Model	Size	Method	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
XGLM	0.6B	Direct	33.4	41.3	44.5	39.6	48.3	42.0	45.5	38.7	44.6	36.1	38.8	40.2	34.5	38.5	33.5	39.4
		Self-translate	40.2	43.9	43.9	42.2	_	43.3	43.3	41.4	43.0	39.0	41.9	40.6	40.6	41.5	35.8	41.5
		MT	46.9	47.1	46.6	46.6	_	47.5	46.5	45.6	45.7	45.6	46.3	46.4	43.8	46.8	47.1	46.3
	1.7B	Direct	33.5	44.7	45.3	40.1	49.7	43.6	45.7	42.6	46.0	42.0	41.7	43.0	39.5	45.0	33.8	41.9
		Self-translate	44.2	46.8	47.0	46.1	_	45.9	46.8	44.1	45.7	43.8	44.0	42.7	42.0	44.7	44.3	44.9
		MT	47.3	47.8	48.8	48.1	_	48.5	48.6	47.1	47.2	45.9	46.5	48.3	44.2	48.6	47.3	47.4
	2.9B	Direct	33.7	46.0	48.3	41.4	51.1	46.7	45.0	44.0	45.3	44.4	42.0	45.0	40.1	46.0	34.8	43.0
		Self-translate	43.9	48.1	48.4	47.3	_	48.2	48.5	44.1	46.5	44.8	45.8	45.2	42.4	46.6	46.7	46.2
		MT	48.9	49.5	50.0	49.4	_	50.5	50.0	48.5	47.9	47.7	47.5	48.6	45.4	49.6	49.0	48.8
	7.5B	Direct	33.4	44.9	49.0	40.7	53.9	47.7	46.9	47.2	46.3	45.8	43.7	46.3	42.1	46.3	35.4	44.0
		Self-translate	47.0	51.6	50.4	50.7	_	51.8	51.6	46.8	50.0	47.3	47.4	47.5	44.5	48.9	48.6	48.9
		MT	50.6	51.8	51.8	51.6	_	52.8	52.1	51.0	50.5	48.7	48.6	51.8	46.9	50.2	51.2	50.7
LLaMA	7B	Direct	33.6	37.0	44.8	34.9	51.1	40.6	43.8	36.1	39.4	33.7	34.5	35.6	33.4	35.6	36.2	37.1
		Self-translate	40.7	48.7	50.6	43.5	_	49.8	49.5	39.7	48.0	34.8	36.3	38.0	36.4	39.9	46.1	43.0
		MT	48.6	49.3	49.9	50.1	_	50.4	50.1	48.5	48.3	46.5	46.4	48.0	45.5	49.2	49.3	48.6
	13B	Direct	34.1	34.1	35.3	34.8	35.7	33.4	33.4	35.5	34.1	33.0	34.5	34.0	34.3	34.0	34.4	34.2
		Self-translate	35.3	34.7	35.3	35.1	_	36.0	35.8	35.4	35.0	34.9	34.8	34.6	34.9	35.4	34.4	35.1
		MT	34.1	35.3	35.3	35.5	_	35.2	35.2	35.3	35.3	35.2	34.1	34.6	35.0	34.8	36.1	35.1
	30B	Direct	34.4	38.6	44.0	35.1	47.9	40.4	42.9	36.6	38.2	34.2	34.0	36.3	34.3	35.6	33.6	37.0
		Self-translate	42.2	47.6	47.7	44.8	_	48.1	47.8	41.4	47.3	37.3	37.4	42.0	38.9	41.6	44.3	43.5
		MT	46.2	46.4	47.3	46.9	_	47.7	47.4	45.7	46.3	44.8	45.0	45.3	43.8	46.5	46.6	46.1

