Model	en	ar	de	el	es	hi	ru	th	tr	vi	zh	**avg**
Fine-tuned Baselines												
mBERT	83.5 / 72.2	61.5 / 45.1	70.6 / 54.0	62.6 / 44.9	75.5 / 56.9	59.2 / 46.0	71.3 / 53.3	42.7 / 33.5	55.4 / 40.1	69.5 / 49.6	58.0 / 48.3	64.5 / 49.4
mT5-Base	84.6 / 71.7	63.8 / 44.3	73.8 / 54.5	59.6 / 35.6	74.8 / 56.1	60.3 / 43.4	57.8 / 34.7	57.6 / 45.7	67.9 / 48.2	70.7 / 50.3	66.1 / 54.1	67.0 / 49.0
XLM-R Large	86.5 / 75.7	68.6 / 49.0	80.4 / 63.4	79.8 / 61.7	82.0 / 63.9	76.7 / 59.7	80.1 / 64.3	74.2 / 62.8	75.9 / 59.3	79.1 / 59.0	59.3 / 50.0	76.6 / 60.8
TuLRv6 - XXL	90.1 / 80.6	**85.4 / 69.6**	**86.1 / 70.4**	**86.3 / 70.4**	**87.6 / 71.0**	**85.9 / 70.5**	**86.8 / 73.2**	**87.0 / 81.1**	**84.3 / 71.0**	**87.6 / 71.3**	79.2 / 73.2	**86.0 / 72.9**
Prompt-Based Baselines												
BLOOMZ	**92.1 / 83.8**	82.8 / 69.7	76.3 / 60.4	49.7 / 37.6	86.8 / 71.4	83.4 / 72.9	65.7 / 47.2	20.5 / 15.5	51.4 / 37.2	86.9 / 72.7	**82.4 / 78.6**	70.7 / 58.8
Open AI Models												
`gpt-3.5-turbo`	79.3 / 58.7	59.6 / 35.1	70.6 / 46.6	49.0 / 22.8	70.3 / 40.8	54.0 / 29.0	58.0 / 31.3	41.9 / 30.4	61.8 / 35.0	69.1 / 42.4	50.4 / 48.3	60.4 / 38.2
`text-davinci-003`	77.2 / 61.8	36.8 / 22.5	55.2 / 39.7	31.8 / 19.7	61.8 / 41.3	19.9 / 10.0	29.4 / 17.6	11.5 / 8.7	44.8 / 29.2	41.7 / 25.4	35.6 / 32.8	40.5 / 28.1
`gpt-4-32k`	83.2 / 65.6	67.8 / 42.4	71.9 / 48.7	62.3 / 36.6	77.5 / 50.7	63.9 / 36.7	63.8 / 35.8	54.6 / 42.0	70.8 / 46.6	75.8 / 49.7	60.0 / 57.5	68.3 / 46.6

Table 13: Comparing performance of different models on all languages in XQuAD. Metric: F1 Score / Exact Match.

