Model Evaluation

Model Statistics:

Model	Correct Answers	Partially Correct Answers	Wrong Answers
bert-large-cased-whole-word-masking-finetuned-squad	144 (72.0%)	8 (4.0%)	48 (24.0%)
deepset/bert-large-uncased-whole-word-masking-squad2	144 (72.0%)	7 (3.5%)	49 (24.5%)
bert-large-uncased-whole-word-masking-finetuned-squad	144 (72.0%)	6 (3.0%)	50 (25.0%)
deepset/roberta-base-squad2	142 (71.0%)	7 (3.5%)	51 (25.5%)
dmis-lab/biobert-large-cased-v1.1-squad	141 (70.5%)	9 (4.5%)	50 (25.0%)
deepset/minilm-uncased-squad2	134 (67.0%)	10 (5.0%)	56 (28.0%)
rsvp-ai/bertserini-bert-base-squad	134 (67.0%)	8 (4.0%)	58 (29.0%)
distilbert-base-uncased-distilled-squad	130 (65.0%)	8 (4.0%)	62 (31.0%)
distilbert-base-cased-distilled-squad	125 (62.5%)	8 (4.0%)	67 (33.5%)
deepset/bert-base-cased-squad2	123 (61.5%)	11 (5.5%)	66 (33.0%)

Response Times Statistics:

Model	Average	Median	Min	Max
distilbert-base-uncased-distilled-squad	11.0	8.1	7.3	37.7
deepset/minilm-uncased-squad2	11.2	8.3	7.7	31.0
distilbert-base-cased-distilled-squad	12.0	8.5	7.8	99.8
deepset/bert-base-cased-squad2	12.6	9.4	8.2	35.3
rsvp-ai/bertserini-bert-base-squad	12.6	9.5	8.2	37.0
deepset/roberta-base-squad2	13.4	10.6	9.3	31.6
bert-large-cased-whole-word-masking-finetuned-squad	14.8	13.0	9.5	41.9
bert-large-uncased-whole-word-masking-finetuned-squad	15.9	13.2	10.2	59.8
deepset/bert-large-uncased-whole-word-masking-squad2	16.2	13.7	10.1	45.9

dmis-lab/biobert-large-cased-v1.1-squad	*16.5*	*14.2*	*10.3*	*37.3*

***All the response times are in seconds*

Model answers confidence scores:

For deepset/roberta-base-squad2 model:

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*6 (4.2%)*	*1 (14.3%)*	*13 (25.5%)*
0.1 - 0.2	*6 (4.2%)*	*1 (14.3%)*	*6 (11.8%)*
0.2 - 0.3	*17 (12.0%)*	*0 (0.0%)*	*5 (9.8%)*
0.3 - 0.4	*10 (7.0%)*	*4 (57.1%)*	*14 (27.5%)*
0.4 - 0.5	*25 (17.6%)*	*1 (14.3%)*	*2 (3.9%)*
0.5 - 0.6	*25 (17.6%)*	*0 (0.0%)*	*6 (11.8%)*
0.6 - 0.7	*12 (8.5%)*	*0 (0.0%)*	*0 (0.0%)*
0.7 - 0.8	*13 (9.2%)*	*0 (0.0%)*	*3 (5.9%)*
0.8 - 0.9	*14 (9.9%)*	*0 (0.0%)*	*0 (0.0%)*
0.9 - 1.0	*14 (9.9%)*	*0 (0.0%)*	*2 (3.9%)*

For bert-large-uncased-whole-word-masking-finetuned-squad model:

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*2 (1.4%)*	*0 (0.0%)*	*7 (14.0%)*
0.1 - 0.2	*7 (4.9%)*	*0 (0.0%)*	*7 (14.0%)*
0.2 - 0.3	*12 (8.3%)*	*0 (0.0%)*	*6 (12.0%)*
0.3 - 0.4	*14 (9.7%)*	*1 (16.7%)*	*6 (12.0%)*
0.4 - 0.5	*15 (10.4%)*	*3 (50.0%)*	*6 (12.0%)*
0.5 - 0.6	*23 (16.0%)*	*1 (16.7%)*	*4 (8.0%)*
0.6 - 0.7	*17 (11.8%)*	*1 (16.7%)*	*6 (12.0%)*
0.7 - 0.8	*14 (9.7%)*	*0 (0.0%)*	*4 (8.0%)*

0.8 - 0.9	*14 (9.7%)*	*0 (0.0%)*	*1 (2.0%)*
0.9 - 1.0	*26 (18.1%)*	*0 (0.0%)*	*3 (6.0%)*

