Few-shot Learning with Multilingual Generative Language Models

Meta AI

Presented by Abdullah Al Ishtiaq

Overview

- Motivation
- Model architecture and training
- Dataset
- Tasks
- Results
- Strengths
- Weaknesses

Motivation

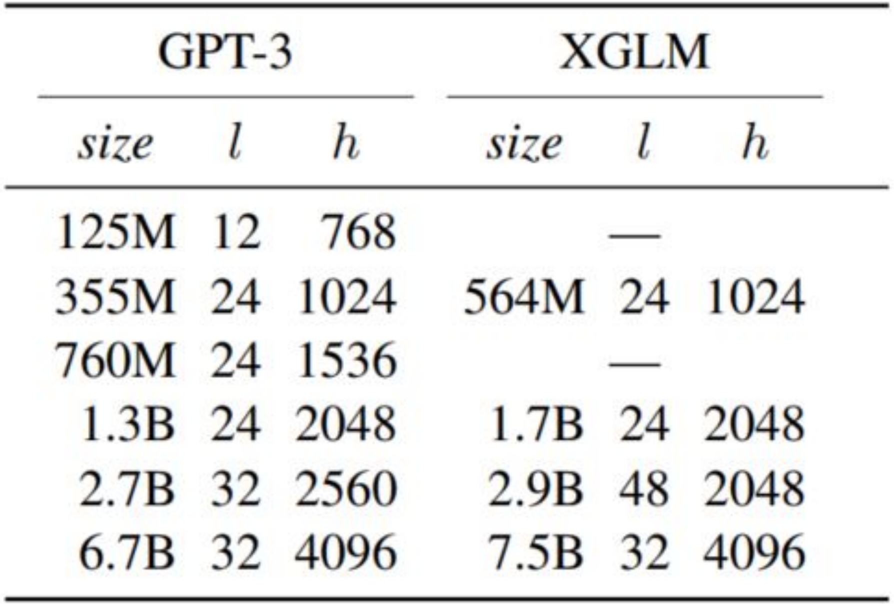
- Natural Language Research are, in many cases, dominated by high-resource languages like English
- Training data of Large Language Models (LLM), e.g., GPT-3, mostly include English (~93%)
- Performance on mid- and low-resource languages is not adequate
- Other multilingual models, e.g., mBERT, XLM-R, mT5, mBART, etc., require finetuning for downstream applications
- Multilingual few-shot learning capabilities of LLMs are not well studied

XGLM

- XGLM presents four multilingual generative models of different sizes
- Training corpus includes 30 diverse languages with 500B tokens
- Achieves state-of-the-art on diverse multilingual NLP tasks
    - Commonsense reasoning
    - Anaphora resolution
    - Natural language inference
    - Paraphrasing
    - Machine translation
- Comprehensively studies zero-shot and few-shot applications and prompt generation techniques

XGLM Models

- Decoder-only Causal Language Model (CLM)
- Transformer architecture similar to GPT-3
- Four models with 564M, 1.7B, 2.9B and 7.5B parameters
- 256 A100 GPUs for about 3 weeks



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

Fig: Model details. l: layers, h: hidden dim

Data

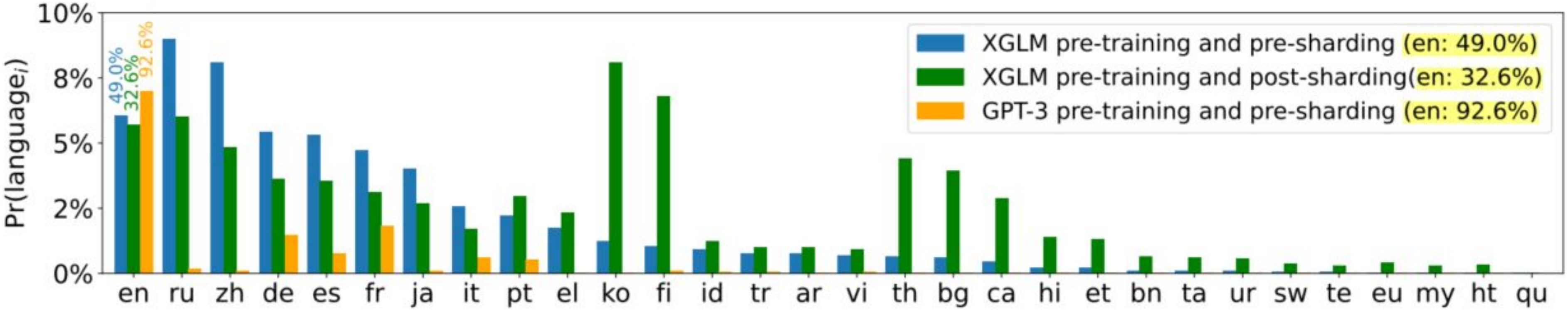
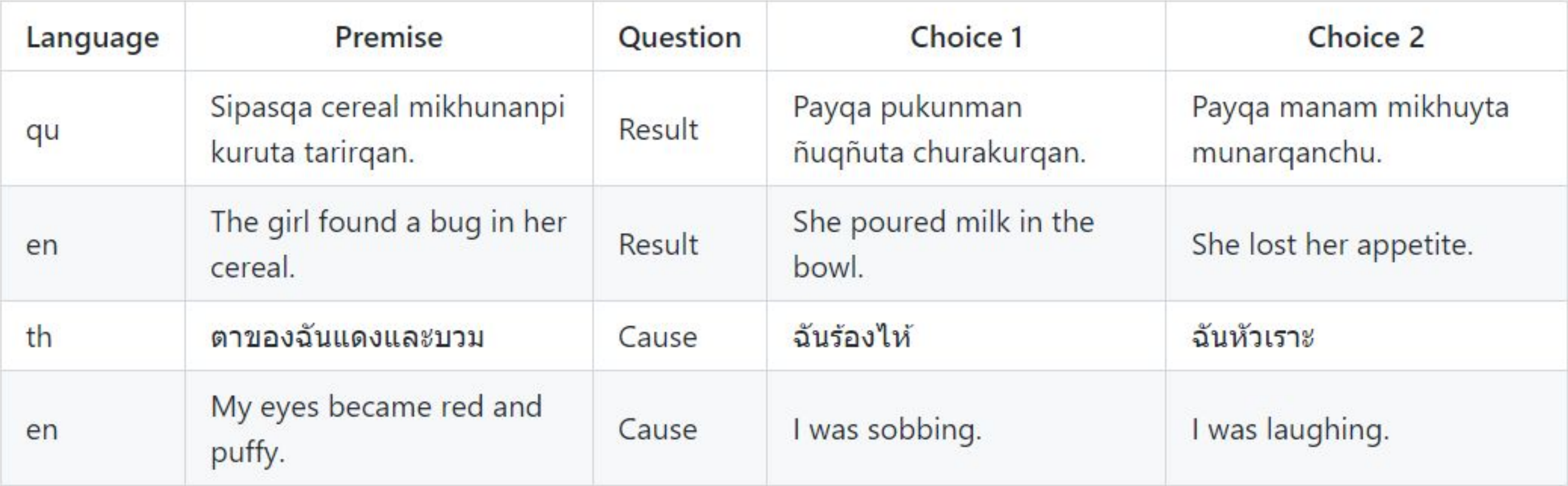


