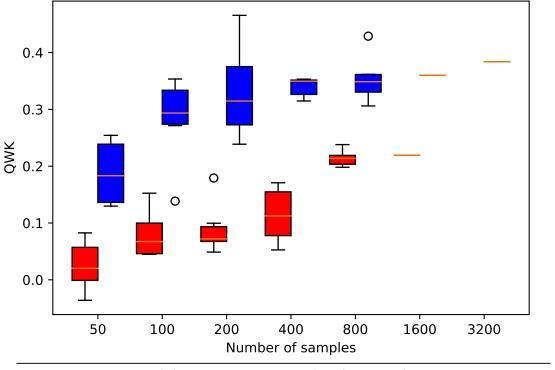
The QWK for score\_1 when comparing two XGB models versus the QWK for score\_1 between a different pair of XGB models.

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



Each box represents exactly 6 kappa values davinci\_vs\_experts is represented by red turbo\_vs\_gpt4 is represented by blue