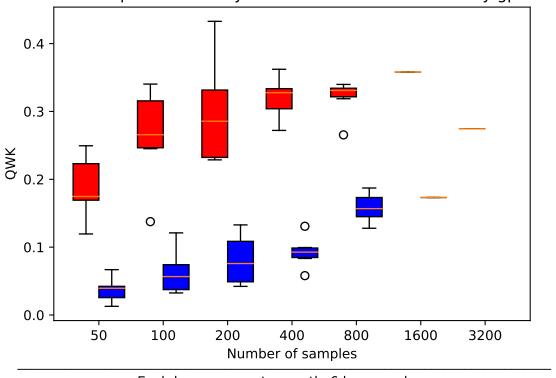
The QWK for score_1 when comparing two XGB models versus the QWK for score_1 between a different pair of XGB models.

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



Each box represents exactly 6 kappa values gpt4_vs_turbo is represented by red turbo_vs_experts is represented by blue