Model	en	ar	bn	fi	id	ko	ru	sw	te	**avg**
Fine-tuned Baselines										
mBERT	75.3 / 63.6	62.2 / 42.8	49.3 / 32.7	59.7 / 45.3	64.8 / 45.8	58.8 / 50.0	60.0 / 38.8	57.5 / 37.9	49.6 / 38.4	59.7 / 43.9
mT5-Base	71.8 / 60.9	67.1 / 50.4	40.7 / 22.1	67.0 / 52.2	71.3 / 54.5	49.5 / 37.7	54.9 / 32.6	60.4 / 43.9	40.6 / 31.1	58.1 / 42.8
XLM-R Large	71.5 / 56.8	67.6 / 40.4	64.0 / 47.8	70.5 / 53.2	77.4 / 61.9	31.9 / 10.9	67.0 / 42.1	66.1 / 48.1	70.1 / 43.6	65.1 / 45.0
TuLRv6 - XXL	**85.4 / 76.4**	**84.1 / 70.4**	86.9 / 79.6	**83.8 / 72.8**	**88.8 / 77.9**	**78.5 / 67.8**	**81.9 / 68.6**	**87.2 / 79.6**	85.2 / 71.6	**84.6 / 73.8**
Prompt-Based Baselines										
BLOOMZ	82.4 / 70.9	81.9 / 62.2	**87.8 / 82.3**	43.6 / 28.6	85.0 / 71.0	52.3 / 43.1	67.4 / 51.5	86.0 / 77.2	**90.3 / 81.6**	75.2 / 63.2
Open AI Models										
`gpt-3.5-turbo`	54.8 / 30.7	50.9 / 24.2	60.7 / 32.7	66.6 / 49.0	67.2 / 43.4	59.7 / 45.3	45.8 / 20.0	64.3 / 47.7	70.9 / 53.1	60.1 / 38.4
`text-davinci-003`	73.7 / 59.1	56.2 / 38.7	16.1 / 10.6	70.3 / 58.8	68.6 / 51.2	40.6 / 32.2	42.3 / 28.9	74.1 / 62.3	5.8 / 3.0	49.8 / 38.3
`gpt-4-32k`	72.9 / 51.4	60.8 / 32.7	68.0 / 42.5	75.4 / 57.7	80.8 / 61.1	69.7 / 58.5	61.4 / 30.5	81.8 / 68.7	72.5 / 54.9	71.5 / 50.9

Table 14: Comparing performance of different models on all languages in TyDiQA. Metric: F1 Score / Exact Match.

Model	en	ar	de	es	hi	vi	zh	**avg**
Fine-tuned Baselines								
mBERT	80.2 / 67.0	52.3 / 34.6	59.0 / 43.8	67.4 / 49.2	50.2 / 35.3	61.2 / 40.7	59.6 / 38.6	61.4 / 44.2
mT5-Base	81.7 / 66.9	57.1 / 36.9	62.1 / 43.2	67.1 / 47.2	55.4 / 37.9	65.9 / 44.1	61.6 / 38.6	64.4 / 45.0
XLM-R Large	83.5 / 70.6	66.6 / 47.1	70.1 / 54.9	74.1 / 56.6	70.6 / 53.1	74.0 / 52.9	62.1 / 37.0	71.6 / 53.2
TuLRv6 - XXL	**86.6 / 74.4**	**76.2 / 56.5**	**80.2 / 67.0**	**81.7 / 65.1**	**82.2 / 64.8**	**82.3 / 63.2**	**78.1 / 56.5**	**81.0 / 63.9**
Open AI Models								
`gpt-3.5-turbo`	72.8 / 53.2	48.5 / 23.9	51.0 / 29.6	53.8 / 29.4	50.7 / 28.9	58.9 / 35.1	56.7 / 29.4	56.1 / 32.8
`gpt-3.5-turbo` (TT)	72.8 / 53.2	37.8 / 18.4	44.3 / 26.2	54.1 / 31.8	37.3 / 20.0	41.6 / 22.5	36.5 / 17.2	46.4 / 27.0
`text-davinci-003`	74.8 / 59.0	38.4 / 21.7	57.7 / 38.1	62.9 / 37.8	24.9 / 14.1	47.7 / 29.7	32.3 / 31.7	48.4 / 33.1
`text-davinci-003` (TT)	74.8 / 59.0	48.2 / 25.6	53.5 / 33.9	62.9 / 40.9	49.2 / 28.7	51.0 / 30.4	45.2 / 24.1	55.0 / 34.7
`gpt-4-32k`	80.3 / 62.8	59.1 / 33.5	64.7 / 44.4	70.0 / 45.9	57.3 / 35.6	72.2 / 49.0	67.1 / 38.4	67.2 / 44.2

Table 15: Comparing performance of different models on all languages in MLQA. Metric: F1 Score / Exact Match.

