Model Evaluation

Model Statistics:

Model	Correct Answers	Partially Correct Answers	Wrong Answers
bert-large-cased-whole-word-masking-finetuned-squad	943 (80.0%)	47 (4.0%)	189 (16.0%)
bert-large-uncased-whole-word-masking-finetuned-squad	938 (79.6%)	50 (4.2%)	191 (16.2%)
deepset/bert-large-uncased-whole-word-masking-squad2	939 (79.6%)	48 (4.1%)	192 (16.3%)
deepset/roberta-base-squad2	930 (78.9%)	50 (4.2%)	199 (16.9%)
dmis-lab/biobert-large-cased-v1.1-squad	921 (78.1%)	45 (3.8%)	213 (18.1%)
deepset/minilm-uncased-squad2	894 (75.8%)	51 (4.3%)	234 (19.8%)
distilbert-base-cased-distilled-squad	882 (74.8%)	53 (4.5%)	244 (20.7%)
distilbert-base-uncased-distilled-squad	876 (74.3%)	58 (4.9%)	245 (20.8%)
rsvp-ai/bertserini-bert-base-squad	882 (74.8%)	43 (3.6%)	254 (21.5%)
deepset/bert-base-cased-squad2	856 (72.6%)	54 (4.6%)	269 (22.8%)

Response Times Statistics:

Model	Average	Median	Min	Max
distilbert-base-uncased-distilled-squad	7.4	7.4	7.0	10.7
deepset/minilm-uncased-squad2	7.8	7.8	7.3	17.7
distilbert-base-cased-distilled-squad	8.0	8.0	7.5	17.4
deepset/bert-base-cased-squad2	8.4	8.3	7.8	10.4
rsvp-ai/bertserini-bert-base-squad	8.4	8.4	7.8	11.5
deepset/roberta-base-squad2	9.1	9.1	8.6	10.1
bert-large-uncased-whole-word-masking-finetuned-squad	9.6	9.5	8.7	11.4
bert-large-cased-whole-word-masking-finetuned-squad	9.6	9.5	8.9	11.8
deepset/bert-large-uncased-whole-word-masking-squad2	10.2	10.1	9.4	15.6

dmis-lab/biobert-large-cased-v1.1-squad	*10.5*	*10.4*	*9.6*	*25.3*

***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	*154 (16.6%)*	*22 (44.0%)*	*87 (43.7%)*
0.1 - 0.2	*66 (7.1%)*	*7 (14.0%)*	*30 (15.1%)*
0.2 - 0.3	*61 (6.6%)*	*4 (8.0%)*	*20 (10.1%)*
0.3 - 0.4	*78 (8.4%)*	*2 (4.0%)*	*13 (6.5%)*
0.4 - 0.5	*81 (8.7%)*	*4 (8.0%)*	*17 (8.5%)*
0.5 - 0.6	*84 (9.0%)*	*2 (4.0%)*	*9 (4.5%)*
0.6 - 0.7	*84 (9.0%)*	*3 (6.0%)*	*9 (4.5%)*
0.7 - 0.8	*70 (7.5%)*	*1 (2.0%)*	*6 (3.0%)*
0.8 - 0.9	*99 (10.6%)*	*2 (4.0%)*	*3 (1.5%)*
0.9 - 1.0	*153 (16.5%)*	*3 (6.0%)*	*5 (2.5%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*18 (1.9%)*	*4 (8.0%)*	*16 (8.4%)*
0.1 - 0.2	*36 (3.8%)*	*7 (14.0%)*	*24 (12.6%)*
0.2 - 0.3	*41 (4.4%)*	*4 (8.0%)*	*41 (21.5%)*
0.3 - 0.4	*77 (8.2%)*	*6 (12.0%)*	*22 (11.5%)*
0.4 - 0.5	*82 (8.7%)*	*7 (14.0%)*	*22 (11.5%)*
0.5 - 0.6	*102 (10.9%)*	*6 (12.0%)*	*24 (12.6%)*
0.6 - 0.7	*82 (8.7%)*	*5 (10.0%)*	*10 (5.2%)*
0.7 - 0.8	*84 (9.0%)*	*4 (8.0%)*	*12 (6.3%)*

