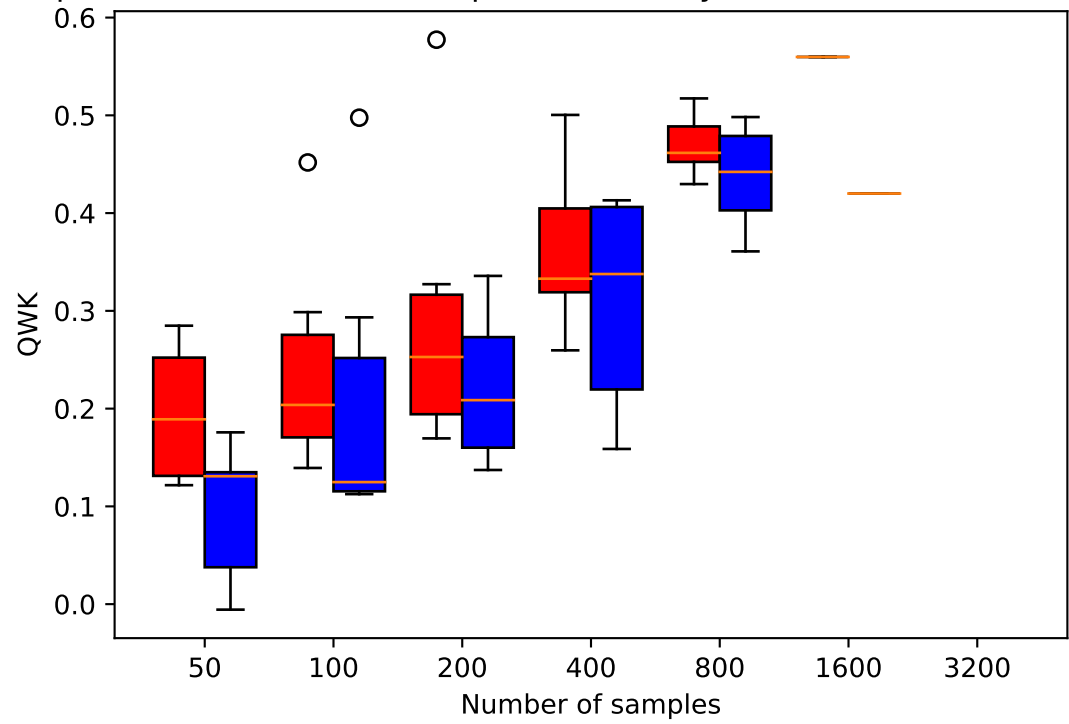The QWK for score_1 when comparing two XGB models versus the QWK for score_1 between a different pair of XGB models. The first pair was trained with samples created by human experts and gpt4, the second pair was trained with samples created by text-davinci-003 and human experts

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