Model	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	**avg**
Fine-tuned Baselines												
BLOOMZ	40.6 / 31.7	42.9 / 36.6	37.2 / 29.9	44.0 / 45.1	37.8 / 26.6	30.5 / 28.4	39.2 / 33.0	25.4 / 22.0	26.4 / 33.5	39.7 / 35.9	38.9 / 34.7	36.6 / 32.5
Open AI Models												
`gpt-3.5-turbo`	35.3 / 21.4	49.5 / 30.2	40.5 / 25.5	55.9 / 39.3	35.3 / 20.4	30.0 / 19.2	50.0 / 32.0	22.1 / 12.7	35.8 / 15.1	32.7 / 21.6	32.9 / 19.7	38.2 / 23.4
`text-davinci-003`	6.7 / 3.2	10.3 / 5.8	5.4 / 3.5	16.8 / 11.8	7.1 / 3.9	3.6 / 2.3	14.6 / 8.5	6.9 / 3.4	10.7 / 4.1	4.2 / 2.5	6.8 / 3.6	8.4 / 4.8
`gpt-4-32k`	**58.8 / 40.4**	**67.1 / 47.4**	**59.4 / 42.4**	**75.2 / 62.2**	**47.1 / 31.6**	**48.3 / 33.7**	**60.7 / 43.1**	**29.9 / 16.7**	**56.1 / 34.1**	**54.0 / 39.7**	**47.9 / 27.8**	**55.0 / 38.1**

Table 16: Comparing performance of different models on all languages in IndicQA. Metric: F1 Score / Exact Match.

Model	en	af	ar	bg	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	kk	
Fine-tuned Baselines																				
mBERT	96.4	86.7	50.0	84.7	88.7	80.9	86.6	79.9	62.1	65.5	73.3	81.2	55.5	66.0	78.6	74.2	87.8	47.2	70.4	
XLM-R Large	**97.0**	**89.2**	**63.0**	**88.3**	**91.2**	**86.5**	**89.2**	**87.3**	74.9	**70.8**	**82.7**	**86.7**	**67.5**	**75.2**	**83.4**	**75.7**	**89.2**	29.3	**78.3**	
Open AI Models																				
`gpt-3.5-turbo`	78.5	74.3	38.3	79.1	80.7	47.1	34.8	76.0	72.0	46.7	79.5	78.0	53.8	50.7	65.4	63.6	75.4	47.4	64.8	
`gpt-4-32k`	84.1	77.6	42.0	83.1	86.3	49.8	68.4	80.2	**79.3**	46.4	**82.7**	85.4	60.4	52.2	68.3	68.6	84.1	**60.2**	71.8	
	ko	lt	mr	nl	pl	pt	ro	ru	ta	te	th	tl	tr	uk	ur	vi	wo	yo	zh	**avg**
Fine-tuned Baselines																				
mBERT	51.7	78.8	68.7	88.6	80.7	88.0	71.5	82.4	58.5	75.2	41.3	80.5	70.5	80.6	56.6	55.4	0.0	**56.6**	**59.6**	71.9
XLM-R Large	**57.1**	**84.2**	**81.8**	**89.5**	**86.8**	**90.2**	**82.6**	**87.3**	**64.0**	84.2	**48.5**	**92.4**	**81.2**	**85.8**	**70.8**	58.5	0.0	24.8	44.1	**76.2**
Open AI Models																				
`gpt-3.5-turbo`	39.0	71.3	57.9	78.3	81.7	76.7	66.7	69.9	32.6	79.8	25.5	54.3	77.2	58.9	39.9	57.7	50.4	7.0	57.2	60.2
`gpt-4-32k`	51.2	73.7	79.1	81.8†	80.7	81.0	66.3†	74.7	34.7	**84.6**	31.2†	58.4†	77.0	61.9	41.3	**64.7**	**59.1**	33.8†	63.5	66.6

Table 17: Comparing performance of different models on all languages in POS. Metric: F1 Score. (All numbers are Monolingual results except the ones marked with † symbol which indicate Zero-Shot Cross-Lingual results (due to the absence of training data in those languages)

