

Models	SQuAD1.1 dev	SQuAD2.0 dev	SQuAD2.0 test	RACE test (Middle/High)
<i>Single model (from leaderboard as of Sept. 23, 2019)</i>				
BERT-large	90.9/84.1	81.8/79.0	89.1/86.3	72.0 (76.6/70.1)
XLNet	94.5/89.0	88.8/86.1	89.1/86.3	81.8 (85.5/80.2)
RoBERTa	94.6/88.9	89.4/86.5	89.8/86.8	83.2 (86.5/81.3)
UPM	-	-	89.9/87.2	-
XLNet + SG-Net Verifier++	-	-	90.1/87.2	-
ALBERT (1M)	94.8/89.2	89.9/87.2	-	86.0 (88.2/85.1)
ALBERT (1.5M)	<b>94.8/89.3</b>	<b>90.2/87.4</b>	<b>90.9/88.1</b>	<b>86.5 (89.0/85.5)</b>
<i>Ensembles (from leaderboard as of Sept. 23, 2019)</i>				
BERT-large	92.2/86.2	-	-	-
XLNet + SG-Net Verifier	-	-	90.7/88.2	-
UPM	-	-	90.7/88.2	-
XLNet + DAAF + Verifier	-	-	90.9/88.6	-
DCMN+	-	-	-	84.1 (88.5/82.3)
ALBERT	<b>95.5/90.1</b>	<b>91.4/88.9</b>	<b>92.2/89.7</b>	<b>89.4 (91.2/88.6)</b>

Table 10: State-of-the-art results on the SQuAD and RACE benchmarks.