

Table 7: **Updated Zero-Shot Performance on SuperGLUE.** ReCoRD uses the F1 score as the evaluation metric. All other benchmarks use accuracy.

Model	BoolQ	RTE	CB	COPA	ReCoRD	WIC	WSC	MultiRC	Average
GPT-2	48.38	51.99	41.07	62.0	71.07	49.53	43.27	53.47	52.6
Kraken 124M (2-way)	53.85	54.15	44.64	68.0	72.42	49.53	36.54	56.89	54.5
Kraken 124M (4-way)	50.15	53.79	41.07	67.0	71.47	50.31	37.5	57.16	53.56
Kraken 124M (6-way)	47.92	56.68	8.93	69.0	70.69	50.16	53.85	53.82	51.38
GPT-2 Medium	58.53	53.07	42.86	68.0	79.43	50.0	41.35	52.58	55.73
Kraken 355M (2-way)	60.06	51.62	41.07	69.0	80.04	50.0	35.58	56.91	55.54
Kraken 355M (4-way)	61.68	55.23	35.71	72.0	79.01	50.0	36.54	57.2	55.92
GPT-2 Large	60.55	52.71	41.07	72.0	81.95	49.69	44.23	48.56	56.34
Kraken 760M (4-way)	60.58	49.1	10.71	73.0	82.04	50.0	36.54	51.65	51.7

Table 8: **Updated Three-Shot Performance on SuperGLUE.** ReCoRD uses the F1 score as the evaluation metric. All other benchmarks use accuracy.

Model	BoolQ	RTE	CB	COPA	ReCoRD	WIC	WSC	MultiRC	Average
GPT-2	53.76	47.65	42.86	60.0	70.12	50.0	51.92	50.62	53.37
Kraken 124M (2-way)	55.14	48.74	44.64	61.0	68.83	49.69	47.12	50.56	53.22
Kraken 124M (4-way)	59.05	44.04	48.21	64.0	69.94	47.49	51.92	54.17	54.85
Kraken 124M (6-way)	54.28	46.93	42.86	63.0	69.17	49.37	44.23	51.96	52.72
GPT-2 Medium	60.58	47.65	44.64	68.0	78.35	47.65	50.0	53.22	56.26
Kraken 355M (2-way)	50.73	53.07	37.5	68.0	79.01	49.06	66.35	51.34	56.88
Kraken 355M (4-way)	61.8	51.99	50.0	73.0	77.97	47.49	44.23	54.17	57.58
GPT-2 Large	60.58	51.99	39.29	70.0	81.05	47.81	60.58	51.53	57.85
Kraken 760M (4-way)	55.75	53.79	42.86	68.0	81.03	45.77	62.5	53.47	57.9