Model	en	af	ar	az	bg	bn	de	el	es	et	eu	fa	fi	fr	gu	he	hi	hu	id	it	ja	jv	ka	kk	
Fine-tuned Baselines																									
mBERT	**86.4**	76.1	42.9	65.5	76.7	69.7	79.5	70.9	**75.3**	75.8	**64.4**	40.0	76.6	79.6	51.3	**56.2**	65.9	76.1	61.0	**81.3**	**29.2**	62.4	65.1	50.3	
XLM-R Large	85.4	**78.6**	47.3	**69.4**	**80.9**	**74.7**	**80.7**	**79.2**	71.8	**78.7**	61.6	55.2	**79.6**	**79.8**	**62.7**	55.5	**70.9**	**80.2**	51.8	80.3	18.5	61.9	**70.9**	**54.4**	
Open AI Models																									
`gpt-3.5-turbo`	43.2	43.8	45.4	42.1	51.6	40.3	52.7	41.0	60.2	58.7	31.5	39.3	59.1	50.7	18.4	34.3	45.5	53.7	58.4	60.0	7.4	57.7	25.1	30.9	
`gpt-4-32k`	49.7	55.9	**59.4**	59.6	62.6	52.7	69.2	54.4	68.6	74.4	57.8	**67.6**	71.1	68.5	23.8	48.0	59.4	71.9	**72.7**	72.8	9.2	**68.8**	31.6	45.3	
	ko	lt	ml	mr	ms	my	nl	pa	pl	pt	qu	ro	ru	sw	ta	te	th	tl	tr	uk	ur	vi	yo	zh	**avg**
Fine-tuned Baselines																									
mBERT	**59.5**	**75.8**	53.0	57.0	67.1	45.7	81.0	30.5	79.2	**80.4**	58.5	74.0	63.9	**71.4**	50.7	48.9	0.4	72.6	73.4	69.7	35.4	74.5	45.8	**42.5**	62.3
XLM-R Large	59.2	**75.8**	**60.2**	**63.4**	**68.5**	**55.2**	**83.2**	49.4	**79.3**	79.9	58.5	**78.7**	**71.9**	68.9	**58.4**	**53.8**	0.7	**74.7**	**80.3**	**78.0**	60.3	**78.3**	37.0	26.6	**65.2**
Open AI Models																									
`gpt-3.5-turbo`	27.9	51.9	25.2	34.4	52.0	8.7	59.4	36.7	58.4	48.9	41.9	42.7	29.4	57.7	26.0	22.0	1.7	36.5	50.5	34.4	35.7	33.5	56.9	13.3	40.3
`gpt-4-32k`	51.4	71.3	35.6	47.4	64.1	16.3	67.9	**49.8**	70.3	64.5	**69.8**	59.6	64.8	68.9	36.9	33.0	**2.5**	61.9	72.9	58.4	**69.6**	58.4	**73.9**	18.5	55.5

Table 18: Comparing performance of different models on all languages in PAN-X. Metric: F1 Score.

Model	ar	en	es	eu	hi	id	my	ru	sw	te	zh	**avg**
Prompt-Based Baselines												
BLOOMZ	79.7	95.7	87.3	70.5	79.9	85.6	49.9	67.3	65.3	67.4	90.0	76.2
XGLM	59.8	75.9	69.2	63.8	62.5	70.8	61.2	72.4	65.2	63.4	67.7	66.5
Open AI Models												
`gpt-3.5-turbo`	92.5	96.8	95.8	78.4	91.1	95.0	57.2	96.6	92.3	73.1	95.6	87.7
`gpt-3.5-turbo` (TT)	94.3	96.8	96.1	92.5	94.7	95.2	88.6	96.2	88.7	93.6	95.6	93.9
`text-davinci-003`	87.4	98.3	97.6	78.1	77.8	96.4	47.4	94.2	78.1	57.6	95.0	82.5
`text-davinci-003` (TT)	95.0	98.3	96.2	94.1	95.1	95.9	90.1	96.9	90.7	94.3	96.2	94.8
`gpt-4-32k`	**99.1**	**99.6**	**99.5**	**97.6**	**98.8**	**99.0**	77.6	99.1	**98.4**	93.4	**99.2**	96.5
`gpt-4-32k` (TT)	97.7	**99.6**	98.7	96.8	97.9	98.1	**93.2**	**99.2**	93.6	**96.4**	98.3	**97.0**

Table 19: Comparing performance of different models on all languages in XStoryCloze. Metric: Accuracy.