Table 7: **XGLM and LLaMA results on XNLI 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	de	en	es	fr	ja	ko	zh	avg
XGLM	0.6B	Direct	49.1	50.6	52.5	50.8	44.1	46.2	47.8	48.4
		Self-translate	51.1	_	50.1	50.3	50.9	50.4	51.0	50.6
		MT	53.5	_	52.8	51.0	51.2	50.4	51.2	51.7
	1.7B	Direct	57.6	52.6	53.8	47.3	46.1	51.4	48.1	50.7
		Self-translate	50.0	_	51.6	51.6	49.6	49.1	49.4	50.2
		MT	51.9	_	51.6	52.8	50.2	51.1	49.5	51.2
	2.9B	Direct	50.6	54.8	53.1	49.7	50.9	46.8	53.7	50.8
		Self-translate	54.9	_	53.9	54.2	52.1	51.6	52.7	53.2
		MT	56.5	_	57.0	56.2	54.8	54.5	55.4	55.7
	7.5B	Direct	55.9	58.9	52.8	51.8	52.0	46.0	51.3	51.6
		Self-translate	57.7	_	56.1	56.1	54.5	53.0	54.9	55.4
		MT	59.6	_	58.4	59.0	54.6	55.2	57.7	57.4
LLaMA	7B	Direct	54.6	61.9	56.1	52.9	56.7	49.7	49.1	53.2
		Self-translate	59.8	_	60.7	59.2	53.9	52.5	55.8	57.0
		MT	59.9	_	60.6	60.1	57.6	57.5	57.3	58.8
	13B	Direct	52.9	53.1	52.4	54.6	45.0	46.9	45.2	49.5
		Self-translate	52.9	_	52.5	52.9	51.2	51.6	51.5	52.1
		MT	53.6	_	54.4	53.8	55.3	54.4	53.8	54.2
	30B	Direct	58.4	58.5	56.0	52.5	46.6	45.6	46.2	50.9
		Self-translate	56.5	_	56.8	58.1	54.5	52.1	55.5	55.6
		MT	56.6	_	57.8	56.9	55.1	54.8	54.2	55.9

Table 8: **XGLM and LLaMA results on PAWS-X 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	bn	de	en	es	fr	ja	ru	sw	te	th	zh	avg
XGLM	0.6B	Direct	1.2	0.8	2.0	1.2	1.6	4.0	0.4	2.4	0.4	1.6	3.2	1.7
		Self-translate	0.0	2.0	_	2.0	1.6	0.8	1.2	2.0	2.4	0.8	1.6	1.4
		MT	1.2	1.2	_	0.8	0.8	2.0	1.6	1.2	0.4	1.6	0.0	1.1
	1.7B	Direct	0.8	1.2	2.0	2.4	2.0	1.6	0.8	1.2	2.0	2.0	2.8	1.7
		Self-translate	1.2	2.0	_	2.8	1.6	2.4	2.8	1.2	1.2	0.8	1.2	1.7
		MT	2.0	2.4	_	2.0	0.8	2.8	2.0	2.8	3.2	2.8	2.4	2.3
	2.9B	Direct	0.0	0.8	2.4	2.0	1.2	2.0	2.0	2.0	2.0	0.8	1.2	1.4
		Self-translate	0.8	1.2	_	1.6	1.6	1.6	1.2	2.0	1.2	2.4	2.0	1.6
		MT	2.8	2.4	_	2.8	2.4	1.2	1.6	2.0	3.2	0.8	2.4	2.2
	7.5B	Direct	0.0	1.2	0.0	0.0	0.0	0.4	2.4	0.4	1.2	1.6	1.2	0.8
		Self-translate	0.0	0.4	_	0.0	0.0	0.0	0.4	0.0	0.4	0.0	0.0	0.1
		MT	0.0	0.0	_	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0
LLaMA	7B	Direct	0.0	9.6	13.6	10.4	8.8	5.2	10.0	2.0	0.0	0.0	4.4	5.0
		Self-translate	2.0	11.2	_	11.2	12.4	4.8	10.8	1.2	0.4	2.4	4.8	6.1
		MT	10.0	12.4	_	12.0	9.6	10.8	10.8	12.0	9.6	8.4	11.2	10.7
	13B	Direct	0.0	16.0	20.8	15.2	15.6	5.2	10.0	3.6	0.0	0.0	8.8	7.4
		Self-translate	3.6	17.6	_	20.4	18.0	9.2	15.2	3.6	0.0	1.6	10.4	10.0
		MT	16.8	20.0	_	20.8	15.2	15.2	15.6	19.2	14.0	14.0	14.4	16.5
	30B	Direct	0.0	29.2	39.6	33.2	30.4	7.2	27.2	5.2	0.0	0.0	22.8	15.5
		Self-translate	8.0	34.4	_	9.6	24.4	20.8	29.6	6.4	0.4	3.6	25.6	16.3
		MT	28.4	32.4	_	31.2	35.2	29.2	26.4	32.0	25.6	20.0	25.6	28.6

