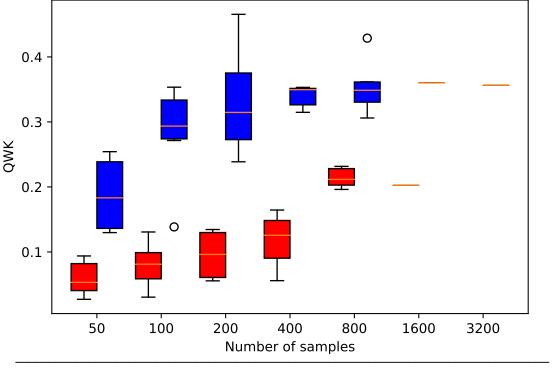
The QWK for score\_1 when comparing two XGB models versus the QWK for score\_1 between a different pair of XGB models.

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



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