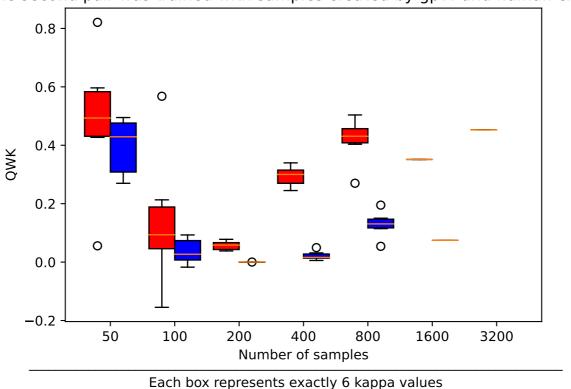
The QWK for score_1 when comparing two BERT models versus the QWK for score_1 between a different pair of BERT models.

The first pair was trained with samples created by text-davinci-003 and gpt4,
the second pair was trained with samples created by gpt4 and human experts



denote the desired days and the desired days are desired by red gpt4_vs_experts is represented by blue