Table 9: **XGLM and LLaMA results on MGSM 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	XStoryC	XCOPA	XNLI	PAWS-X	MGSM	Avg
NLLB	0.6B	86.9	80.3	84.6	85.4	80.2	83.5
	1.3B	88.2	82.9	85.6	86.0	83.8	85.3
	1.3B	88.3	82.1	85.5	86.0	83.5	85.1
	3.3B	88.7	83.3	85.9	86.2	84.5	85.7
XGLM	0.6B	63.4	61.3	66.2	66.0	54.7	62.3
	1.7B	77.1	74.1	75.8	75.9	68.4	74.3
	2.9B	81.1	77.6	78.5	79.2	73.5	78.0
	7.5B	84.2	79.8	81.7	81.6	79.2	81.3
BLOOM	0.6B	61.5	54.0	63.6	60.6	48.2	57.6
	1.7B	73.6	61.9	67.4	72.1	61.7	67.3
	3B	76.3	63.3	69.5	74.7	69.1	70.6
	7.1B	78.8	66.4	73.1	78.8	74.5	74.3
LLaMA	7B	66.8	59.4	71.5	80.9	66.0	68.9
	13B	68.8	61.8	75.0	82.6	69.6	71.6
	30B	71.7	65.0	78.4	83.8	67.5	73.3

Table 10: COMET translation metrics for different models.

Model	Size	XStoryC	XCOPA	XNLI	PAWS-X	MGSM	Avg
NLLB	0.6B	38.0	32.1	38.0	49.0	32.1	37.8
	1.3B	40.6	36.6	40.3	51.3	41.3	42.0
	1.3B	40.9	35.6	40.1	50.9	40.9	41.7
	3.3B	41.8	37.6	41.5	51.9	43.7	43.3
XGLM	0.6B	7.1	6.5	10.4	18.0	5.4	9.5
	1.7B	18.5	18.1	20.3	28.3	17.1	20.5
	2.9B	23.8	24.1	24.1	33.1	23.5	25.7
	7.5B	29.0	28.4	28.8	37.0	28.3	30.3
BLOOM	0.6B	7.9	4.8	11.8	16.2	5.4	9.2
	1.7B	17.3	10.5	14.9	27.2	12.6	16.5
	3B	20.2	13.0	17.1	31.1	20.3	20.3
	7.1B	25.2	16.5	21.4	36.1	27.7	25.4
LLaMA	7B	14.7	8.9	19.9	39.1	23.9	21.3
	13B	17.7	12.4	24.1	42.5	27.9	24.9
	30B	21.2	15.4	27.7	45.4	25.5	27.0

Table 11: BLEU translation metrics for different models.

