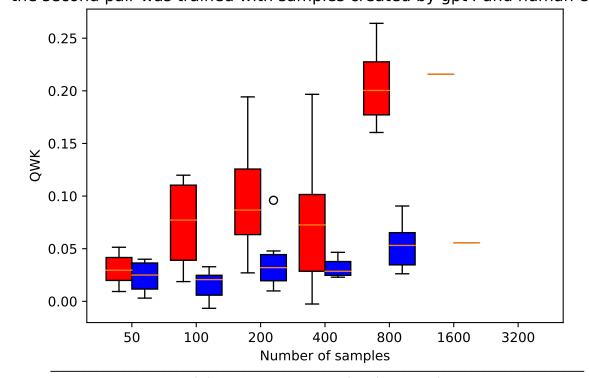
The QWK for score\_1 when comparing two XGB models versus the QWK for score\_1 between a different pair of XGB models.

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

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



Each box represents exactly 6 kappa values experts\_vs\_davinci is represented by red gpt4\_vs\_experts is represented by blue