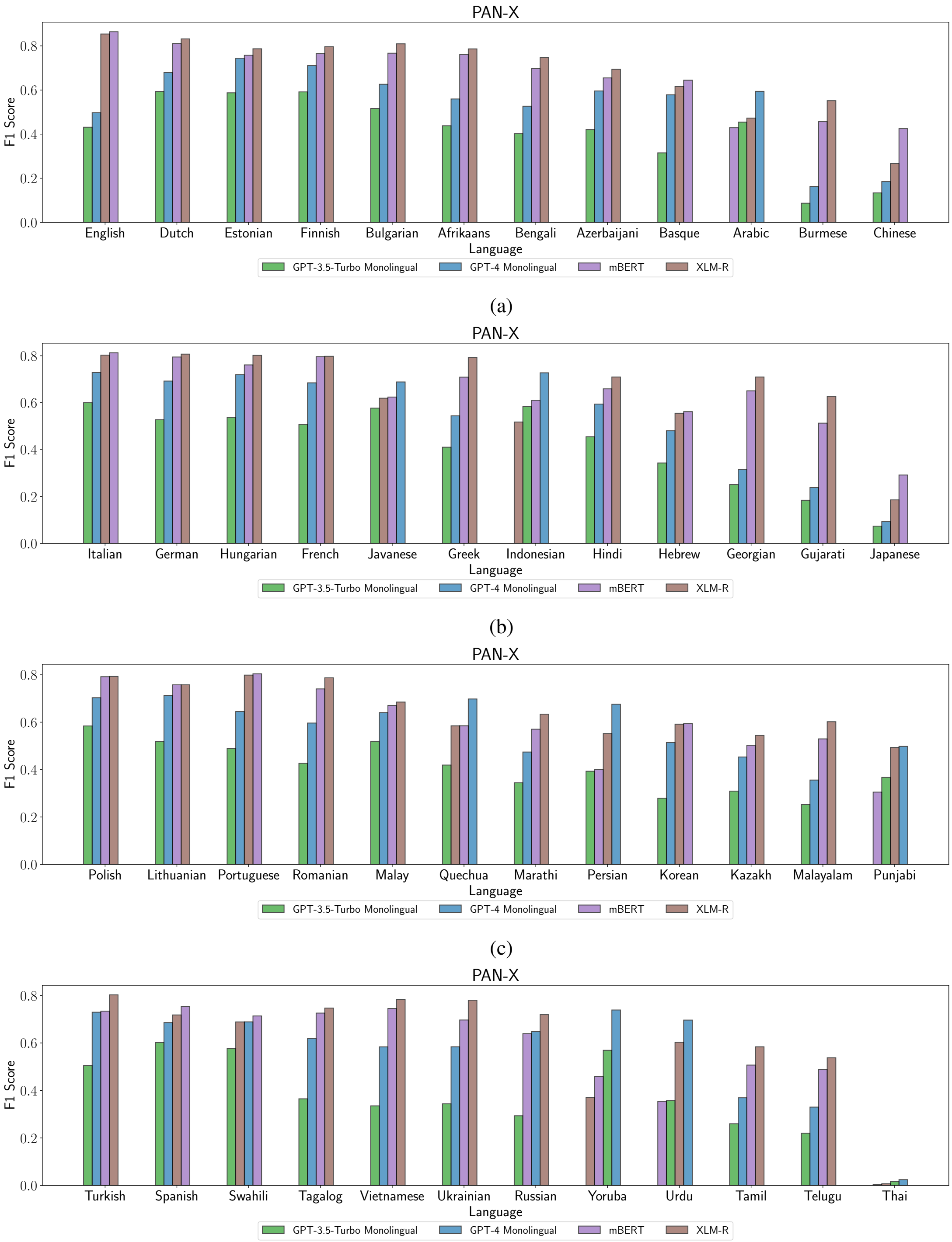

(a)

(b)

(c)

(d)

