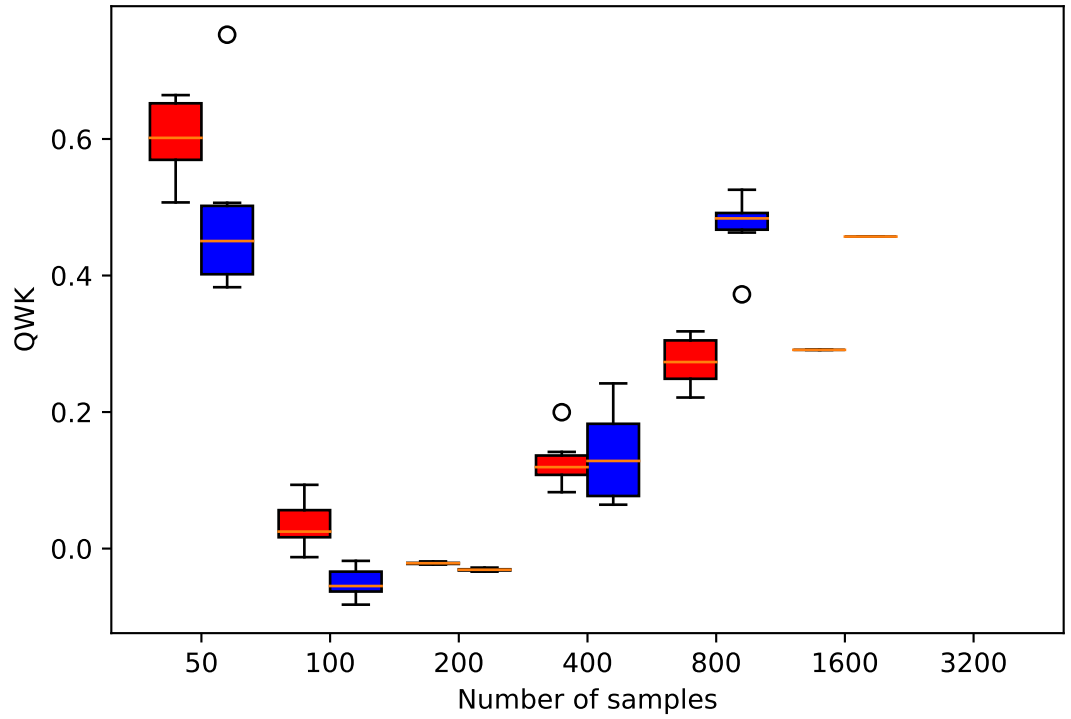


The QWK for score\_1 when comparing two BERT models versus the QWK for score\_1 between a different pair of BERT models.  
The first pair was trained with samples created by human experts and gpt4,  
the second pair was trained with samples created by human experts and text-davinci-003 annotated by gpt-3.5-turbo



Each box represents exactly 6 kappa values  
experts\_vs\_gpt4 is represented by red  
experts\_vs\_turbo is represented by blue