0.8 - 0.9	*104 (11.1%)*	*3 (6.0%)*	*7 (3.7%)*
0.9 - 1.0	*312 (33.3%)*	*4 (8.0%)*	*13 (6.8%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*19 (2.2%)*	*7 (13.2%)*	*25 (10.2%)*
0.1 - 0.2	*33 (3.7%)*	*2 (3.8%)*	*43 (17.6%)*
0.2 - 0.3	*49 (5.6%)*	*5 (9.4%)*	*40 (16.4%)*
0.3 - 0.4	*65 (7.4%)*	*5 (9.4%)*	*24 (9.8%)*
0.4 - 0.5	*81 (9.2%)*	*8 (15.1%)*	*28 (11.5%)*
0.5 - 0.6	*76 (8.6%)*	*5 (9.4%)*	*19 (7.8%)*
0.6 - 0.7	*76 (8.6%)*	*5 (9.4%)*	*14 (5.7%)*
0.7 - 0.8	*84 (9.5%)*	*4 (7.5%)*	*16 (6.6%)*
0.8 - 0.9	*101 (11.5%)*	*2 (3.8%)*	*15 (6.1%)*
0.9 - 1.0	*298 (33.8%)*	*10 (18.9%)*	*20 (8.2%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*168 (17.9%)*	*18 (37.5%)*	*85 (44.3%)*
0.1 - 0.2	*63 (6.7%)*	*7 (14.6%)*	*19 (9.9%)*
0.2 - 0.3	*52 (5.5%)*	*3 (6.2%)*	*20 (10.4%)*
0.3 - 0.4	*67 (7.1%)*	*3 (6.2%)*	*12 (6.2%)*
0.4 - 0.5	*64 (6.8%)*	*3 (6.2%)*	*12 (6.2%)*
0.5 - 0.6	*89 (9.5%)*	*2 (4.2%)*	*10 (5.2%)*
0.6 - 0.7	*82 (8.7%)*	*4 (8.3%)*	*12 (6.2%)*
0.7 - 0.8	*75 (8.0%)*	*2 (4.2%)*	*7 (3.6%)*
0.8 - 0.9	*77 (8.2%)*	*2 (4.2%)*	*5 (2.6%)*
0.9 - 1.0	*202 (21.5%)*	*4 (8.3%)*	*10 (5.2%)*

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

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*15 (1.7%)*	*6 (10.3%)*	*28 (11.4%)*
0.1 - 0.2	*37 (4.2%)*	*7 (12.1%)*	*39 (15.9%)*
0.2 - 0.3	*54 (6.2%)*	*5 (8.6%)*	*33 (13.5%)*
0.3 - 0.4	*68 (7.8%)*	*8 (13.8%)*	*37 (15.1%)*
0.4 - 0.5	*83 (9.5%)*	*4 (6.9%)*	*23 (9.4%)*
0.5 - 0.6	*99 (11.3%)*	*8 (13.8%)*	*26 (10.6%)*
0.6 - 0.7	*83 (9.5%)*	*6 (10.3%)*	*17 (6.9%)*
0.7 - 0.8	*77 (8.8%)*	*5 (8.6%)*	*16 (6.5%)*
0.8 - 0.9	*77 (8.8%)*	*6 (10.3%)*	*10 (4.1%)*
0.9 - 1.0	*283 (32.3%)*	*3 (5.2%)*	*16 (6.5%)*

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

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*23 (2.6%)*	*2 (4.7%)*	*39 (15.4%)*
0.1 - 0.2	*37 (4.2%)*	*4 (9.3%)*	*41 (16.1%)*
0.2 - 0.3	*68 (7.7%)*	*6 (14.0%)*	*47 (18.5%)*
0.3 - 0.4	*57 (6.5%)*	*8 (18.6%)*	*30 (11.8%)*
0.4 - 0.5	*81 (9.2%)*	*3 (7.0%)*	*22 (8.7%)*
0.5 - 0.6	*73 (8.3%)*	*5 (11.6%)*	*20 (7.9%)*
0.6 - 0.7	*72 (8.2%)*	*0 (0.0%)*	*17 (6.7%)*
0.7 - 0.8	*94 (10.7%)*	*8 (18.6%)*	*11 (4.3%)*
0.8 - 0.9	*120 (13.6%)*	*2 (4.7%)*	*13 (5.1%)*
0.9 - 1.0	*257 (29.1%)*	*5 (11.6%)*	*14 (5.5%)*

For deepset/minilm-uncased-squad2 model:

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*95 (10.6%)*	*16 (31.4%)*	*76 (32.5%)*

