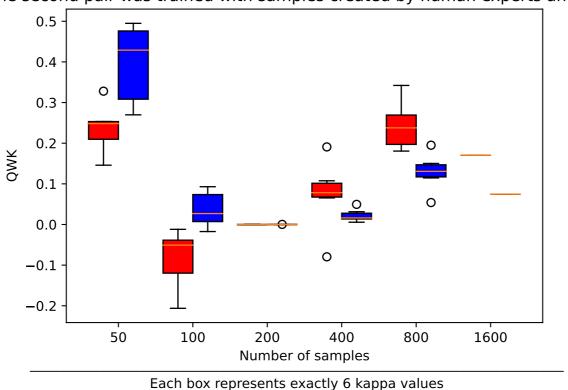
The QWK for score_1 when comparing two BERT models versus the QWK for score_1 between a different pair of BERT models.

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

the second pair was trained with samples created by human experts and gpt4



experts_vs_davinci is represented by red experts vs_gpt4 is represented by blue