For distilbert-base-cased-distilled-squad model:

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*2 (1.6%)*	*0 (0.0%)*	*6 (9.0%)*
0.1 - 0.2	*4 (3.2%)*	*0 (0.0%)*	*18 (26.9%)*
0.2 - 0.3	*12 (9.6%)*	*1 (12.5%)*	*4 (6.0%)*
0.3 - 0.4	*14 (11.2%)*	*2 (25.0%)*	*8 (11.9%)*
0.4 - 0.5	*12 (9.6%)*	*1 (12.5%)*	*7 (10.4%)*
0.5 - 0.6	*21 (16.8%)*	*0 (0.0%)*	*9 (13.4%)*
0.6 - 0.7	*10 (8.0%)*	*2 (25.0%)*	*2 (3.0%)*
0.7 - 0.8	*15 (12.0%)*	*0 (0.0%)*	*4 (6.0%)*
0.8 - 0.9	*8 (6.4%)*	*2 (25.0%)*	*3 (4.5%)*
0.9 - 1.0	*27 (21.6%)*	*0 (0.0%)*	*6 (9.0%)*

For deepset/bert-large-uncased-whole-word-masking-squad2 model:

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*7 (4.9%)*	*2 (28.6%)*	*4 (8.2%)*
0.1 - 0.2	*4 (2.8%)*	*3 (42.9%)*	*9 (18.4%)*
0.2 - 0.3	*10 (6.9%)*	*0 (0.0%)*	*6 (12.2%)*
0.3 - 0.4	*16 (11.1%)*	*1 (14.3%)*	*4 (8.2%)*
0.4 - 0.5	*23 (16.0%)*	*0 (0.0%)*	*4 (8.2%)*
0.5 - 0.6	*15 (10.4%)*	*1 (14.3%)*	*7 (14.3%)*
0.6 - 0.7	*14 (9.7%)*	*0 (0.0%)*	*5 (10.2%)*
0.7 - 0.8	*18 (12.5%)*	*0 (0.0%)*	*7 (14.3%)*
0.8 - 0.9	*16 (11.1%)*	*0 (0.0%)*	*0 (0.0%)*
0.9 - 1.0	*21 (14.6%)*	*0 (0.0%)*	*3 (6.1%)*

For distilbert-base-uncased-distilled-squad model:

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*4 (3.1%)*	*0 (0.0%)*	*14 (22.6%)*
0.1 - 0.2	*4 (3.1%)*	*0 (0.0%)*	*8 (12.9%)*
0.2 - 0.3	*12 (9.2%)*	*2 (25.0%)*	*7 (11.3%)*
0.3 - 0.4	*13 (10.0%)*	*1 (12.5%)*	*6 (9.7%)*
0.4 - 0.5	*17 (13.1%)*	*3 (37.5%)*	*6 (9.7%)*
0.5 - 0.6	*12 (9.2%)*	*0 (0.0%)*	*5 (8.1%)*
0.6 - 0.7	*17 (13.1%)*	*0 (0.0%)*	*3 (4.8%)*
0.7 - 0.8	*18 (13.8%)*	*2 (25.0%)*	*2 (3.2%)*
0.8 - 0.9	*9 (6.9%)*	*0 (0.0%)*	*5 (8.1%)*
0.9 - 1.0	*24 (18.5%)*	*0 (0.0%)*	*6 (9.7%)*

For rsvp-ai/bertserini-bert-base-squad model:

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*2 (1.5%)*	*2 (25.0%)*	*10 (17.2%)*
0.1 - 0.2	*7 (5.2%)*	*0 (0.0%)*	*12 (20.7%)*
0.2 - 0.3	*8 (6.0%)*	*2 (25.0%)*	*13 (22.4%)*
0.3 - 0.4	*15 (11.2%)*	*1 (12.5%)*	*7 (12.1%)*
0.4 - 0.5	*20 (14.9%)*	*0 (0.0%)*	*3 (5.2%)*
0.5 - 0.6	*17 (12.7%)*	*2 (25.0%)*	*2 (3.4%)*
0.6 - 0.7	*14 (10.4%)*	*1 (12.5%)*	*3 (5.2%)*
0.7 - 0.8	*15 (11.2%)*	*0 (0.0%)*	*2 (3.4%)*
0.8 - 0.9	*14 (10.4%)*	*0 (0.0%)*	*4 (6.9%)*
0.9 - 1.0	*22 (16.4%)*	*0 (0.0%)*	*2 (3.4%)*

For deepset/minilm-uncased-squad2 model:

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*3 (2.2%)*	*2 (20.0%)*	*9 (16.1%)*

