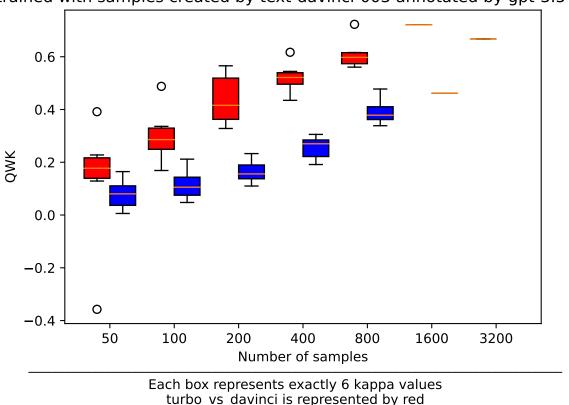
The QWK for score 1 when comparing two XGB models versus the QWK for score 1 between a different pair of XGB models. The first pair was trained with samples created by text-davinci-003 annotated by gpt-3.5-turbo and text-davinci-003, the second pair was trained with samples created by text-davinci-003 annotated by gpt-3.5-turbo and human experts



turbo vs experts is represented by blue