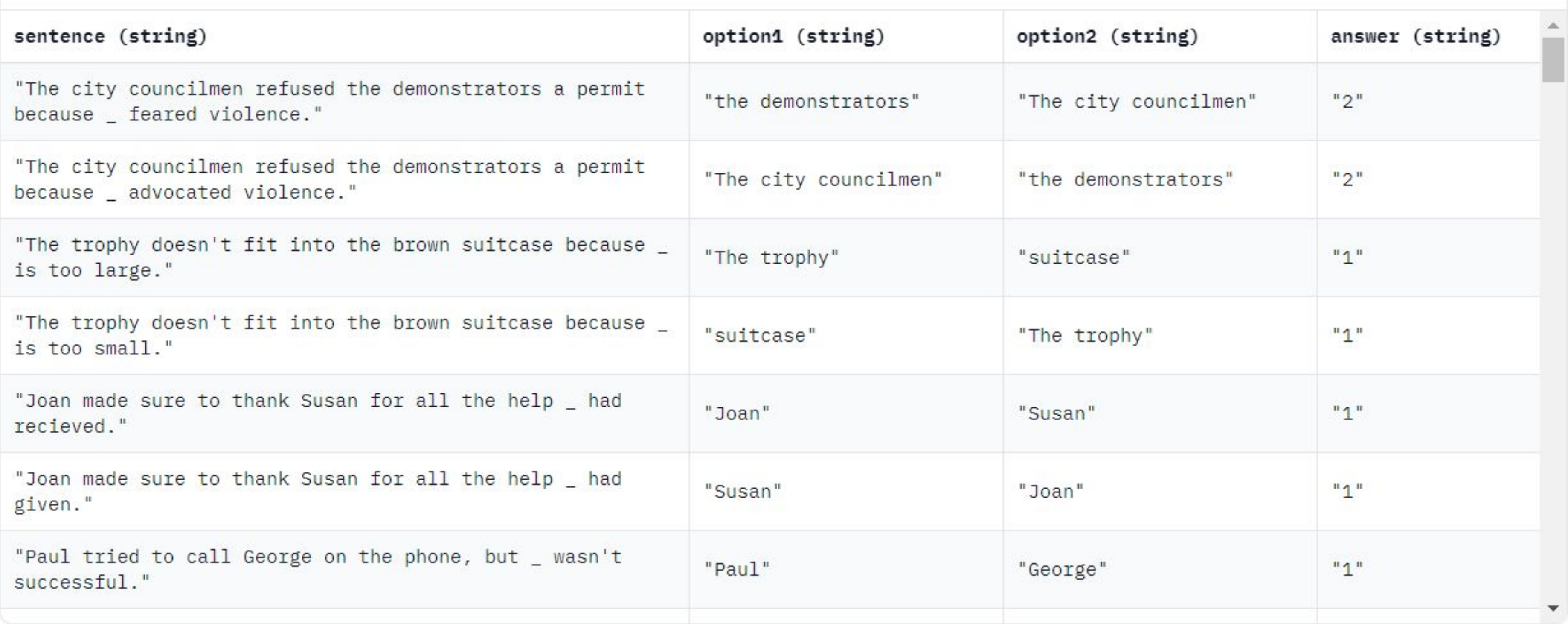
Figure 1: The % of each language l ($l = 1, 2, ..., 30$) in XGLM's pre-training data pre-upsampling (blue), post-upsampling (green), and its corresponding % in GPT-3's training data (orange). We truncate the y-axis at 10% to better visualize the tail distribution.

Tasks: Commonsense reasoning (XCOPA) [1]



