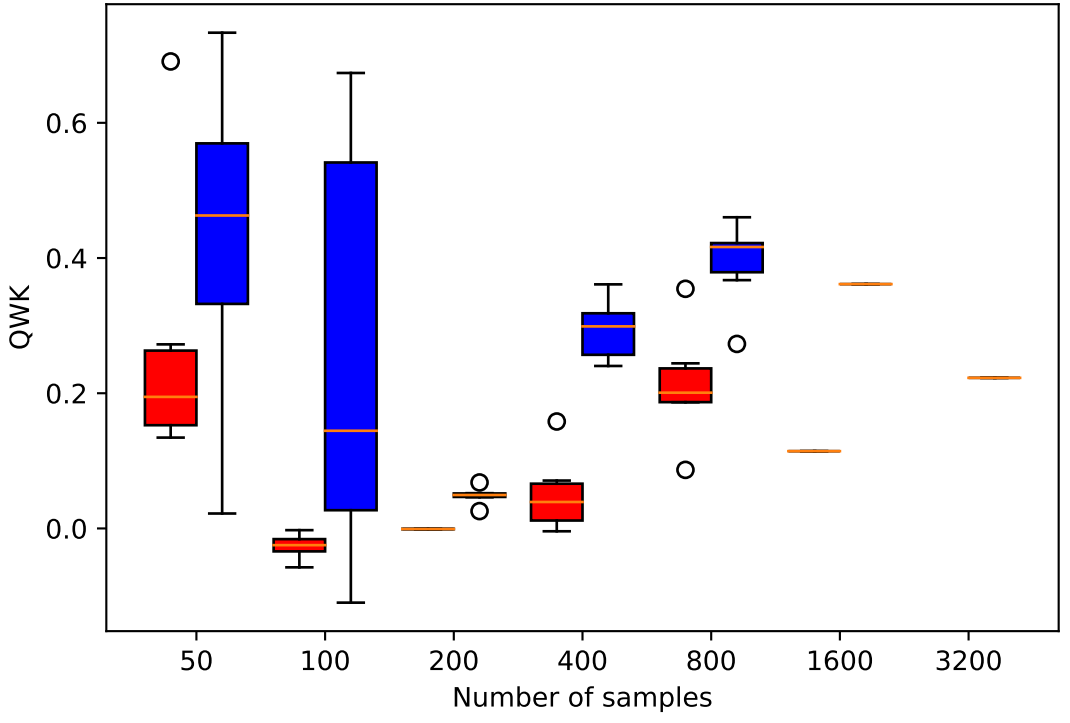
The QWK for score_1 when comparing two BERT models versus the QWK for score_1 between a different pair of BERT models. The first pair was trained with samples created by human experts and text-davinci-003 annotated by gpt-3.5-turbo, the second pair was trained with samples created by text-davinci-003 and gpt4



Each box represents exactly 6 kappa values
experts_vs_turbo is represented by red
davinci_vs_gpt4 is represented by blue