Model	Size	ru	zh	es	ar	hi	id	te	sw	eu	my	avg
NLLB	0.6B	87.07	85.00	89.36	88.39	90.52	88.08	86.44	86.04	86.87	81.35	86.9
	1.3B	88.44	86.02	90.33	89.85	91.56	89.14	87.64	87.31	86.92	85.26	88.2
	1.3B	88.18	86.36	90.22	89.83	91.39	89.05	87.30	87.21	87.25	85.99	88.3
	3.3B	88.63	87.54	90.54	90.36	91.70	89.54	88.00	87.46	86.92	86.60	88.7
XGLM	0.6B	73.05	54.47	72.08	61.44	68.85	77.52	57.04	58.63	59.52	50.99	63.4
	1.7B	80.96	77.26	81.95	76.35	77.48	83.96	74.09	75.15	71.25	73.03	77.1
	2.9B	83.36	82.11	85.61	79.84	82.99	85.66	75.43	79.71	79.32	77.47	81.1
	7.5B	85.76	84.25	87.81	83.81	86.25	87.60	80.66	82.92	82.05	81.36	84.2
BLOOM	0.6B	43.20	70.47	73.65	72.18	73.40	79.31	58.06	42.03	55.73	47.25	61.5
	1.7B	60.47	82.81	85.44	80.40	81.05	85.06	72.48	66.06	71.98	50.69	73.6
	3B	63.44	84.45	87.16	82.20	83.16	85.72	75.11	71.03	76.99	53.68	76.3
	7.1B	68.97	86.63	88.42	84.68	86.76	87.87	78.86	75.15	80.88	49.80	78.8
LLaMA	7B	85.66	79.10	88.56	65.12	67.96	77.08	50.39	52.14	49.66	52.55	66.8
	13B	87.02	82.66	89.37	70.64	72.86	81.15	48.62	53.14	51.36	51.17	68.8
	30B	87.98	84.37	90.13	77.37	81.64	84.55	49.38	59.99	52.50	49.04	71.7

Table 12: XStoryCloze COMET translation metrics for different models.

Model	Size	et	ht	it	id	qu	sw	zh	ta	th	tr	vi	avg
NLLB	0.6B	82.78	75.42	86.49	85.23	62.17	79.74	84.66	83.93	76.30	84.54	81.97	80.3
	1.3B	86.57	78.88	88.95	87.44	64.26	82.01	87.07	86.50	78.79	86.97	84.29	82.9
	1.3B	85.38	77.84	88.50	86.86	62.97	81.43	86.44	85.79	77.72	86.31	83.55	82.1
	3.3B	86.76	79.16	89.16	87.56	63.87	82.08	87.85	86.60	80.10	87.42	85.23	83.3
XGLM	0.6B	68.27	58.08	65.79	73.98	34.54	54.72	50.21	64.52	71.24	64.44	68.33	61.3
	1.7B	78.78	67.84	79.09	81.47	50.98	69.01	80.06	77.22	77.88	74.84	77.87	74.1
	2.9B	83.16	71.97	82.96	84.22	50.82	74.41	83.93	79.67	81.37	78.98	82.23	77.6
	7.5B	85.49	72.47	85.19	86.04	55.33	77.29	85.41	83.47	82.36	81.38	83.61	79.8
BLOOM	0.6B	41.78	41.47	48.71	75.73	37.32	40.93	75.23	65.09	42.51	50.09	75.22	54.0
	1.7B	45.41	46.04	65.38	82.57	45.08	58.94	84.71	76.72	46.41	48.74	81.43	61.9
	3B	46.22	48.21	70.61	83.61	43.38	63.68	86.20	80.41	43.01	47.86	83.56	63.3
	7.1B	47.93	50.22	75.59	86.24	47.02	67.57	87.99	83.99	47.90	50.54	85.17	66.4
LLaMA	7B	51.26	48.89	85.89	70.59	49.65	50.03	80.04	49.16	53.79	59.32	54.76	59.4
	13B	52.17	49.01	87.22	75.13	48.00	50.14	83.16	49.02	58.65	67.93	59.71	61.8
	30B	55.41	52.29	88.42	79.85	48.48	54.73	85.10	52.96	59.66	71.51	66.20	65.0

Table 13: XCOPA COMET translation metrics for different models.