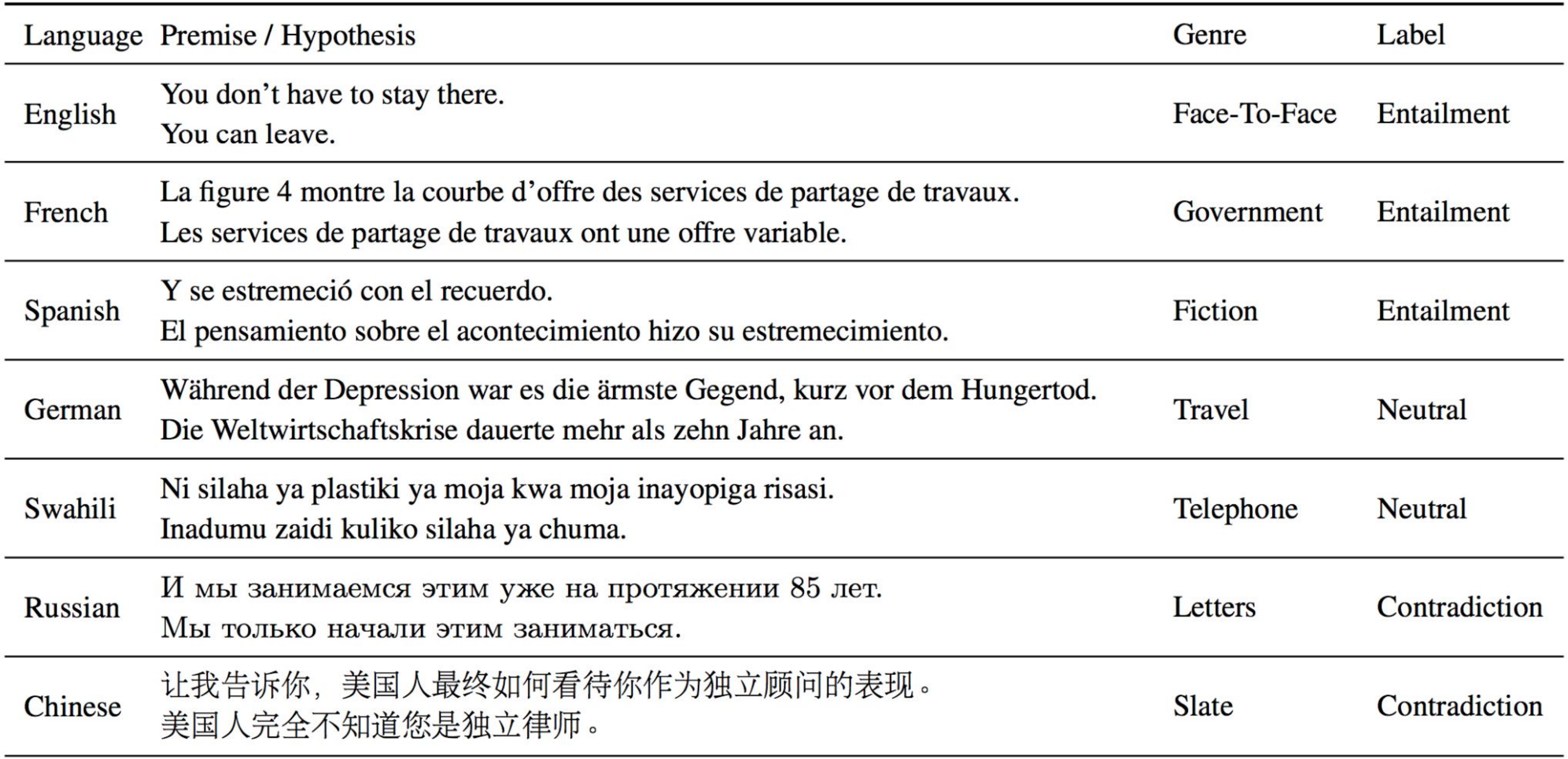
Language	Premise	Question	Choice 1	Choice 2
qu	Sipasqa cereal mikhunanpi kuruta tarirqan.	Result	Payqa pukunman ñuqñuta churakurqan.	Payqa manam mikhuyta munarqanchu.
en	The girl found a bug in her cereal.	Result	She poured milk in the bowl.	She lost her appetite.
th	ตาของฉันแดงและบวม	Cause	ฉันร้องไห้	ฉันหัวเราะ
en	My eyes became red and puffy.	Cause	I was sobbing.	I was laughing.

Tasks: Anaphora resolution (XWinograd) [2]



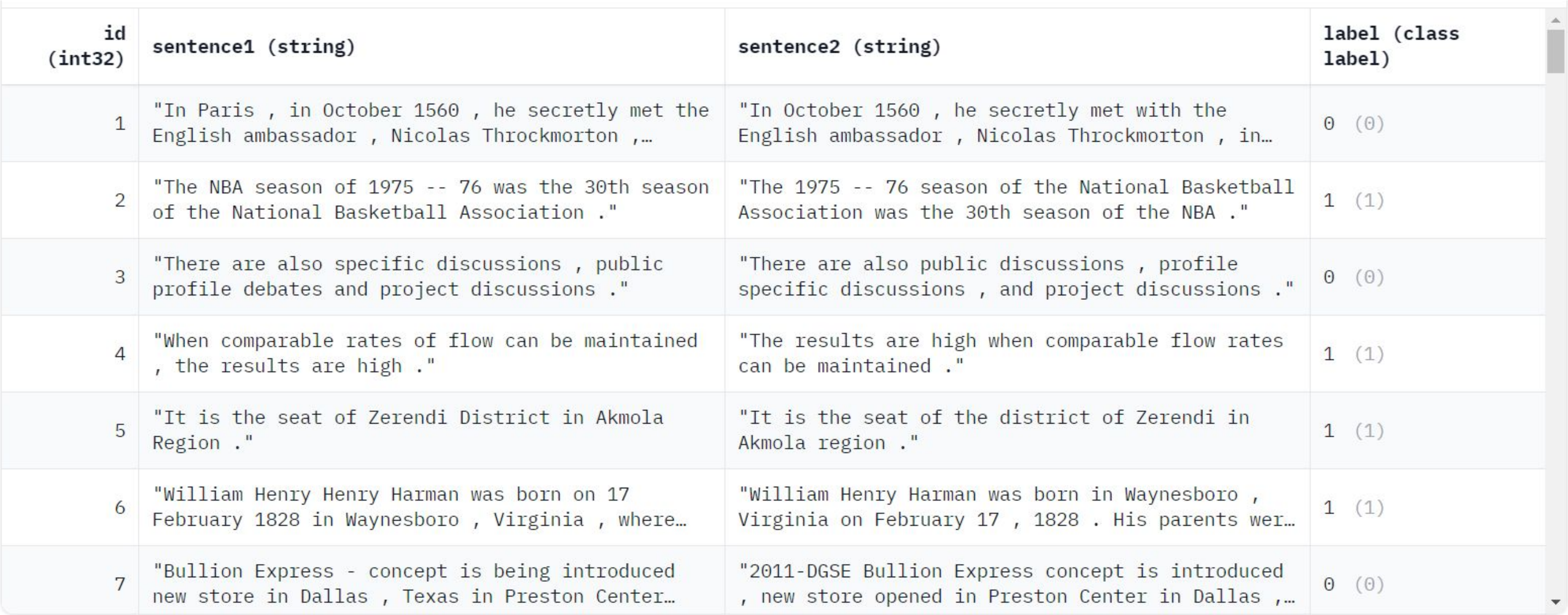
sentence (string)	option1 (string)	option2 (string)	answer (string)
"The city councilmen refused the demonstrators a permit because _ feared violence."	"the demonstrators"	"The city councilmen"	"2"
"The city councilmen refused the demonstrators a permit because _ advocated violence."	"The city councilmen"	"the demonstrators"	"2"
"The trophy doesn't fit into the brown suitcase because _ is too large."	"The trophy"	"suitcase"	"1"
"The trophy doesn't fit into the brown suitcase because _ is too small."	"suitcase"	"The trophy"	"1"
"Joan made sure to thank Susan for all the help _ had recieved."	"Joan"	"Susan"	"1"
"Joan made sure to thank Susan for all the help _ had given."	"Susan"	"Joan"	"1"
"Paul tried to call George on the phone, but _ wasn't successful."	"Paul"	"George"	"1"

