Model type	**Original**	**Unpatched**	**Patched**			
			$\mathcal{T}_{\text{CheckList}}$	$\mathcal{T}_{\text{TestAug}}$	$\mathcal{T}_{\text{TestAug}}$ $\setminus\mathcal{T}_{\text{GPT}-3}$	$\mathcal{T}_{\text{TestAug}}$ $\setminus\mathcal{T}_{\text{Expansion}}$
Sentiment Classification						
ALBERT	7.3	32.6±5.7	13.4±6.5	11.3±10.0	10.6±6.6	**9.6±8.2**
BERT$_{\text{Base}}$	7.6	33.9±6.1	9.0±4.2	**8.3±4.2**	8.5±1.6	9.9±4.9
DistillBERT	10.0	29.5±10.9	6.5±3.4	**3.9±2.1**	4.9±2.1	5.1±3.3
RoBERTa$_{\text{Base}}$	5.7	14.2±6.1	3.7±2.3	1.6±1.0	2.7±2.7	**1.4±1.2**
Paraphrase Detection						
ALBERTA	9.3	38.1±3.8	7.1±0.8	0.6±0.4	5.8±1.8	**0.4±0.4**
BERT$_{\text{Base}}$	9.1	36.0±4.9	6.2±1.5	0.5±0.4	5.6±1.1	**0.4±0.3**
DistillBERT	10.3	49.8±10.2	12.5±16.4	**1.1±2.4**	6.4±3.9	7.3±15.8
Natural Language Inference						
ALBERT	9.9	42.8±1.9	30.1±4.2	**23.0±1.6**	/	/
DistillBERT	12.6	34.7±3.6	23.6±6.1	**16.5±3.9**	/	/
RoBERTa$_{\text{Large}}$	8.1	17.8±4.0	8.3±3.1	**8.0±3.1**	/	/