	Google			Microsoft			Amazon			Systran			GPT Turbo 3.5			Bloomz		
	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S
es	50.9	23.2	20.9	45	36.5	22.9	**57.2**	15.3	21.7	42.5	46.2	15.6	54.9	22.7	26.2	55.6	17.2	32.5
fr	**61.6**	6.1	22.3	44.5	34.2	15.8	54.2	16.4	15	43.4	41.8	-0.1	52.7	21.4	26.1	52	17.8	24.6
it	38.6	32.9	18.6	38.8	41.8	10.5	40.2	26.8	14.7	38.1	47.3	6.3	45.1	21.9	26.7	**45.7**	9	18.5
ru	37.8	36.7	11.4	36.9	42	8.4	39.8	34.8	9.4	37.3	44.1	9.2	**41**	31.6	10.2	5.9	INV	0
uk	38.4	43.5	10.7	41.3	46.8	11.9	-	-	-	28.9	22.4	12.9	**42.9**	34.2	12.1	16.8	22.7	2.2
he	50.8	11.7	35.5	44	22	29.8	48	13.6	45.9	43.1	26.9	23.1	**57.5**	7.6	40.8	27.5	31.4	5
ar	45.8	42.5	16.2	45	47.1	14.2	48.3	37.8	18.8	45.6	49.4	-4.1	**61.1**	13.9	27.9	48.1	23	25.6
de	59.4	12.5	12.6	**74.1**	0	8.8	62.4	12	16.7	48.5	34.5	10	57.5	19.5	14.2	47.6	56.2	6.6

Table 20: Performance of commercial MT systems and LLMs on the WinoMT corpus on 8 target languages. Results are categorized by language family. Acc indicates overall gender accuracy (% of instances the translation had the correct gender), Δ_G denotes the difference in performance (F1 score) between masculine and feminine scores, and Δ_S is the difference in performance (F1 score) between pro-stereotypical and anti-stereotypical gender role assignments (higher numbers in the two latter metrics indicate stronger biases). Numbers in bold indicate best accuracy for the language across all systems. Notes: [1. For Google, Microsoft, Amazon, and Systran we use the translations provided by (Stanovsky et al., 2019). Some values differ from the original paper due to updated Spcay modules. 2. For Ru in Bloomz, Precision in male predictions is 0 leading to Invalid (INV) in Δ_G]

Model	es	fr	it	pt	ru	tr	**avg**
LLM Baselines							
PALM (0-Shot)	79.83	78.99	-	77.58	80.35	84.1	80.17
PALM (10-Shot Monolingual)	**91.23**	86.16	-	90.99	92.47	84.5	89.07
PALM-2 (0-Shot)	88.6	84.11	-	87.68	90.5	93.42	88.86
PALM-2 (10-Shot Monolingual)	89.68	**87.94**	-	**92.05**	**94.25**	**94.34**	**91.65**
OpenAI Models							
gpt-3.5-turbo (Crosslingual)	77.27	73.64	80.05	81.16	74.99	85.65	78.79
gpt-3.5-turbo (TT)	74.20	70.09	76.67	72.66	73.68	82.99	75.05
text-davinci-003 (Crosslingual)	79	74.55	**81.11**	81.63	79.13	93.55	81.50
text-davinci-003 (TT)	79.06	72.93	78.93	75.18	80.48	93.22	79.97

Table 21: Comparing performance of different models on all languages in Jigsaw. Metric: Accuracy.