	Tokenizer Fertility				Pre-training Size			
	GPT-3.5-Turbo		GPT-4		GPT-3.5-Turbo		GPT-4	
Task	ρ	P-value	ρ	P-value	ρ	P-value	ρ	P-value
XNLI+IndicXNLI	-0.784	$6.9e-05$	-0.803	$3.4e-05$	0.893	$4.1e-09$	0.836	$3.5e-07$
XCOPA	-0.982	$7.9e-05$	-0.957	0.00	0.70	0.035	0.489	0.181
XstoryCloze	-0.745	0.033	-0.918	0.001	0.603	0.064	0.407	0.242
PAWS-X	-0.587	0.219	-0.61	0.198	0.85	0.031	0.94	0.005
MLQA	-0.451	0.368	-0.674	0.141	0.71	0.085	0.808	0.051
TyDiQA-GoldP	0.543	0.163	0.049	0.907	-0.464	0.207	-0.159	0.682
XQuAD	-0.865	0.00	-0.818	0.002	0.782	0.004	0.736	0.009
IndicQA	-0.960	0.002	-0.856	0.029	0.628	0.051	0.690	0.027
WinoMT	-0.36	0.379	-	-	0.249	0.589	-	-
Jigsaw	0.306	0.554	-	-	-0.674	0.141	-	-
PAN-X	-0.456	0.003	-0.442	0.004	0.41	0.006	0.326	0.032
UDPOS	-0.216	0.198	-0.304	0.066	0.29	0.095	0.359	0.036
XLSum	-0.821	$4.8e-07$	-0.578	0.002	0.448	0.011	0.49	0.005

Table 5: Pearson Correlation coefficient ρ between performance and tokenizer fertility and performance and pre-training data size for different datasets and models. We also provide the p-values, to see which correlations are statistically significant.

NaturalInstructions (SNI)(Wang et al., 2022)[7]. All the explanations that we used were written in English. For XStoryCloze, the results are plotted in Figure 6d, and we observe that while there is a slight gain upon using explanations for Telugu, for all other languages the performance remains largely unchanged if not slightly worse. Interestingly, upon manual inspection of the model's prediction, we observe that the model often first translates the problem to English and then proceeds with the explanation, without having prompted to do so. We have similar observations for the XCOPA dataset as well, where adding explanations doesn't help improve performance and ends up hurting the performance by a slight margin (Figure 9)

Effect of the language of the prompt templates. While all our experiments were run using prompt templates written in English, we initially evaluated DV003 on Native-Language-Templates as well, which were obtained by translating English templates using Bing-Translator. As can be seen in Table 6, the performance is much worse when using templates in the native language compared to English. This is consistent with the results in Muennighoff et al. (2022) for BLOOMZ and Lin et al.

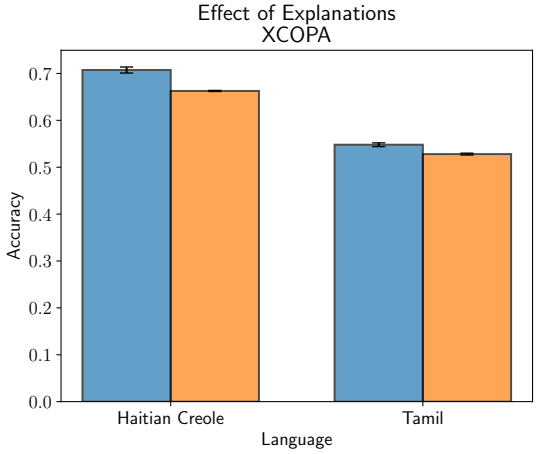


Figure 9: Effect of using explanations in XCOPA dataset. Blue bars mean no explanations in the prompt and orange bars correspond to prompting with explanations.

[7]At the time of writing this paper, XStoryCloze wasn't included in SNI, hence we use the few-shot examples and explanations available for StoryCloze dataset(Mostafazadeh et al., 2016), making the prompting setup Zero-Shot Cross-Lingual.

Task	English-Template	Native-Language-Template
XNLI	**58.3**	54.4
Indic-XNLI	**49.6**	38.7
PAWS-X	**67.1**	64.2
XCOPA	**77.6**	73.1

Table 6: Average performance on non-English languages with the Monolingual Prompting strategy using English-Template and Native-Language-Template prompts for the classification tasks for DV003.

Model	IndicXNLI	IndicQA
MuRIL	**72.4**	47.7
`text-davinci-003`	49.6	8.45
`text-davinci-003` (TT)	62.4	×
`gpt-3.5-turbo`	50.7	38.6
`gpt-3.5-turbo` (TT)	59.7	×
`gpt-4-32k`	66.8	**55.0**

Table 7: Comparing performance of `text-davinci-003`, `gpt-3.5-turbo`, and `gpt-4-32k` with fine-tuned baseline MuRIL on Indic datasets (Doddapaneni et al., 2022). For IndicXNLI we report Accuracy and F1-score for IndicQA.