Tasks: Natural language inference (XNLI) [3]



Language	Premise / Hypothesis	Genre	Label
English	You don’t have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d’offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction

Tasks: Paraphrasing (PAWS-X) [4]

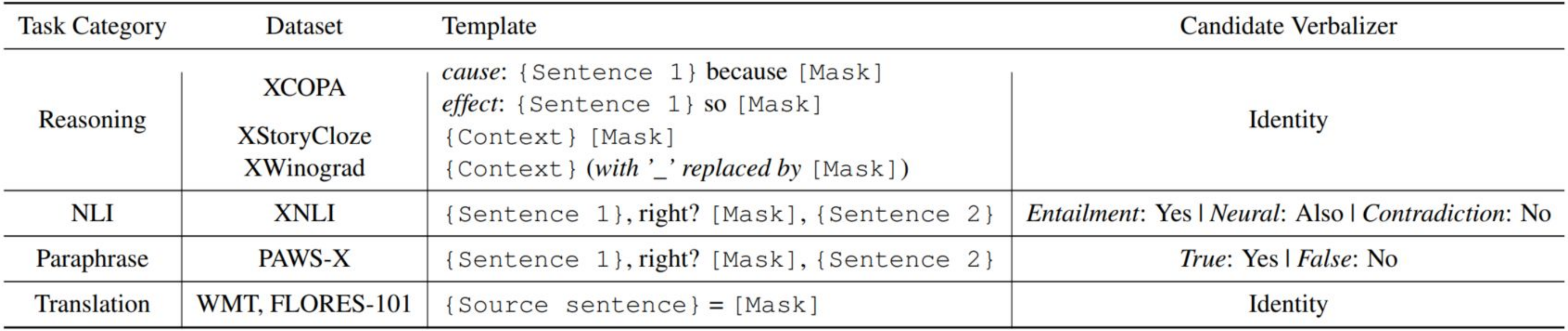


id (int32)	sentence1 (string)	sentence2 (string)	label (class label)
1	"In Paris , in October 1560 , he secretly met the English ambassador , Nicolas Throckmorton ,…	"In October 1560 , he secretly met with the English ambassador , Nicolas Throckmorton , in…	0 (0)
2	"The NBA season of 1975 -- 76 was the 30th season of the National Basketball Association ."	"The 1975 -- 76 season of the National Basketball Association was the 30th season of the NBA ."	1 (1)
3	"There are also specific discussions , public profile debates and project discussions ."	"There are also public discussions , profile specific discussions , and project discussions ."	0 (0)
4	"When comparable rates of flow can be maintained , the results are high ."	"The results are high when comparable flow rates can be maintained ."	1 (1)
5	"It is the seat of Zerendi District in Akmola Region ."	"It is the seat of the district of Zerendi in Akmola region ."	1 (1)
6	"William Henry Henry Harman was born on 17 February 1828 in Waynesboro , Virginia , where…	"William Henry Harman was born in Waynesboro , Virginia on February 17 , 1828 . His parents wer…	1 (1)
7	"Bullion Express - concept is being introduced new store in Dallas , Texas in Preston Center…	"2011-DGSE Bullion Express concept is introduced , new store opened in Preston Center in Dallas ,…	0 (0)

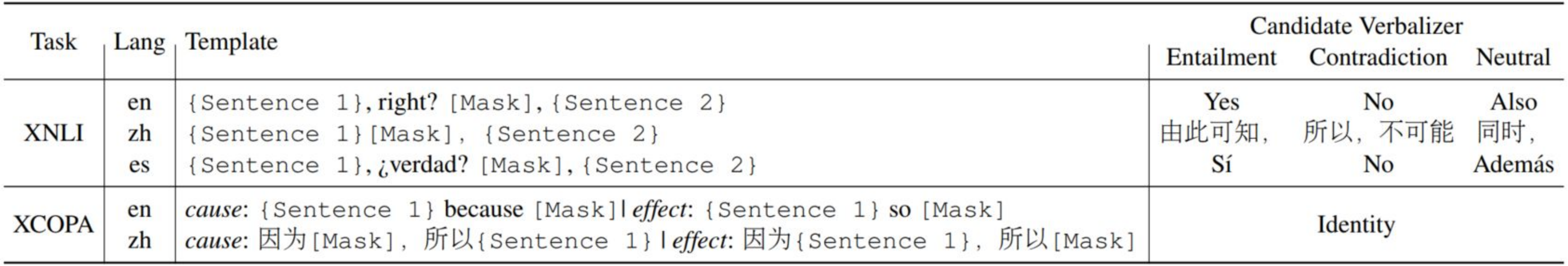
Prompts

- Three approaches for obtaining prompts for non-English tasks
  - Handcrafted prompts
  - Translating from English prompts
  - Cross-lingual prompts
- Also evaluates Cross lingual demonstrations
- This enables cheap transfer from high-resource to low-resource language

