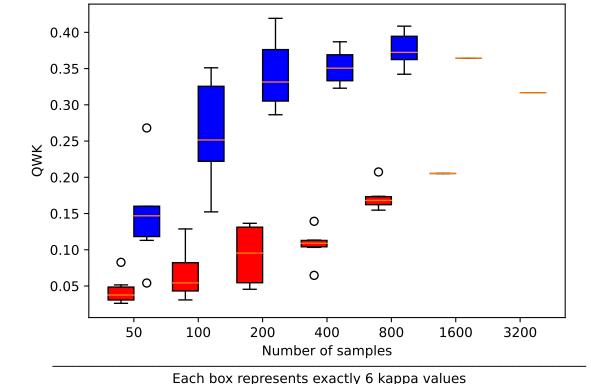
The QWK for score\_1 when comparing two XGB models versus the QWK for score\_1 between a different pair of XGB models.

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



ech box represents exactly 6 kappa values experts\_vs\_turbo is represented by red davinci\_vs\_gpt4 is represented by blue