55.2	**55.1**	39.2	47.0	**2.3**	39.8
		Self-translate	**55.5**	54.7	**41.9**	**48.0**	1.8	**40.4**
	3.0B	Direct	56.4	56.1	39.8	49.4	2.0	40.7
		Self-translate	**57.2**	**56.7**	**44.1**	**52.1**	**2.1**	**42.4**
	7.1B	Direct	58.2	56.9	40.7	50.2	**3.2**	41.8
		Self-translate	**59.3**	**59.7**	**45.4**	**54.4**	3.1	**44.4**
LLaMA 2	7B	Direct	55.6	56.7	39.2	57.9	1.8	42.2
		Self-translate	**57.8**	**59.3**	**47.6**	**61.3**	**7.2**	**46.6**
	13B	Direct	57.2	58.2	39.8	52.4	13.2	44.2
		Self-translate	**59.9**	**61.3**	**46.0**	**55.2**	**19.2**	**48.3**
RedPajama	3B	Direct	51.4	53.0	36.3	52.6	1.1	38.9
		Self-translate	**52.3**	**53.1**	**41.8**	**56.8**	**1.4**	**41.1**
	7B	Direct	53.3	52.5	38.2	54.5	2.0	40.1
		Self-translate	**53.9**	**55.2**	**42.6**	**57.4**	**3.2**	**42.5**
OpenLLaMA	3B	Direct	51.0	52.4	35.7	48.4	1.1	37.7
		Self-translate	**53.4**	**52.5**	**39.7**	**53.1**	**1.9**	**40.1**
	7B	Direct	52.4	52.9	37.0	51.8	1.9	39.2
		Self-translate	**55.5**	**53.9**	**43.1**	**56.9**	**3.6**	**42.6**
	13B	Direct	53.8	54.0	38.6	52.7	3.5	40.5
		Self-translate	**55.4**	**56.0**	**44.2**	**58.0**	**5.3**	**43.8**
OpenLLaMA V2	3B	Direct	52.2	53.7	36.8	49.0	2.2	38.8
		Self-translate	**54.5**	**55.6**	**43.4**	**52.8**	**3.0**	**41.9**
	7B	Direct	53.9	54.4	38.2	52.3	3.6	40.5
		Self-translate	**55.7**	**56.9**	**44.6**	**56.2**	**5.7**	**43.8**
PolyLM	1.7B	Direct	51.8	**54.3**	37.4	48.2	1.4	38.6
		Self-translate	**52.6**	53.2	**40.6**	**49.4**	**1.6**	**39.5**
	13B	Direct	56.3	58.9	41.4	55.0	4.4	43.2
		Self-translate	**57.4**	**60.4**	**45.6**	**57.3**	**5.3**	**45.2**

Table 3: **Direct vs. self-translate.** Task accuracy for different sizes of BLOOM, OpenLLaMA, OpenLLaMA V2, Redpajama and PolyLM, using direct inference and self-translate. The last column shows the average accuracy over all tasks. We highlight the best results for each model and task in bold.

Model	Size	Method	XStoryC	XCOPA	XNLI	PAWS-X	MGSM	Avg
XGLM	0.6B	Self-translate	52.8	53.4	41.5	50.6	**1.4**	39.9
		MT	**57.3**	**59.8**	**46.3**	**51.7**	1.1	**43.2**
	1.7B	Self-translate	55.9	58.4	44.9	50.2	1.7	42.2
		MT	**60.7**	**62.3**	**47.4**	**51.2**	**2.3**	**44.8**
	2.9B	Self-translate	58.2	62.5	46.2	53.2	1.6	44.3
		MT	**62.3**	**65.3**	**48.8**	**55.7**	**2.2**	**46.9**
	7.5B	Self-translate	60.9	64.4	48.9	55.4	**0.1**	45.9
		MT	**63.6**	**66.3**	**50.7**	**57.4**	0.0	**47.6**
LLaMA	7B	Self-translate	55.8	54.9	43.0	57.0	6.1	43.4
		MT	**66.8**	**68.6**	**48.6**	**58.8**	**10.7**	**50.7**
	13B	Self-translate	57.7	56.5	**35.1**	52.1	10.0	42.3
		MT	**68.1**	**70.4**	**35.1**	**54.2**	**16.5**	**48.9**
	30B	Self-translate	59.0	58.4	43.5	55.6	16.3	46.6
		MT	**68.7**	**71.5**	**46.1**	**55.9**	**28.6**	**54.2**

Table 4: **Self-translate vs. MT.** Task accuracy for different sizes of XGLM and LLaMA, using self-translate and MT. The last column shows the average accuracy over all tasks. We highlight the best results for each model and task in bold.