Figure 14: Comparing performance of different models on PAN-X

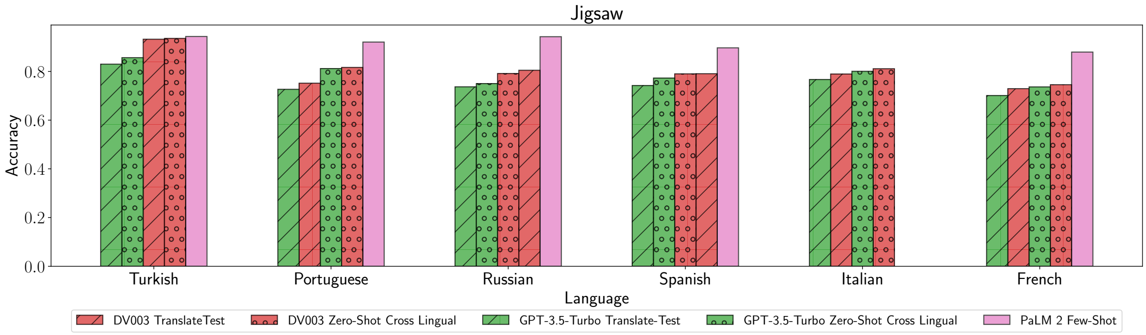

Figure 15: Comparing performance of different models on the Jigsaw dataset.

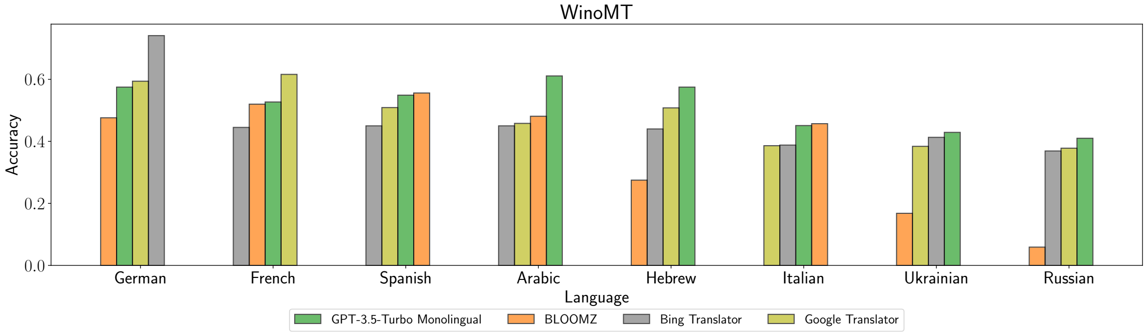

Figure 16: Comparing performance of different models on the WinoMT dataset.

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	**avg**
Fine-tuned Baselines																
mBERT	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9	67.8	49.7	54.1	60.9	57.2	69.3	67.8	65.4
mT5-Base	84.7	73.3	78.6	77.4	77.1	80.3	79.1	70.8	77.1	69.4	73.2	72.8	68.3	74.2	74.1	75.4
XLM-R Large	88.7	77.2	83.0	82.5	80.8	83.7	82.2	75.6	79.1	71.2	77.4	78.0	71.7	79.3	78.2	79.2
TuLRv6 - XXL	**93.3**	**89.0**	**90.6**	**90.0**	**90.2**	**91.1**	**90.7**	**86.2**	**89.2**	**85.5**	**87.5**	**88.4**	**82.7**	**89.0**	**88.4**	**88.8**
Prompt-Based Baselines																
BLOOMZ	67.5	60.7	46.5	54.0	47.4	61.2	61.4	56.8	53.3	50.4	43.8	42.7	50.0	61.0	56.7	54.2
XGLM	52.6	46.4	48.9	45.6	48.7	45.8	49.4	46.8	48.6	44.5	46.6	45.4	43.4	48.5	48.8	47.3
Open AI Models																
`gpt-3.5-turbo`	76.2	59.0	63.5	67.3	65.1	70.3	67.7	55.5	62.5	56.3	54.0	62.6	49.1	60.9	62.1	62.1
`gpt-3.5-turbo` (TT)	76.2	62.7	67.3	69.4	67.2	69.6	69.0	59.9	63.7	55.8	59.6	63.8	54.0	63.9	62.6	64.3
`text-davinci-003`	79.5	52.2	61.8	65.8	59.7	71.0	65.7	47.6	62.2	50.2	51.1	57.9	50.0	56.4	58.0	59.3
`text-davinci-003` (TT)	79.5	65.1	70.8	71.7	69.3	72.2	71.8	63.3	67.3	57.3	62.0	67.6	55.1	66.9	65.8	67.1
`gpt-4-32k`	84.9	73.1	77.3	78.8	79.0	78.8	79.5	72.0	74.3	70.9	68.8	76.3	68.1	74.3	74.6	75.4

Table 9: Comparing performance of different models on all languages in XNLI. Metric: Accuracy.

