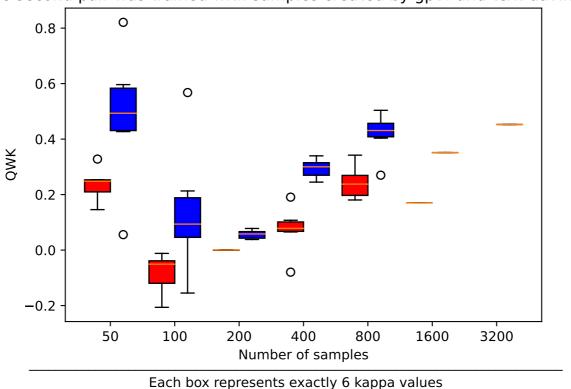
The QWK for score_1 when comparing two BERT models versus the QWK for score_1 between a different pair of BERT models.

The first pair was trained with samples created by text-davinci-003 and human experts,

the second pair was trained with samples created by gpt4 and text-davinci-003



ach box represents exactly 6 kappa values davinci_vs_experts is represented by red gpt4_vs_davinci is represented by blue