Model Evaluation

Model Statistics:

Model	Correct Answers	Partially Correct Answers	Wrong Answers
dmis-lab/biobert-large-cased-v1.1-squad	135 (67.5%)	13 (6.5%)	52 (26.0%)
deepset/bert-large-uncased-whole-word-masking-squad2	134 (67.0%)	11 (5.5%)	55 (27.5%)
bert-large-cased-whole-word-masking-finetuned-squad	130 (65.0%)	15 (7.5%)	55 (27.5%)
deepset/roberta-base-squad2	130 (65.0%)	11 (5.5%)	59 (29.5%)
bert-large-uncased-whole-word-masking-finetuned-squad	130 (65.0%)	11 (5.5%)	59 (29.5%)
deepset/minilm-uncased-squad2	122 (61.0%)	9 (4.5%)	69 (34.5%)
distilbert-base-uncased-distilled-squad	120 (60.0%)	10 (5.0%)	70 (35.0%)
distilbert-base-cased-distilled-squad	117 (58.5%)	10 (5.0%)	73 (36.5%)
rsvp-ai/bertserini-bert-base-squad	114 (57.0%)	15 (7.5%)	71 (35.5%)
deepset/bert-base-cased-squad2	109 (54.5%)	11 (5.5%)	80 (40.0%)

Response Times Statistics:

Model	Average	Median	Min	Max
distilbert-base-uncased-distilled-squad	7.5	7.5	7.2	8.0
deepset/minilm-uncased-squad2	7.9	7.9	7.5	8.3
distilbert-base-cased-distilled-squad	8.1	8.0	7.8	17.8
rsvp-ai/bertserini-bert-base-squad	8.6	8.5	8.1	9.5
deepset/bert-base-cased-squad2	8.6	8.5	8.2	9.7
deepset/roberta-base-squad2	9.3	9.3	8.9	10.3
bert-large-uncased-whole-word-masking-finetuned-squad	10.4	10.3	9.3	12.2
bert-large-cased-whole-word-masking-finetuned-squad	10.4	10.3	9.2	12.9
deepset/bert-large-uncased-whole-word-masking-squad2	11.0	10.8	9.8	14.5

dmis-lab/biobert-large-cased-v1.1-squad	*11.3*	*11.1*	*10.1*	*13.9*

***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	*8 (6.2%)*	*0 (0.0%)*	*14 (23.7%)*
0.1 - 0.2	*11 (8.5%)*	*1 (9.1%)*	*10 (16.9%)*
0.2 - 0.3	*12 (9.2%)*	*1 (9.1%)*	*8 (13.6%)*
0.3 - 0.4	*13 (10.0%)*	*2 (18.2%)*	*6 (10.2%)*
0.4 - 0.5	*22 (16.9%)*	*1 (9.1%)*	*6 (10.2%)*
0.5 - 0.6	*18 (13.8%)*	*2 (18.2%)*	*2 (3.4%)*
0.6 - 0.7	*10 (7.7%)*	*2 (18.2%)*	*4 (6.8%)*
0.7 - 0.8	*9 (6.9%)*	*1 (9.1%)*	*3 (5.1%)*
0.8 - 0.9	*11 (8.5%)*	*1 (9.1%)*	*4 (6.8%)*
0.9 - 1.0	*16 (12.3%)*	*0 (0.0%)*	*2 (3.4%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*2 (1.5%)*	*1 (9.1%)*	*6 (10.2%)*
0.1 - 0.2	*6 (4.6%)*	*1 (9.1%)*	*11 (18.6%)*
0.2 - 0.3	*10 (7.7%)*	*1 (9.1%)*	*12 (20.3%)*
0.3 - 0.4	*15 (11.5%)*	*1 (9.1%)*	*7 (11.9%)*
0.4 - 0.5	*17 (13.1%)*	*1 (9.1%)*	*2 (3.4%)*
0.5 - 0.6	*16 (12.3%)*	*2 (18.2%)*	*5 (8.5%)*
0.6 - 0.7	*15 (11.5%)*	*1 (9.1%)*	*6 (10.2%)*
0.7 - 0.8	*12 (9.2%)*	*1 (9.1%)*	*2 (3.4%)*

