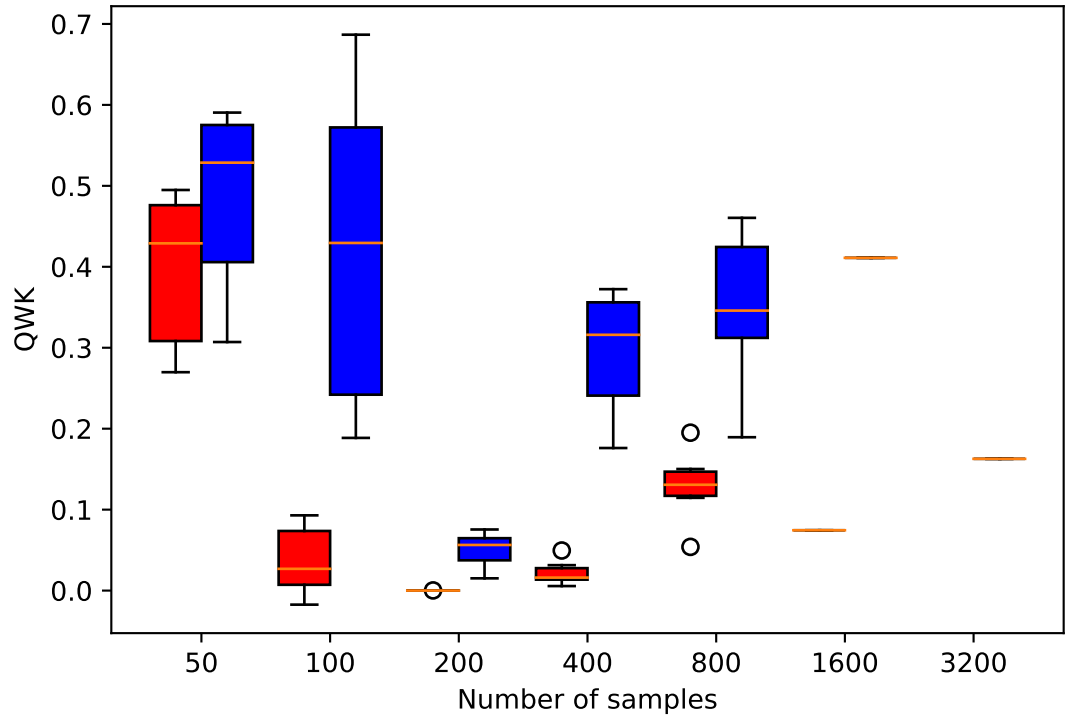


The QWK for score\_1 when comparing two BERT models versus the QWK for score\_1 between a different pair of BERT models.  
The first pair was trained with samples created by gpt4 and human experts,  
the second pair was trained with samples created by text-davinci-003 annotated by gpt-3.5-turbo and gpt4



Each box represents exactly 6 kappa values  
gpt4\_vs\_experts is represented by red  
turbo\_vs\_gpt4 is represented by blue