Model Evaluation

Model Statistics:

Model	*Correct Answers*	*Partially Correct Answers*	*Wrong Answers*
deepset-xlm-roberta-large-squad2	*1016 (85.4%)*	*79 (6.6%)*	*95 (8.0%)*

Response Times Statistics:

Model	*Average*	*Median*	*Min*	*Max*
deepset-xlm-roberta-large-squad2	*13.1*	*13.0*	*12.3*	*25.8*

***All the response times are in seconds*

Model answers confidence scores:

For deepset-xlm-roberta-large-squad2 model:

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*108 (10.6%)*	*24 (30.4%)*	*33 (34.7%)*
0.1 - 0.2	*39 (3.8%)*	*3 (3.8%)*	*13 (13.7%)*
0.2 - 0.3	*42 (4.1%)*	*10 (12.7%)*	*13 (13.7%)*
0.3 - 0.4	*71 (7.0%)*	*5 (6.3%)*	*5 (5.3%)*
0.4 - 0.5	*65 (6.4%)*	*6 (7.6%)*	*8 (8.4%)*
0.5 - 0.6	*87 (8.6%)*	*8 (10.1%)*	*7 (7.4%)*
0.6 - 0.7	*90 (8.9%)*	*2 (2.5%)*	*5 (5.3%)*
0.7 - 0.8	*115 (11.3%)*	*7 (8.9%)*	*2 (2.1%)*
0.8 - 0.9	*131 (12.9%)*	*8 (10.1%)*	*5 (5.3%)*
0.9 - 1.0	*268 (26.4%)*	*6 (7.6%)*	*4 (4.2%)*

Question Statistics:

1016 questions were answered correctly by all models

95 questions were answered incorrectly by all models

1016 questions were answered correctly by at least one model

1016 questions were answered correctly by exactly one model

1095 questions were answered correctly or partially correct by all models