0.8 - 0.9	*13 (10.0%)*	*0 (0.0%)*	*4 (6.8%)*
0.9 - 1.0	*24 (18.5%)*	*2 (18.2%)*	*4 (6.8%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*5 (4.3%)*	*0 (0.0%)*	*6 (8.2%)*
0.1 - 0.2	*2 (1.7%)*	*2 (20.0%)*	*9 (12.3%)*
0.2 - 0.3	*10 (8.5%)*	*0 (0.0%)*	*9 (12.3%)*
0.3 - 0.4	*8 (6.8%)*	*2 (20.0%)*	*11 (15.1%)*
0.4 - 0.5	*16 (13.7%)*	*3 (30.0%)*	*8 (11.0%)*
0.5 - 0.6	*14 (12.0%)*	*0 (0.0%)*	*5 (6.8%)*
0.6 - 0.7	*9 (7.7%)*	*2 (20.0%)*	*4 (5.5%)*
0.7 - 0.8	*15 (12.8%)*	*1 (10.0%)*	*7 (9.6%)*
0.8 - 0.9	*12 (10.3%)*	*0 (0.0%)*	*10 (13.7%)*
0.9 - 1.0	*26 (22.2%)*	*0 (0.0%)*	*4 (5.5%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*4 (3.0%)*	*0 (0.0%)*	*10 (18.2%)*
0.1 - 0.2	*10 (7.5%)*	*0 (0.0%)*	*8 (14.5%)*
0.2 - 0.3	*12 (9.0%)*	*0 (0.0%)*	*8 (14.5%)*
0.3 - 0.4	*21 (15.7%)*	*2 (18.2%)*	*3 (5.5%)*
0.4 - 0.5	*15 (11.2%)*	*4 (36.4%)*	*10 (18.2%)*
0.5 - 0.6	*15 (11.2%)*	*1 (9.1%)*	*2 (3.6%)*
0.6 - 0.7	*14 (10.4%)*	*2 (18.2%)*	*5 (9.1%)*
0.7 - 0.8	*13 (9.7%)*	*1 (9.1%)*	*2 (3.6%)*
0.8 - 0.9	*11 (8.2%)*	*0 (0.0%)*	*5 (9.1%)*
0.9 - 1.0	*19 (14.2%)*	*1 (9.1%)*	*2 (3.6%)*

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

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*1 (0.8%)*	*1 (10.0%)*	*13 (18.6%)*
0.1 - 0.2	*8 (6.7%)*	*1 (10.0%)*	*10 (14.3%)*
0.2 - 0.3	*7 (5.8%)*	*2 (20.0%)*	*2 (2.9%)*
0.3 - 0.4	*12 (10.0%)*	*1 (10.0%)*	*12 (17.1%)*
0.4 - 0.5	*16 (13.3%)*	*1 (10.0%)*	*14 (20.0%)*
0.5 - 0.6	*10 (8.3%)*	*2 (20.0%)*	*2 (2.9%)*
0.6 - 0.7	*12 (10.0%)*	*1 (10.0%)*	*5 (7.1%)*
0.7 - 0.8	*15 (12.5%)*	*0 (0.0%)*	*5 (7.1%)*
0.8 - 0.9	*14 (11.7%)*	*0 (0.0%)*	*4 (5.7%)*
0.9 - 1.0	*25 (20.8%)*	*1 (10.0%)*	*3 (4.3%)*

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

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*3 (2.6%)*	*1 (6.7%)*	*6 (8.5%)*
0.1 - 0.2	*3 (2.6%)*	*4 (26.7%)*	*13 (18.3%)*
0.2 - 0.3	*10 (8.8%)*	*3 (20.0%)*	*10 (14.1%)*
0.3 - 0.4	*14 (12.3%)*	*2 (13.3%)*	*11 (15.5%)*
0.4 - 0.5	*15 (13.2%)*	*3 (20.0%)*	*8 (11.3%)*
0.5 - 0.6	*15 (13.2%)*	*1 (6.7%)*	*7 (9.9%)*
0.6 - 0.7	*8 (7.0%)*	*0 (0.0%)*	*7 (9.9%)*
0.7 - 0.8	*16 (14.0%)*	*0 (0.0%)*	*2 (2.8%)*
0.8 - 0.9	*5 (4.4%)*	*1 (6.7%)*	*5 (7.0%)*
0.9 - 1.0	*25 (21.9%)*	*0 (0.0%)*	*2 (2.8%)*

For deepset/minilm-uncased-squad2 model:

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*3 (2.5%)*	*0 (0.0%)*	*12 (17.4%)*