0.1 - 0.2	*6 (4.5%)*	*0 (0.0%)*	*9 (16.1%)*
0.2 - 0.3	*11 (8.2%)*	*1 (10.0%)*	*5 (8.9%)*
0.3 - 0.4	*9 (6.7%)*	*1 (10.0%)*	*9 (16.1%)*
0.4 - 0.5	*18 (13.4%)*	*2 (20.0%)*	*3 (5.4%)*
0.5 - 0.6	*20 (14.9%)*	*1 (10.0%)*	*6 (10.7%)*
0.6 - 0.7	*18 (13.4%)*	*2 (20.0%)*	*6 (10.7%)*
0.7 - 0.8	*15 (11.2%)*	*0 (0.0%)*	*5 (8.9%)*
0.8 - 0.9	*10 (7.5%)*	*0 (0.0%)*	*3 (5.4%)*
0.9 - 1.0	*24 (17.9%)*	*1 (10.0%)*	*1 (1.8%)*

For dmis-lab/biobert-large-cased-v1.1-squad model:

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*2 (1.4%)*	*1 (11.1%)*	*8 (16.0%)*
0.1 - 0.2	*8 (5.7%)*	*0 (0.0%)*	*4 (8.0%)*
0.2 - 0.3	*9 (6.4%)*	*2 (22.2%)*	*6 (12.0%)*
0.3 - 0.4	*15 (10.6%)*	*1 (11.1%)*	*8 (16.0%)*
0.4 - 0.5	*10 (7.1%)*	*0 (0.0%)*	*8 (16.0%)*
0.5 - 0.6	*16 (11.3%)*	*1 (11.1%)*	*6 (12.0%)*
0.6 - 0.7	*19 (13.5%)*	*1 (11.1%)*	*2 (4.0%)*
0.7 - 0.8	*17 (12.1%)*	*1 (11.1%)*	*3 (6.0%)*
0.8 - 0.9	*16 (11.3%)*	*2 (22.2%)*	*2 (4.0%)*
0.9 - 1.0	*29 (20.6%)*	*0 (0.0%)*	*3 (6.0%)*

For deepset/bert-base-cased-squad2 model:

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*3 (2.4%)*	*4 (36.4%)*	*20 (30.3%)*
0.1 - 0.2	*6 (4.9%)*	*1 (9.1%)*	*9 (13.6%)*
0.2 - 0.3	*13 (10.6%)*	*0 (0.0%)*	*10 (15.2%)*

0.3 - 0.4	*10 (8.1%)*	*3 (27.3%)*	*7 (10.6%)*
0.4 - 0.5	*11 (8.9%)*	*1 (9.1%)*	*7 (10.6%)*
0.5 - 0.6	*18 (14.6%)*	*0 (0.0%)*	*4 (6.1%)*
0.6 - 0.7	*16 (13.0%)*	*2 (18.2%)*	*2 (3.0%)*
0.7 - 0.8	*8 (6.5%)*	*0 (0.0%)*	*1 (1.5%)*
0.8 - 0.9	*11 (8.9%)*	*0 (0.0%)*	*0 (0.0%)*
0.9 - 1.0	*27 (22.0%)*	*0 (0.0%)*	*6 (9.1%)*

For bert-large-cased-whole-word-masking-finetuned-squad model:

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*2 (1.4%)*	*1 (12.5%)*	*8 (16.7%)*
0.1 - 0.2	*3 (2.1%)*	*3 (37.5%)*	*5 (10.4%)*
0.2 - 0.3	*15 (10.4%)*	*1 (12.5%)*	*5 (10.4%)*
0.3 - 0.4	*13 (9.0%)*	*0 (0.0%)*	*8 (16.7%)*
0.4 - 0.5	*17 (11.8%)*	*1 (12.5%)*	*7 (14.6%)*
0.5 - 0.6	*19 (13.2%)*	*0 (0.0%)*	*4 (8.3%)*
0.6 - 0.7	*19 (13.2%)*	*2 (25.0%)*	*2 (4.2%)*
0.7 - 0.8	*16 (11.1%)*	*0 (0.0%)*	*5 (10.4%)*
0.8 - 0.9	*17 (11.8%)*	*0 (0.0%)*	*2 (4.2%)*
0.9 - 1.0	*23 (16.0%)*	*0 (0.0%)*	*2 (4.2%)*

Question Statistics:

84 questions were answered correctly by all models

22 questions were answered incorrectly by all models

175 questions were answered correctly by at least one model

6 questions were answered correctly by exactly one model

94 questions were answered correctly or partially correct by all models