Prompts

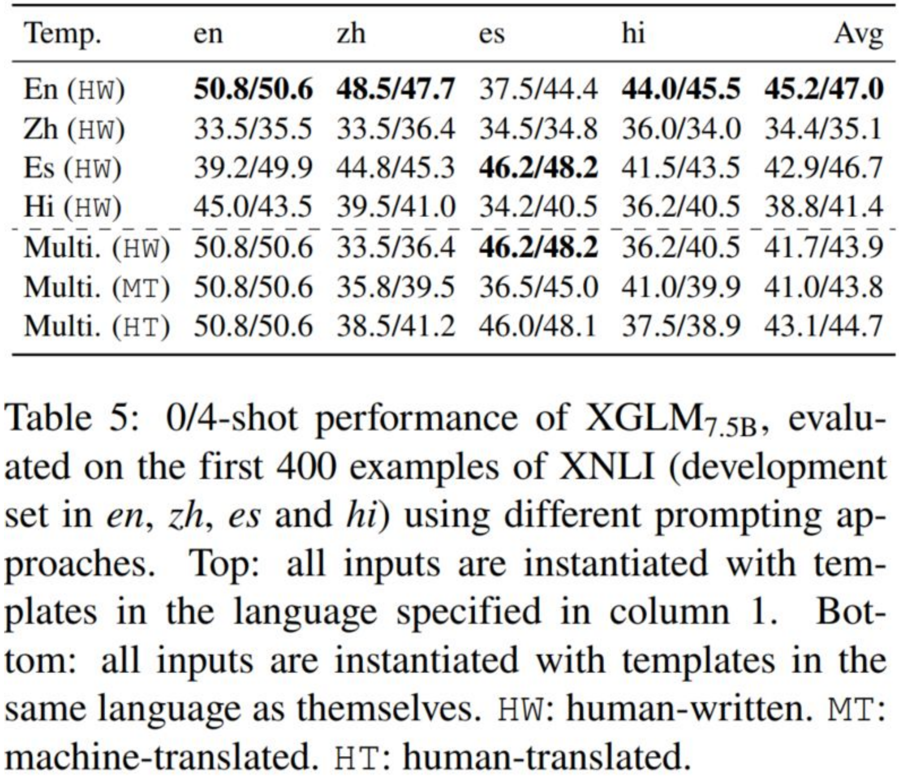


Task Category	Dataset	Template	Candidate Verbalizer
Reasoning	XCOPA	*cause*: {Sentence 1} because [Mask] *effect*: {Sentence 1} so [Mask]	Identity
	XStoryCloze	{Context} [Mask]	
	XWinograd	{Context} (*with ’_’ replaced by* [Mask])	
NLI	XNLI	{Sentence 1}, right? [Mask], {Sentence 2}	*Entailment*: Yes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



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

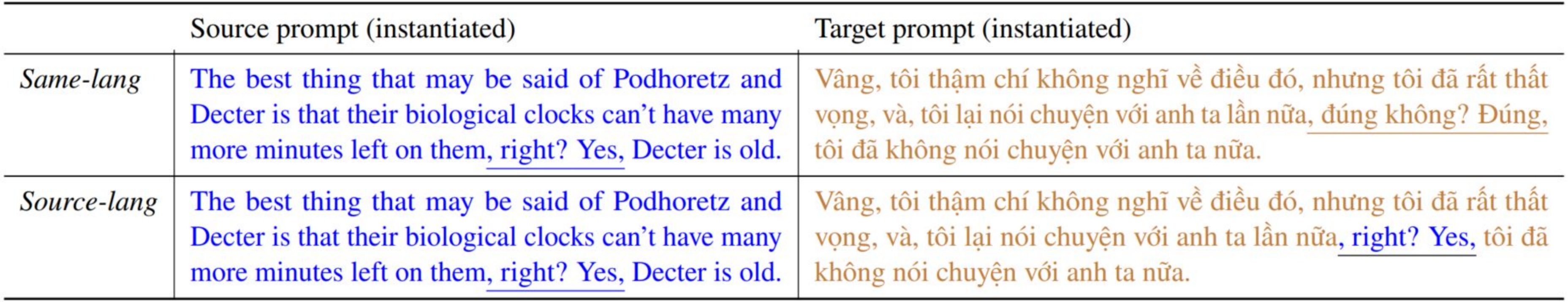
Results: Prompt strategy



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

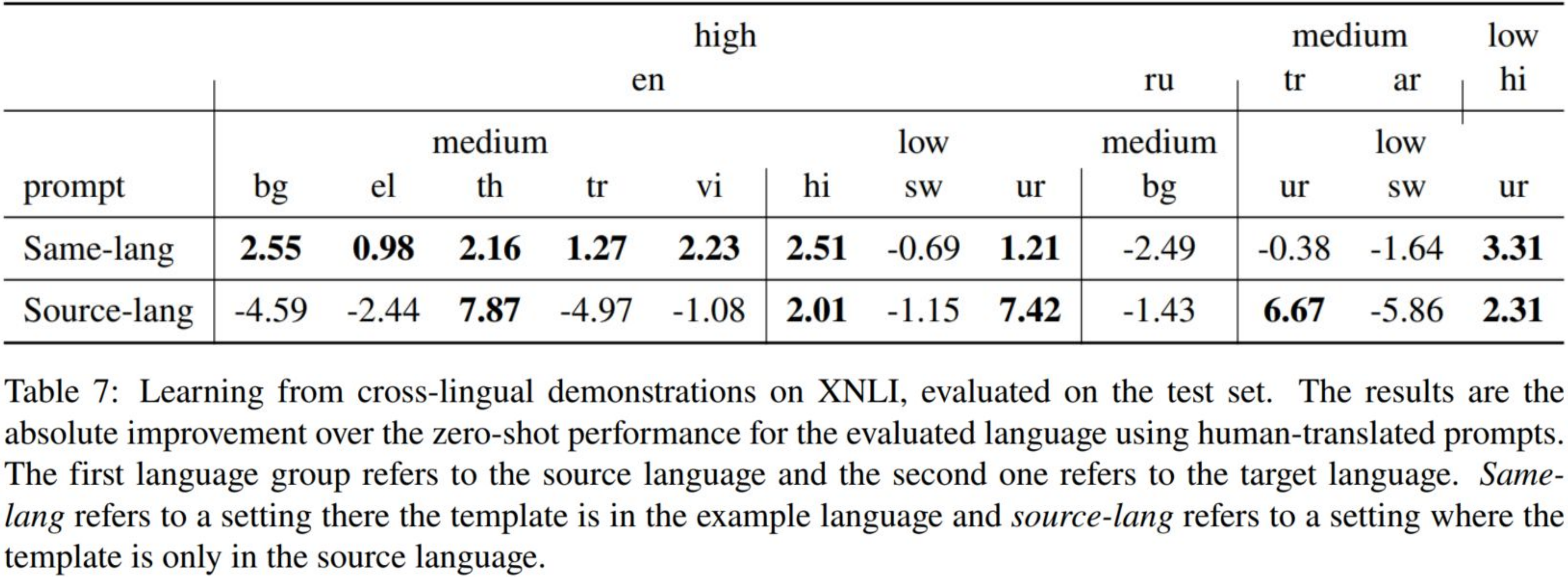
Table 5: 0/4-shot performance of XGLM$_{7.5B}$, evaluated on the first 400 examples of XNLI (development set in *en*, *zh*, *es* and *hi*) using different prompting approaches. Top: all inputs are instantiated with templates in the language specified in column 1. Bottom: all inputs are instantiated with templates in the same language as themselves. HW: human-written. MT: machine-translated. HT: human-translated.