0.1 - 0.2	*50 (5.6%)*	*4 (7.8%)*	*32 (13.7%)*
0.2 - 0.3	*55 (6.2%)*	*7 (13.7%)*	*30 (12.8%)*
0.3 - 0.4	*67 (7.5%)*	*3 (5.9%)*	*30 (12.8%)*
0.4 - 0.5	*78 (8.7%)*	*2 (3.9%)*	*16 (6.8%)*
0.5 - 0.6	*75 (8.4%)*	*10 (19.6%)*	*19 (8.1%)*
0.6 - 0.7	*76 (8.5%)*	*5 (9.8%)*	*5 (2.1%)*
0.7 - 0.8	*76 (8.5%)*	*1 (2.0%)*	*13 (5.6%)*
0.8 - 0.9	*100 (11.2%)*	*0 (0.0%)*	*5 (2.1%)*
0.9 - 1.0	*222 (24.8%)*	*3 (5.9%)*	*8 (3.4%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*13 (1.4%)*	*6 (13.3%)*	*25 (11.7%)*
0.1 - 0.2	*24 (2.6%)*	*2 (4.4%)*	*27 (12.7%)*
0.2 - 0.3	*42 (4.6%)*	*1 (2.2%)*	*31 (14.6%)*
0.3 - 0.4	*52 (5.6%)*	*5 (11.1%)*	*26 (12.2%)*
0.4 - 0.5	*60 (6.5%)*	*5 (11.1%)*	*10 (4.7%)*
0.5 - 0.6	*83 (9.0%)*	*4 (8.9%)*	*29 (13.6%)*
0.6 - 0.7	*96 (10.4%)*	*6 (13.3%)*	*16 (7.5%)*
0.7 - 0.8	*83 (9.0%)*	*3 (6.7%)*	*13 (6.1%)*
0.8 - 0.9	*119 (12.9%)*	*5 (11.1%)*	*13 (6.1%)*
0.9 - 1.0	*349 (37.9%)*	*8 (17.8%)*	*23 (10.8%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*207 (24.2%)*	*24 (44.4%)*	*144 (53.5%)*
0.1 - 0.2	*51 (6.0%)*	*8 (14.8%)*	*22 (8.2%)*
0.2 - 0.3	*48 (5.6%)*	*6 (11.1%)*	*21 (7.8%)*

0.3 - 0.4	*50 (5.8%)*	*3 (5.6%)*	*16 (5.9%)*
0.4 - 0.5	*57 (6.7%)*	*3 (5.6%)*	*17 (6.3%)*
0.5 - 0.6	*58 (6.8%)*	*2 (3.7%)*	*14 (5.2%)*
0.6 - 0.7	*67 (7.8%)*	*3 (5.6%)*	*12 (4.5%)*
0.7 - 0.8	*66 (7.7%)*	*2 (3.7%)*	*7 (2.6%)*
0.8 - 0.9	*72 (8.4%)*	*1 (1.9%)*	*4 (1.5%)*
0.9 - 1.0	*180 (21.0%)*	*2 (3.7%)*	*12 (4.5%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*15 (1.6%)*	*5 (10.6%)*	*25 (13.2%)*
0.1 - 0.2	*30 (3.2%)*	*6 (12.8%)*	*24 (12.7%)*
0.2 - 0.3	*42 (4.5%)*	*3 (6.4%)*	*25 (13.2%)*
0.3 - 0.4	*67 (7.1%)*	*4 (8.5%)*	*22 (11.6%)*
0.4 - 0.5	*83 (8.8%)*	*8 (17.0%)*	*18 (9.5%)*
0.5 - 0.6	*85 (9.0%)*	*6 (12.8%)*	*20 (10.6%)*
0.6 - 0.7	*100 (10.6%)*	*2 (4.3%)*	*16 (8.5%)*
0.7 - 0.8	*91 (9.7%)*	*3 (6.4%)*	*14 (7.4%)*
0.8 - 0.9	*93 (9.9%)*	*4 (8.5%)*	*7 (3.7%)*
0.9 - 1.0	*337 (35.7%)*	*6 (12.8%)*	*18 (9.5%)*

Question Statistics:

668 questions were answered correctly by all models

89 questions were answered incorrectly by all models

1051 questions were answered correctly by at least one model

23 questions were answered correctly by exactly one model

733 questions were answered correctly or partially correct by all models