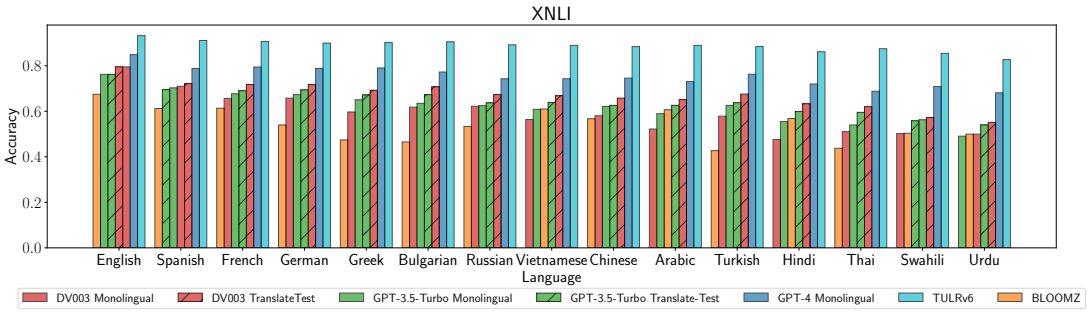
(2022a) for XGLM, which also show better performance when using prompt templates in English.

A.8 Detailed Results

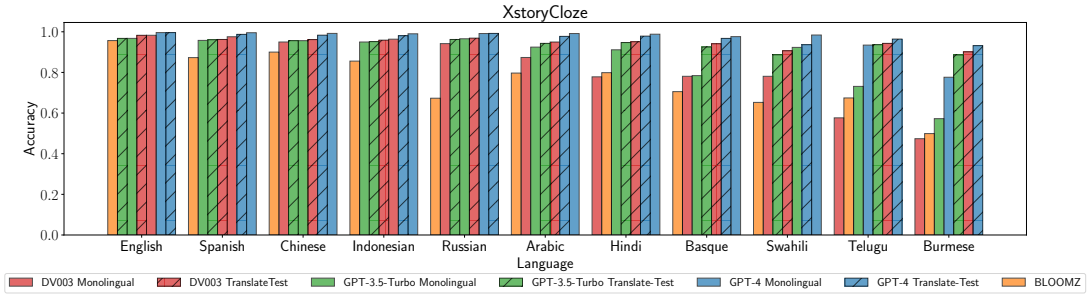
The results for across all tasks, languages and models included in our benchmarking exercise can are provided in Figures 10 (for Classification tasks), 11 (for QA Tasks), 12 (for XLSum), 14 (for PAN-X), 13 (for UDPOS), 15 (for Jigsaw), and finally 16 (for Wino-MT). The results for the Indic Datasets and the two code-mixed datasets GLUECoS NLI and En-ES-CS are provided in Tables 7 8 respectively.

Model	NLI En-Hi	Sentiment En-Es
mBERT	63.1	69.31
`text-davinci-003`	72.1	**68.8**
`gpt-3.5-turbo`	**78.8**	68.0

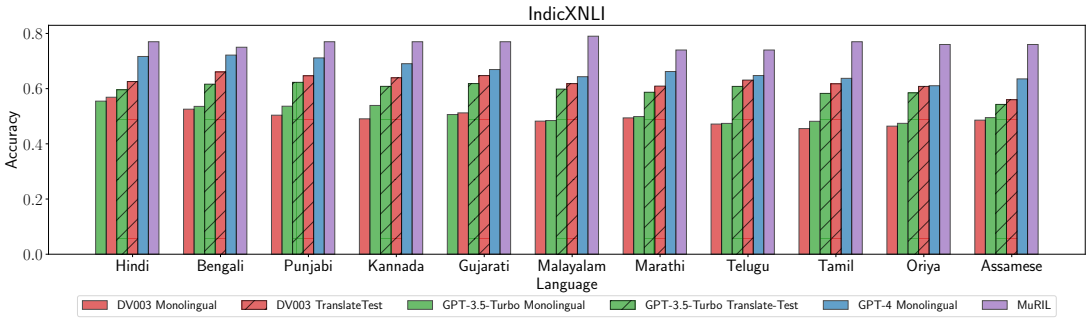
Table 8: Performance of GPT-3.5 models on code-mixing datasets from (Khanuja et al., 2020b). For both the tasks, the metric reported is accuracy.



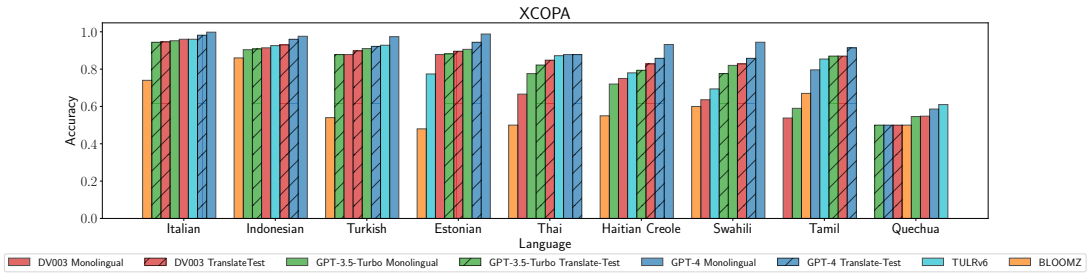
(a)



(b)



(c)



(d)

Figure 10: Comparing the language-wise performance of different models on the classification tasks.