Model	Size	Method	ar	en	es	eu	hi	id	my	ru	sw	te	zh	avg
XGLM	0.6B	Direct	50.1	60.6	55.1	53.1	52.3	54.0	51.5	56.2	53.1	55.9	53.3	53.5
		Self-translate	52.2	_	53.1	54.0	53.5	53.6	52.3	53.9	52.1	53.0	50.0	52.8
		MT	58.1	_	57.2	55.7	57.4	57.9	55.2	58.8	56.5	59.5	56.8	57.3
	1.7B	Direct	52.5	64.3	59.2	56.1	55.8	58.0	53.8	59.8	56.0	58.0	56.2	56.5
		Self-translate	55.4	_	58.4	54.3	55.1	57.1	55.5	58.4	55.3	54.8	54.9	55.9
		MT	61.9	_	60.4	58.3	61.7	61.4	57.8	62.7	60.0	61.3	61.6	60.7
	2.9B	Direct	53.9	67.3	61.0	56.3	57.5	61.4	55.2	62.2	56.7	60.0	57.6	58.2
		Self-translate	56.3	_	61.3	56.9	58.3	60.4	57.6	59.7	57.9	56.3	57.8	58.2
		MT	63.0	_	63.2	61.2	63.3	62.9	58.8	64.7	60.0	62.8	63.0	62.3
	7.5B	Direct	56.2	69.8	64.1	57.7	58.8	62.9	57.1	63.5	59.3	60.2	58.9	59.9
		Self-translate	60.7	_	63.8	59.8	61.3	62.9	57.8	64.4	60.0	57.6	60.4	60.9
		MT	64.3	_	64.7	63.1	64.9	63.4	60.3	65.9	61.4	63.3	65.0	63.6
LLaMA	7B	Direct	48.3	74.8	65.1	50.1	52.7	52.1	48.7	61.4	50.4	52.9	54.3	53.6
		Self-translate	52.2	_	68.0	50.0	51.9	56.5	50.2	66.8	50.6	51.4	60.4	55.8
		MT	67.7	_	68.4	65.4	68.5	68.3	62.5	70.1	64.3	65.5	67.2	66.8
	13B	Direct	49.7	77.3	69.4	50.7	52.3	55.3	47.8	63.4	49.9	53.3	56.5	54.8
		Self-translate	55.2	_	72.1	50.8	53.7	59.3	51.8	70.4	48.4	51.8	63.2	57.7
		MT	68.6	_	70.0	66.4	70.0	69.0	62.8	71.7	66.0	67.7	69.1	68.1
	30B	Direct	50.9	78.2	70.8	51.4	56.7	59.2	48.8	66.7	50.6	53.2	58.6	56.7
		Self-translate	56.4	_	74.0	48.8	60.2	62.6	51.0	71.4	48.9	49.9	67.0	59.0
		MT	70.0	_	71.5	66.6	70.0	69.3	63.6	73.3	67.0	66.9	69.0	68.7

Table 5: **XGLM and LLaMA results on XStoryCloze for each language.** We show task accuracy for different sizes of these models, using **direct** inference **self-translate** and **MT**. The last column shows the average accuracy over all languages except English.

Model	Size	Method	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
XGLM	0.6B	Direct	55.6	55.0	57.2	53.8	49.2	53.2	56.2	55.2	54.4	58.4	55.6	54.9
		Self-translate	52.2	54.2	59.4	51.8	50.0	52.6	55.0	55.2	55.2	51.8	50.4	53.4
		MT	60.0	61.0	60.4	61.8	50.4	59.4	61.6	58.8	62.4	61.8	60.2	59.8
	1.7B	Direct	56.8	55.8	64.6	54.0	52.2	56.6	55.2	58.2	53.4	63.0	58.0	57.1
		Self-translate	59.0	57.0	60.6	60.0	50.8	57.8	58.8	58.4	60.8	61.0	58.4	58.4
		MT	65.6	62.8	63.4	65.6	50.4	62.2	63.8	61.0	63.8	64.0	62.6	62.3
	2.9B	Direct	58.2	55.8	66.8	60.2	50.2	58.8	54.2	57.0	56.6	65.2	60.0	58.5
		Self-translate	64.4	65.2	64.8	64.2	52.0	62.2	59.4	60.8	62.0	65.4	67.4	62.5
		MT	69.2	65.4	67.2	70.8	51.0	64.8	65.2	64.0	66.4	67.2	67.0	65.3
	7.5B	Direct	61.2	57.4	69.4	63.6	48.8	60.0	54.4	59.4	58.4	70.2	63.8	60.6
		Self-translate	66.8	64.6	66.8	68.4	51.0	62.8	65.6	62.8	65.4	65.2	68.6	64.4
		MT	71.8	64.8	67.6	72.8	50.4	66.8	67.4	62.0	69.8	68.6	67.6	66.3
LLaMA	7B	Direct	48.8	51.0	54.6	62.0	51.4	50.8	55.2	55.8	55.6	51.6	56.2	53.9
		Self-translate	54.2	51.2	59.4	73.8	48.4	52.8	47.6	50.8	51.6	47.8	66.0	54.9
		MT	72.6	68.2	71.0	75.4	52.2	67.4	70.2	62.2	72.6	71.2	71.6	68.6
	13B	Direct	48.2	52.8	57.8	67.2	50.2	51.2	54.4	54.6	53.0	53.8	58.4	54.7
		Self-translate	51.8	51.4	62.8	75.8	51.6	49.4	51.2	51.4	56.6	49.2	69.8	56.5
		MT	73.2	70.0	72.8	76.8	51.6	70.2	71.8	64.8	73.2	75.2	75.2	70.4
	30B	Direct	47.2	51.8	60.6	71.4	49.4	52.4	53.2	54.6	52.2	52.4	62.2	55.2
		Self-translate	50.4	53.0	68.0	79.0	49.4	50.2	52.8	48.6	59.8	58.4	73.2	58.4
		MT	75.2	71.2	73.2	80.6	52.6	70.6	72.2	64.6	74.2	75.0	76.8	71.5

Table 6: **XGLM and LLaMA results on XCOPA for each language.** We show task accuracy for different sizes of these models, using **direct** inference **self-translate** and **MT**. The last column shows the average accuracy over all languages.