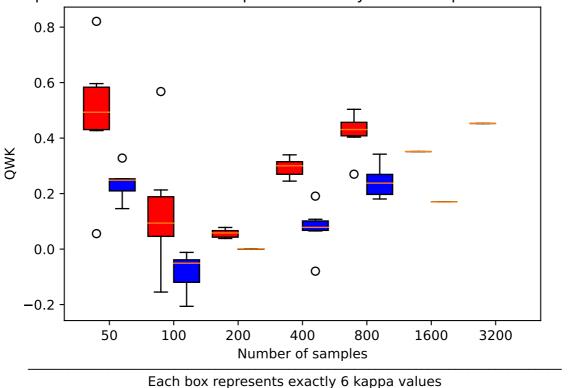The QWK for score_1 when comparing two BERT models versus the QWK for score_1 between a different pair of BERT models. The first pair was trained with samples created by text-davinci-003 and gpt4, the second pair was trained with samples created by human experts and text-davinci-003

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