Model Evaluation

Model Statistics:

Model	*Correct Answers*	*Partially Correct Answers*	*Wrong Answers*
bert-large-cased-whole-word-masking-finetuned-squad	*839 (71.2%)*	*104 (8.8%)*	*236 (20.0%)*
bert-large-uncased-whole-word-masking-finetuned-squad	*828 (70.2%)*	*117 (9.9%)*	*234 (19.8%)*
deepset/bert-large-uncased-whole-word-masking-squad2	*831 (70.5%)*	*111 (9.4%)*	*237 (20.1%)*
deepset/roberta-base-squad2	*825 (70.0%)*	*98 (8.3%)*	*256 (21.7%)*
dmis-lab/biobert-large-cased-v1.1-squad	*815 (69.1%)*	*114 (9.7%)*	*250 (21.2%)*
deepset/minilm-uncased-squad2	*798 (67.7%)*	*104 (8.8%)*	*277 (23.5%)*
rsvp-ai/bertserini-bert-base-squad	*779 (66.1%)*	*108 (9.2%)*	*292 (24.8%)*
distilbert-base-uncased-distilled-squad	*764 (64.8%)*	*130 (11.0%)*	*285 (24.2%)*
distilbert-base-cased-distilled-squad	*760 (64.5%)*	*129 (10.9%)*	*290 (24.6%)*
deepset/bert-base-cased-squad2	*747 (63.4%)*	*120 (10.2%)*	*312 (26.5%)*

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	*116 (14.1%)*	*34 (34.7%)*	*113 (44.1%)*
0.1 - 0.2	*62 (7.5%)*	*10 (10.2%)*	*31 (12.1%)*
0.2 - 0.3	*52 (6.3%)*	*10 (10.2%)*	*23 (9.0%)*
0.3 - 0.4	*69 (8.4%)*	*7 (7.1%)*	*17 (6.6%)*
0.4 - 0.5	*68 (8.2%)*	*10 (10.2%)*	*24 (9.4%)*
0.5 - 0.6	*79 (9.6%)*	*5 (5.1%)*	*11 (4.3%)*
0.6 - 0.7	*71 (8.6%)*	*10 (10.2%)*	*15 (5.9%)*
0.7 - 0.8	*68 (8.2%)*	*1 (1.0%)*	*8 (3.1%)*
0.8 - 0.9	*94 (11.4%)*	*3 (3.1%)*	*7 (2.7%)*
0.9 - 1.0	*146 (17.7%)*	*8 (8.2%)*	*7 (2.7%)*

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

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*12 (1.4%)*	*7 (6.0%)*	*19 (8.1%)*
0.1 - 0.2	*26 (3.1%)*	*15 (12.8%)*	*26 (11.1%)*
0.2 - 0.3	*36 (4.3%)*	*10 (8.5%)*	*40 (17.1%)*
0.3 - 0.4	*63 (7.6%)*	*15 (12.8%)*	*27 (11.5%)*
0.4 - 0.5	*76 (9.2%)*	*10 (8.5%)*	*25 (10.7%)*
0.5 - 0.6	*85 (10.3%)*	*16 (13.7%)*	*31 (13.2%)*
0.6 - 0.7	*74 (8.9%)*	*7 (6.0%)*	*16 (6.8%)*
0.7 - 0.8	*70 (8.5%)*	*13 (11.1%)*	*17 (7.3%)*

