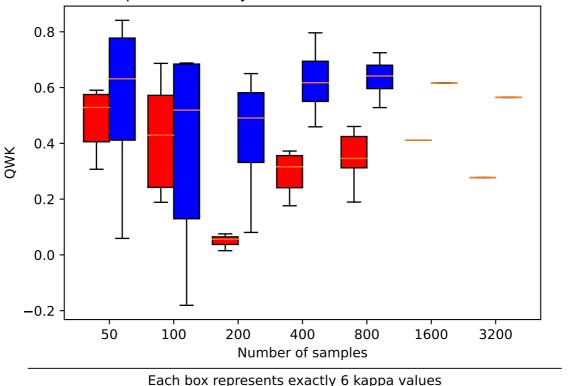
The QWK for score_1 when comparing two BERT models versus the QWK for score_1 between a different pair of BERT models.

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



ach box represents exactly 6 kappa values gpt4_vs_turbo is represented by red davinci_vs_turbo is represented by blue