Model	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	**avg**
Fine-tuned Baselines												
MuRIL	**76.0**	**75.0**	**77.0**	**77.0**	**77.0**	**79.0**	**74.0**	**76.0**	**77.0**	**77.0**	**74.0**	**76.0**
Open AI Models												
`gpt-3.5-turbo`	49.5	53.6	50.6	55.5	53.9	48.4	49.9	47.4	53.6	48.2	47.4	50.7
`gpt-3.5-turbo` (TT)	54.3	61.6	61.8	59.6	60.8	59.9	58.7	58.5	62.3	58.3	60.8	59.7
`text-davinci-003`	48.6	52.6	51.2	56.9	49.1	48.2	49.4	46.4	50.4	45.5	47.2	49.6
`text-davinci-003` (TT)	56.0	66.0	64.7	62.6	63.9	61.8	60.9	60.8	64.7	61.8	63.1	62.4
`gpt-4-32k`	63.5	72.2	66.9	71.7	69.0	64.3	66.2	61.1	71.1	63.7	64.8	66.8

Table 10: Comparing performance of different models on all languages in IndicXNLI. Metric: Accuracy.

Model	en	de	es	fr	ja	ko	zh	**avg**
Fine-tuned Baselines								
mBERT	94.0	85.7	87.4	87.0	73.0	69.6	77.0	81.9
mT5-Base	95.4	89.4	89.6	91.2	79.8	78.5	81.1	86.4
XLM-R Large	94.7	89.7	90.1	90.4	78.7	79.0	82.3	86.4
TuLRv6 - XXL	**97.2**	**95.1**	**94.8**	**95.6**	**89.4**	**90.4**	**90.4**	**93.2**
Prompt-Based Baselines								
BLOOMZ	89.8	84.3	88.9	87.5	74.4	85.8	65.2	82.3
Open AI Models								
`gpt-3.5-turbo`	72.4	70.6	72.0	72.1	67.2	66.5	69.2	70.0
`gpt-3.5-turbo` (TT)	72.4	70.8	69.7	70.1	61.9	62.5	63.1	67.2
`text-davinci-003`	72.5	70.6	72.7	70.7	60.6	61.8	60.8	67.1
`text-davinci-003` (TT)	72.5	69.8	70.1	71.3	65.4	65.8	65.2	68.6
`gpt-4-32k`	76.2	74.0	74.1	72.6	71.5	69.9	72.6	73.0

Table 11: Comparing performance of different models on all languages in PAWS-X. Metric: Accuracy.

Model	en	et	ht	id	it	qu	sw	ta	th	tr	**avg**
Fine-tuned Baselines											
mT5-Base	-	50.3	49.9	49.2	49.6	50.5	50.4	49.2	50.7	49.5	49.9
TuLRv6 - XXL	-	77.4	78.0	92.6	96.0	61.0	69.4	85.4	87.2	92.8	74.0
Prompt-Based Baselines											
BLOOMZ	88.0	48.0	55.0	86.0	74.0	50.0	60.0	67.0	50.0	54.0	63.2
XGLM	-	65.9	58.9	68.9	69.2	47.1	62.9	56.3	62.0	58.5	61.1
Open AI Models											
`gpt-3.5-turbo`	97.8	90.6	72.0	90.4	95.2	54.6	82.0	59.0	77.6	91.0	81.0
`gpt-3.5-turbo` (TT)	97.8	88.2	79.4	90.8	94.4	50.0	77.6	87.0	82.2	87.8	83.5
`text-davinci-003`	98.2	87.8	75.0	91.4	96.0	54.8	63.6	53.8	66.6	87.8	77.5
`text-davinci-003` (TT)	98.2	89.6	82.8	93.0	94.6	50.0	82.8	87.0	84.8	89.8	85.3
`gpt-4-32k`	**99.6**	**98.8**	**93.2**	**97.6**	**99.8**	58.6	**94.4**	79.6	**87.8**	**97.4**	**90.7**
`gpt-4-32k` (TT)	**99.6**	94.4	85.8	96.0	98.2	**85.8**	83.4	**91.4**	**87.8**	92.2	90.6

Table 12: Comparing performance of different models on all languages in XCOPA. Metric: Accuracy.