0.8 - 0.9	*94 (11.4%)*	*10 (8.5%)*	*10 (4.3%)*
0.9 - 1.0	*292 (35.3%)*	*14 (12.0%)*	*23 (9.8%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*11 (1.4%)*	*11 (8.5%)*	*29 (10.0%)*
0.1 - 0.2	*23 (3.0%)*	*13 (10.1%)*	*42 (14.5%)*
0.2 - 0.3	*36 (4.7%)*	*15 (11.6%)*	*43 (14.8%)*
0.3 - 0.4	*52 (6.8%)*	*14 (10.9%)*	*28 (9.7%)*
0.4 - 0.5	*67 (8.8%)*	*23 (17.8%)*	*27 (9.3%)*
0.5 - 0.6	*68 (8.9%)*	*8 (6.2%)*	*24 (8.3%)*
0.6 - 0.7	*64 (8.4%)*	*9 (7.0%)*	*22 (7.6%)*
0.7 - 0.8	*69 (9.1%)*	*8 (6.2%)*	*27 (9.3%)*
0.8 - 0.9	*92 (12.1%)*	*8 (6.2%)*	*18 (6.2%)*
0.9 - 1.0	*278 (36.6%)*	*20 (15.5%)*	*30 (10.3%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*141 (17.0%)*	*28 (25.2%)*	*102 (43.0%)*
0.1 - 0.2	*54 (6.5%)*	*9 (8.1%)*	*26 (11.0%)*
0.2 - 0.3	*44 (5.3%)*	*10 (9.0%)*	*21 (8.9%)*
0.3 - 0.4	*59 (7.1%)*	*7 (6.3%)*	*16 (6.8%)*
0.4 - 0.5	*57 (6.9%)*	*9 (8.1%)*	*13 (5.5%)*
0.5 - 0.6	*80 (9.6%)*	*8 (7.2%)*	*13 (5.5%)*
0.6 - 0.7	*68 (8.2%)*	*14 (12.6%)*	*16 (6.8%)*
0.7 - 0.8	*67 (8.1%)*	*10 (9.0%)*	*7 (3.0%)*
0.8 - 0.9	*68 (8.2%)*	*7 (6.3%)*	*9 (3.8%)*
0.9 - 1.0	*193 (23.2%)*	*9 (8.1%)*	*14 (5.9%)*

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

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*12 (1.6%)*	*13 (10.0%)*	*24 (8.4%)*
0.1 - 0.2	*27 (3.5%)*	*11 (8.5%)*	*45 (15.8%)*
0.2 - 0.3	*43 (5.6%)*	*13 (10.0%)*	*36 (12.6%)*
0.3 - 0.4	*53 (6.9%)*	*16 (12.3%)*	*44 (15.4%)*
0.4 - 0.5	*71 (9.3%)*	*14 (10.8%)*	*25 (8.8%)*
0.5 - 0.6	*89 (11.6%)*	*13 (10.0%)*	*31 (10.9%)*
0.6 - 0.7	*66 (8.6%)*	*16 (12.3%)*	*24 (8.4%)*
0.7 - 0.8	*71 (9.3%)*	*9 (6.9%)*	*18 (6.3%)*
0.8 - 0.9	*67 (8.8%)*	*8 (6.2%)*	*18 (6.3%)*
0.9 - 1.0	*265 (34.7%)*	*17 (13.1%)*	*20 (7.0%)*

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

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*15 (1.9%)*	*10 (9.3%)*	*39 (13.4%)*
0.1 - 0.2	*23 (3.0%)*	*16 (14.8%)*	*43 (14.7%)*
0.2 - 0.3	*57 (7.3%)*	*10 (9.3%)*	*54 (18.5%)*
0.3 - 0.4	*51 (6.5%)*	*11 (10.2%)*	*33 (11.3%)*
0.4 - 0.5	*70 (9.0%)*	*15 (13.9%)*	*21 (7.2%)*
0.5 - 0.6	*63 (8.1%)*	*6 (5.6%)*	*29 (9.9%)*
0.6 - 0.7	*62 (8.0%)*	*9 (8.3%)*	*18 (6.2%)*
0.7 - 0.8	*90 (11.6%)*	*5 (4.6%)*	*18 (6.2%)*
0.8 - 0.9	*106 (13.6%)*	*9 (8.3%)*	*20 (6.8%)*
0.9 - 1.0	*242 (31.1%)*	*17 (15.7%)*	*17 (5.8%)*

For deepset/minilm-uncased-squad2 model:

Range	*Correct*	*Partially Correct*	*Wrong*
0.0 - 0.1	*77 (9.6%)*	*23 (22.1%)*	*87 (31.4%)*

