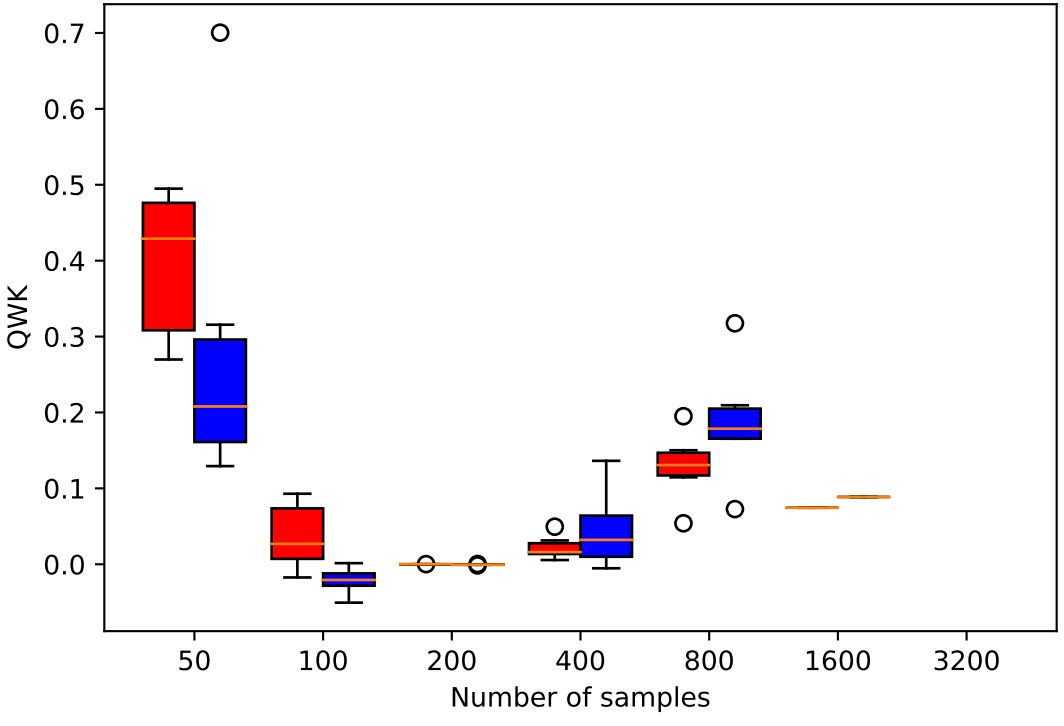
The QWK for score_1 when comparing two BERT models versus the QWK for score_1 between a different pair of BERT models.
The first pair was trained with samples created by gpt4 and human experts,
the second pair was trained with samples created by human experts and text-davinci-003 annotated by gpt-3.5-turbo



Each box represents exactly 6 kappa values
gpt4_vs_experts is represented by red
experts_vs_turbo is represented by blue