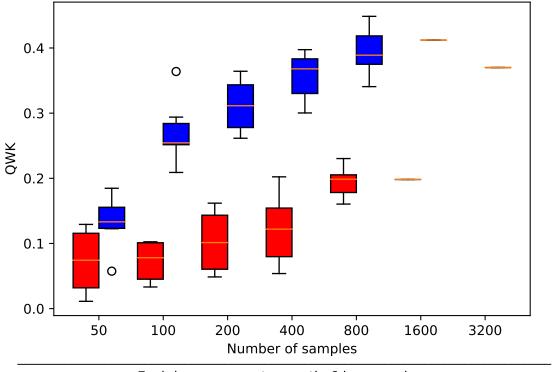
The QWK for score_1 when comparing two XGB models versus the QWK for score_1 between a different pair of XGB models.

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



Each box represents exactly 6 kappa values experts_vs_gpt4 is represented by red davinci_vs_turbo is represented by blue