0.1 - 0.2	*6 (4.9%)*	*0 (0.0%)*	*8 (11.6%)*
0.2 - 0.3	*17 (13.9%)*	*0 (0.0%)*	*14 (20.3%)*
0.3 - 0.4	*14 (11.5%)*	*2 (22.2%)*	*6 (8.7%)*
0.4 - 0.5	*19 (15.6%)*	*1 (11.1%)*	*6 (8.7%)*
0.5 - 0.6	*12 (9.8%)*	*2 (22.2%)*	*6 (8.7%)*
0.6 - 0.7	*14 (11.5%)*	*1 (11.1%)*	*6 (8.7%)*
0.7 - 0.8	*9 (7.4%)*	*1 (11.1%)*	*6 (8.7%)*
0.8 - 0.9	*12 (9.8%)*	*2 (22.2%)*	*2 (2.9%)*
0.9 - 1.0	*16 (13.1%)*	*0 (0.0%)*	*3 (4.3%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*3 (2.2%)*	*1 (7.7%)*	*3 (5.8%)*
0.1 - 0.2	*5 (3.7%)*	*1 (7.7%)*	*7 (13.5%)*
0.2 - 0.3	*8 (5.9%)*	*2 (15.4%)*	*7 (13.5%)*
0.3 - 0.4	*13 (9.6%)*	*0 (0.0%)*	*9 (17.3%)*
0.4 - 0.5	*20 (14.8%)*	*2 (15.4%)*	*5 (9.6%)*
0.5 - 0.6	*20 (14.8%)*	*1 (7.7%)*	*9 (17.3%)*
0.6 - 0.7	*12 (8.9%)*	*2 (15.4%)*	*1 (1.9%)*
0.7 - 0.8	*13 (9.6%)*	*1 (7.7%)*	*7 (13.5%)*
0.8 - 0.9	*15 (11.1%)*	*2 (15.4%)*	*1 (1.9%)*
0.9 - 1.0	*26 (19.3%)*	*1 (7.7%)*	*3 (5.8%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*9 (8.3%)*	*1 (9.1%)*	*20 (25.0%)*
0.1 - 0.2	*9 (8.3%)*	*1 (9.1%)*	*18 (22.5%)*
0.2 - 0.3	*8 (7.3%)*	*4 (36.4%)*	*13 (16.2%)*

0.3 - 0.4	*8 (7.3%)*	*1 (9.1%)*	*8 (10.0%)*
0.4 - 0.5	*8 (7.3%)*	*0 (0.0%)*	*8 (10.0%)*
0.5 - 0.6	*12 (11.0%)*	*2 (18.2%)*	*3 (3.8%)*
0.6 - 0.7	*10 (9.2%)*	*0 (0.0%)*	*2 (2.5%)*
0.7 - 0.8	*11 (10.1%)*	*0 (0.0%)*	*5 (6.2%)*
0.8 - 0.9	*10 (9.2%)*	*1 (9.1%)*	*1 (1.2%)*
0.9 - 1.0	*24 (22.0%)*	*1 (9.1%)*	*2 (2.5%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*1 (0.8%)*	*1 (6.7%)*	*5 (9.1%)*
0.1 - 0.2	*8 (6.2%)*	*2 (13.3%)*	*8 (14.5%)*
0.2 - 0.3	*8 (6.2%)*	*3 (20.0%)*	*8 (14.5%)*
0.3 - 0.4	*16 (12.3%)*	*3 (20.0%)*	*9 (16.4%)*
0.4 - 0.5	*11 (8.5%)*	*1 (6.7%)*	*9 (16.4%)*
0.5 - 0.6	*22 (16.9%)*	*1 (6.7%)*	*5 (9.1%)*
0.6 - 0.7	*14 (10.8%)*	*3 (20.0%)*	*1 (1.8%)*
0.7 - 0.8	*18 (13.8%)*	*0 (0.0%)*	*2 (3.6%)*
0.8 - 0.9	*8 (6.2%)*	*0 (0.0%)*	*3 (5.5%)*
0.9 - 1.0	*24 (18.5%)*	*1 (6.7%)*	*5 (9.1%)*

Question Statistics:

67 questions were answered correctly by all models

26 questions were answered incorrectly by all models

167 questions were answered correctly by at least one model

13 questions were answered correctly by exactly one model

73 questions were answered correctly or partially correct by all models