Model Evaluation

Model Statistics:

Model	Correct Answers	Partially Correct Answers	Wrong Answers
deepset/tinyroberta-squad2	26 (65.0%)	5 (12.5%)	9 (22.5%)
deepset/roberta-base-squad2-covid	26 (65.0%)	5 (12.5%)	9 (22.5%)

Response Times Statistics:

Model	Average	Median	Min	Max
deepset/tinyroberta-squad2	7.2	7.1	6.9	7.5
deepset/roberta-base-squad2-covid	9.5	9.4	9.0	10.3

***All the response times are in seconds*

Model answers confidence scores:

For deepset/tinyroberta-squad2 model:

Range	Correct	Partially Correct	Wrong
0.0 - 0.1	2 (7.7%)	0 (0.0%)	2 (22.2%)
0.1 - 0.2	2 (7.7%)	0 (0.0%)	0 (0.0%)
0.2 - 0.3	1 (3.8%)	1 (20.0%)	1 (11.1%)
0.3 - 0.4	0 (0.0%)	1 (20.0%)	0 (0.0%)
0.4 - 0.5	4 (15.4%)	1 (20.0%)	2 (22.2%)
0.5 - 0.6	2 (7.7%)	1 (20.0%)	1 (11.1%)
0.6 - 0.7	4 (15.4%)	0 (0.0%)	0 (0.0%)
0.7 - 0.8	2 (7.7%)	1 (20.0%)	0 (0.0%)
0.8 - 0.9	3 (11.5%)	0 (0.0%)	2 (22.2%)
0.9 - 1.0	6 (23.1%)	0 (0.0%)	1 (11.1%)

For deepset/roberta-base-squad2-covid model:

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*3 (11.5%)*	*0 (0.0%)*	*5 (55.6%)*
0.1 - 0.2	*0 (0.0%)*	*3 (60.0%)*	*0 (0.0%)*
0.2 - 0.3	*2 (7.7%)*	*0 (0.0%)*	*0 (0.0%)*
0.3 - 0.4	*0 (0.0%)*	*0 (0.0%)*	*0 (0.0%)*
0.4 - 0.5	*1 (3.8%)*	*0 (0.0%)*	*1 (11.1%)*
0.5 - 0.6	*4 (15.4%)*	*0 (0.0%)*	*1 (11.1%)*
0.6 - 0.7	*1 (3.8%)*	*1 (20.0%)*	*0 (0.0%)*
0.7 - 0.8	*3 (11.5%)*	*1 (20.0%)*	*1 (11.1%)*
0.8 - 0.9	*4 (15.4%)*	*0 (0.0%)*	*1 (11.1%)*
0.9 - 1.0	*8 (30.8%)*	*0 (0.0%)*	*0 (0.0%)*

Question Statistics:

22 questions were answered correctly by all models

6 questions were answered incorrectly by all models

30 questions were answered correctly by at least one model

8 questions were answered correctly by exactly one model

28 questions were answered correctly or partially correct by all models