Prompt Language



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

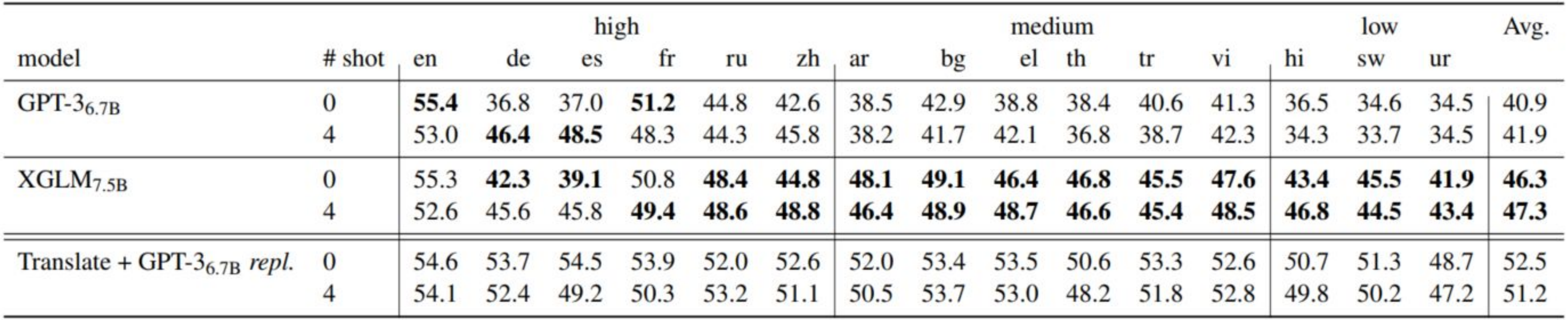
Results: Prompt Language



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

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

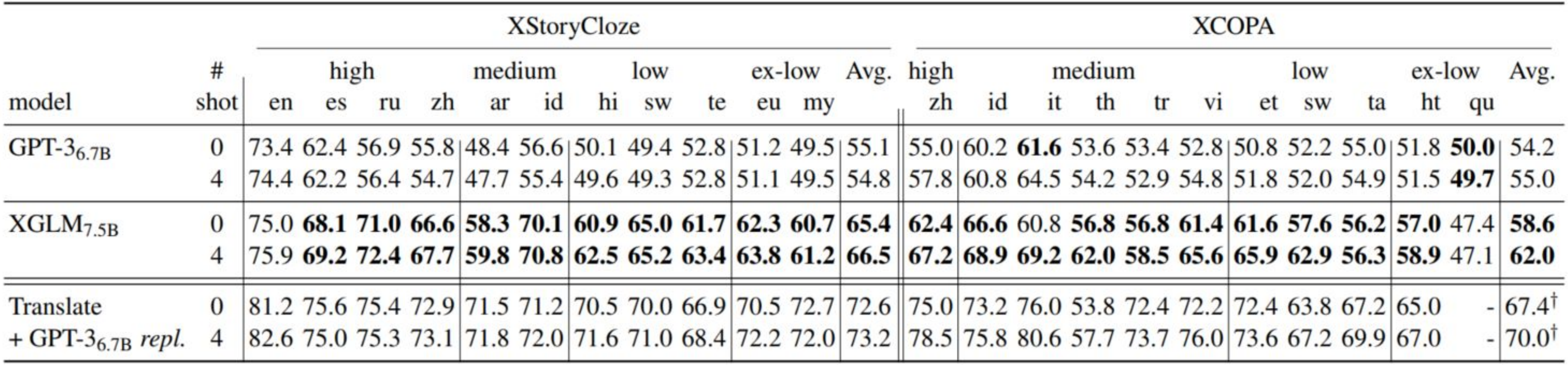
Results: Comparison



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

Table 9: Comparison of different models on XNLI.

