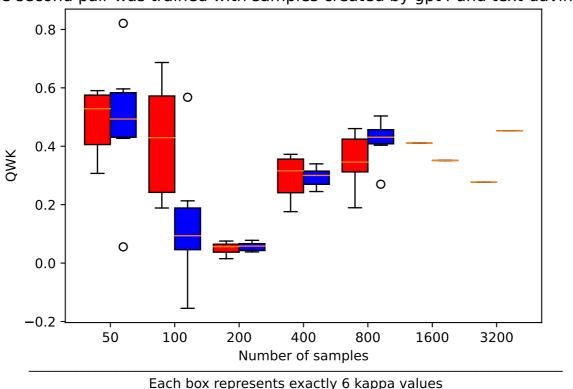
The QWK for score_1 when comparing two BERT models versus the QWK for score_1 between a different pair of BERT models.

The first pair was trained with samples created by text-davinci-003 annotated by gpt-3.5-turbo and gpt4,

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



tch box represents exactly 6 kappa values turbo_vs_gpt4 is represented by red gpt4_vs_davinci is represented by blue