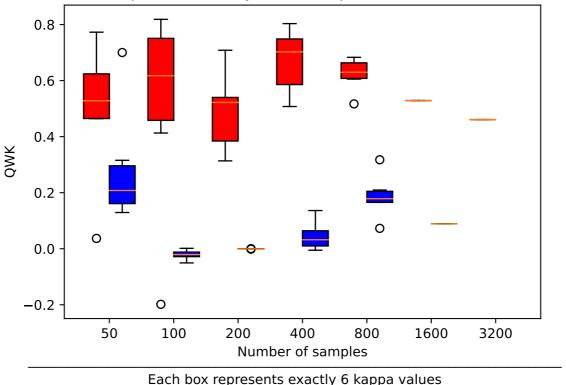
The QWK for score_1 when comparing two BERT models versus the QWK for score_1 between a different pair of BERT models.

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



ach box represents exactly 6 kappa values turbo_vs_davinci is represented by red experts vs turbo is represented by blue