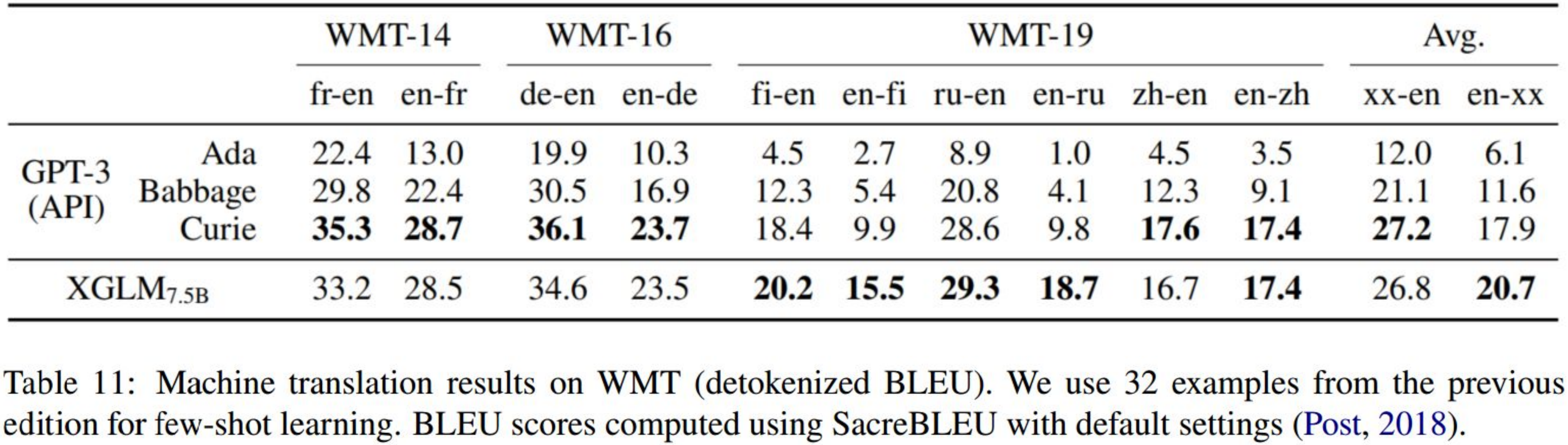
Results: Comparison



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

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

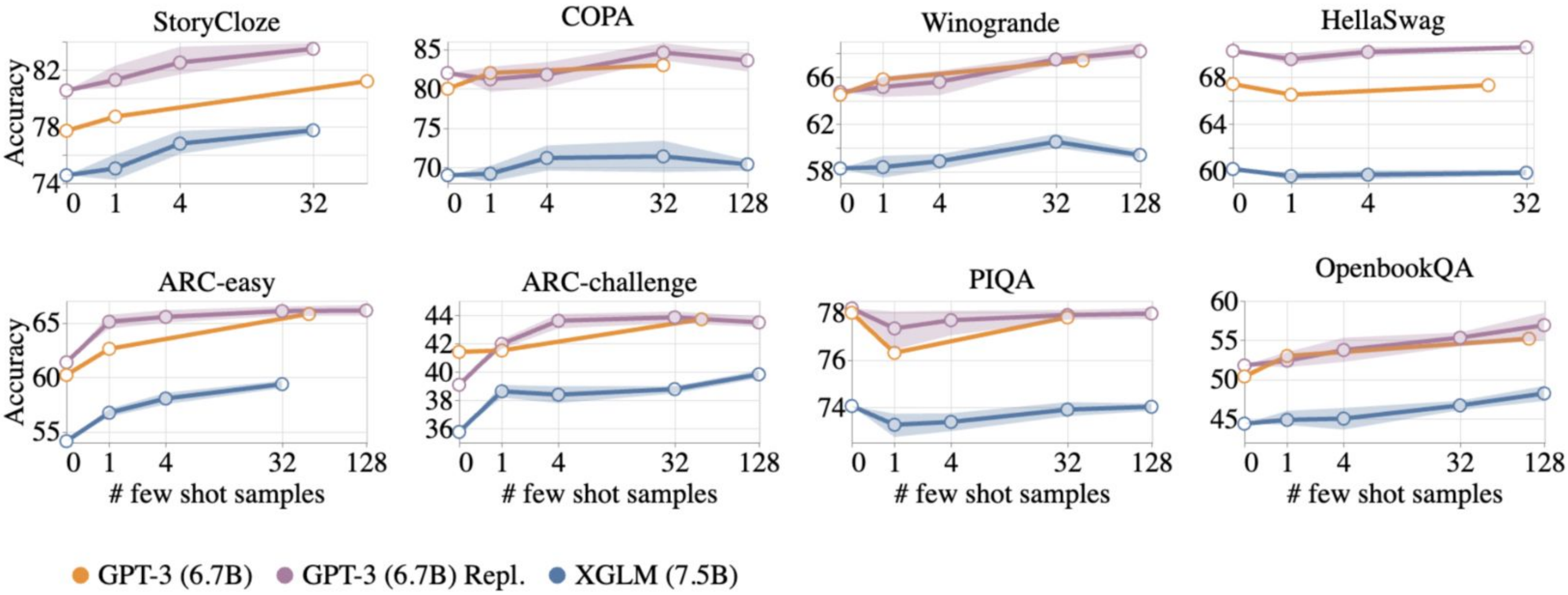
Results: Machine translation



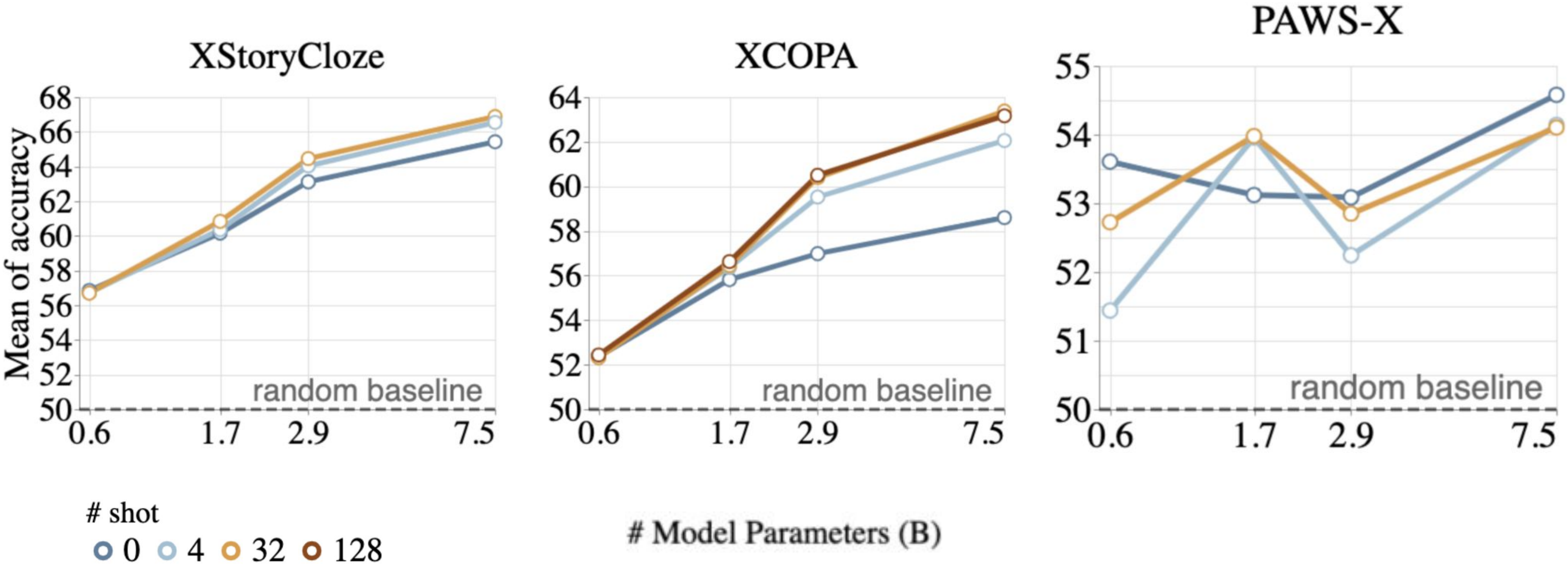
		WMT-14		WMT-16		WMT-19						Avg.	
		fr-en	en-fr	de-en	en-de	fi-en	en-fi	ru-en	en-ru	zh-en	en-zh	xx-en	en-xx
GPT-3 (API)	Ada	22.4	13.0	19.9	10.3	4.5	2.7	8.9	1.0	4.5	3.5	12.0	6.1
	Babbage	29.8	22.4	30.5	16.9	12.3	5.4	20.8	4.1	12.3	9.1	21.1	11.6
	Curie	**35.3**	**28.7**	**36.1**	**23.7**	18.4	9.9	28.6	9.8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

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

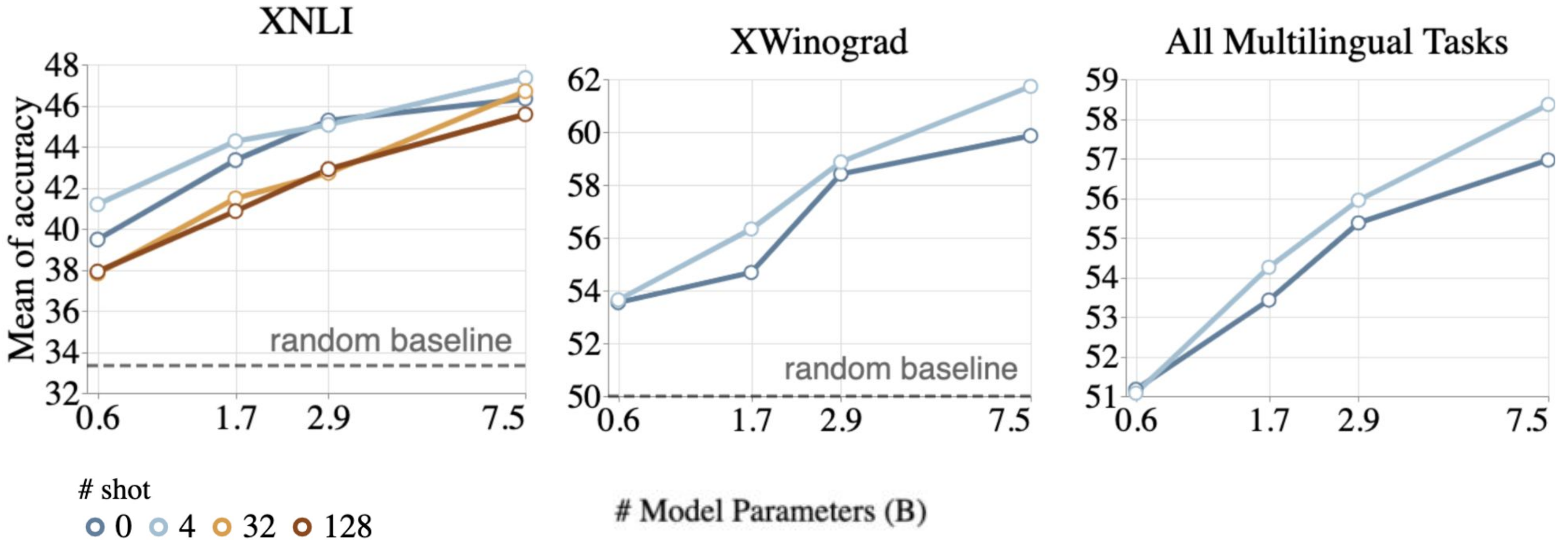
Results: English Tasks



Results: XGLM scale



Results: XGLM scale



Strengths

- Provides a balanced dataset for multilingual NLP research
- Demonstrates different prompting strategies for multilingual tasks
- Evaluates zero and few shot performance of XGLM comprehensively on diverse languages

Weaknesses

- Does not provide any new model or does not discuss how the prior models are appropriate for multilingual generation tasks
- Uses only 30 languages, whereas, despite uneven ratio, GPT-3 is trained on 118 languages
- No comparison or reference to the zero or few-shot performance of other multilingual models like mBERT, XLM-R, mT5, mBART
- Degrades performance on English tasks

Thanks

Questions?

References

[1] Ponti, Edoardo Maria, et al. "XCOPA: A multilingual dataset for causal commonsense reasoning." arXiv preprint arXiv:2005.00333 (2020).

[2] Emelin, Denis, and Rico Sennrich. "Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

[3] Conneau, Alexis, et al. "XNLI: Evaluating cross-lingual sentence representations." arXiv preprint arXiv:1809.05053 (2018).

[4] Yang, Yinfei, et al. "PAWS-X: A cross-lingual adversarial dataset for paraphrase identification." arXiv preprint arXiv:1908.11828 (2019).