0.1 - 0.2	*46 (5.8%)*	*7 (6.7%)*	*33 (11.9%)*
0.2 - 0.3	*47 (5.9%)*	*14 (13.5%)*	*31 (11.2%)*
0.3 - 0.4	*62 (7.8%)*	*8 (7.7%)*	*30 (10.8%)*
0.4 - 0.5	*67 (8.4%)*	*7 (6.7%)*	*22 (7.9%)*
0.5 - 0.6	*70 (8.8%)*	*12 (11.5%)*	*22 (7.9%)*
0.6 - 0.7	*67 (8.4%)*	*8 (7.7%)*	*11 (4.0%)*
0.7 - 0.8	*63 (7.9%)*	*12 (11.5%)*	*15 (5.4%)*
0.8 - 0.9	*87 (10.9%)*	*5 (4.8%)*	*13 (4.7%)*
0.9 - 1.0	*212 (26.6%)*	*8 (7.7%)*	*13 (4.7%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*11 (1.3%)*	*7 (6.1%)*	*26 (10.4%)*
0.1 - 0.2	*19 (2.3%)*	*8 (7.0%)*	*26 (10.4%)*
0.2 - 0.3	*35 (4.3%)*	*5 (4.4%)*	*34 (13.6%)*
0.3 - 0.4	*45 (5.5%)*	*10 (8.8%)*	*28 (11.2%)*
0.4 - 0.5	*44 (5.4%)*	*17 (14.9%)*	*14 (5.6%)*
0.5 - 0.6	*73 (9.0%)*	*8 (7.0%)*	*35 (14.0%)*
0.6 - 0.7	*75 (9.2%)*	*21 (18.4%)*	*22 (8.8%)*
0.7 - 0.8	*76 (9.3%)*	*9 (7.9%)*	*14 (5.6%)*
0.8 - 0.9	*107 (13.1%)*	*12 (10.5%)*	*18 (7.2%)*
0.9 - 1.0	*330 (40.5%)*	*17 (14.9%)*	*33 (13.2%)*

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

Range	***Correct***	***Partially Correct***	***Wrong***
0.0 - 0.1	*165 (22.1%)*	*45 (37.5%)*	*165 (52.9%)*
0.1 - 0.2	*48 (6.4%)*	*5 (4.2%)*	*28 (9.0%)*
0.2 - 0.3	*41 (5.5%)*	*12 (10.0%)*	*22 (7.1%)*

0.3 - 0.4	40 (5.4%)	13 (10.8%)	16 (5.1%)
0.4 - 0.5	47 (6.3%)	14 (11.7%)	16 (5.1%)
0.5 - 0.6	52 (7.0%)	7 (5.8%)	15 (4.8%)
0.6 - 0.7	59 (7.9%)	8 (6.7%)	15 (4.8%)
0.7 - 0.8	62 (8.3%)	3 (2.5%)	10 (3.2%)
0.8 - 0.9	61 (8.2%)	6 (5.0%)	10 (3.2%)
0.9 - 1.0	172 (23.0%)	7 (5.8%)	15 (4.8%)

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

Range	Correct	Partially Correct	Wrong
0.0 - 0.1	11 (1.3%)	9 (8.7%)	25 (10.6%)
0.1 - 0.2	20 (2.4%)	9 (8.7%)	31 (13.1%)
0.2 - 0.3	35 (4.2%)	8 (7.7%)	27 (11.4%)
0.3 - 0.4	56 (6.7%)	12 (11.5%)	25 (10.6%)
0.4 - 0.5	74 (8.8%)	13 (12.5%)	22 (9.3%)
0.5 - 0.6	75 (8.9%)	12 (11.5%)	24 (10.2%)
0.6 - 0.7	81 (9.7%)	15 (14.4%)	22 (9.3%)
0.7 - 0.8	81 (9.7%)	8 (7.7%)	19 (8.1%)
0.8 - 0.9	88 (10.5%)	3 (2.9%)	13 (5.5%)
0.9 - 1.0	318 (37.9%)	15 (14.4%)	28 (11.9%)

Question Statistics:

563 questions were answered correctly by all models

131 questions were answered incorrectly by all models

964 questions were answered correctly by at least one model

36 questions were answered correctly by exactly one model

688 questions were answered correctly or partially correct by all models