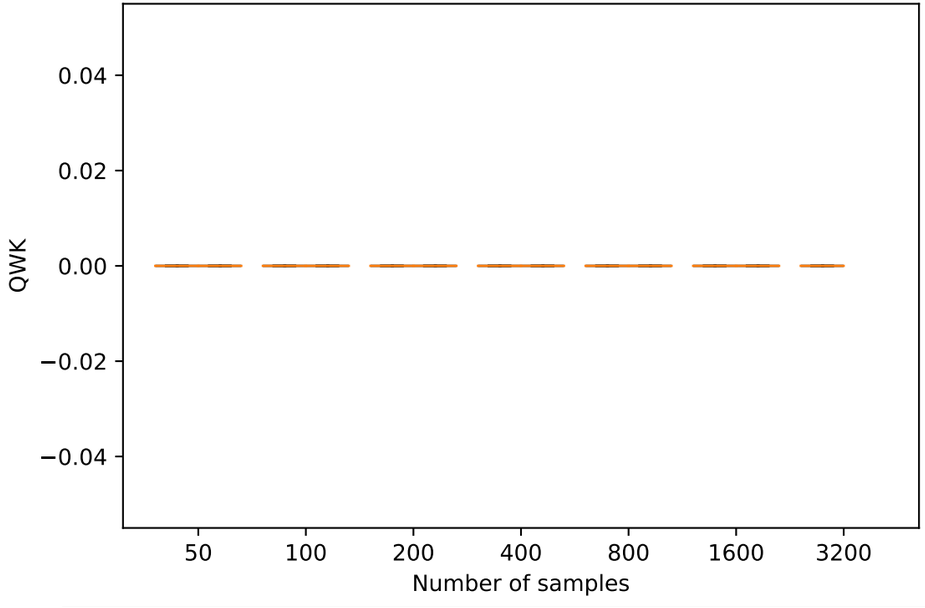
QWK of two BERT Models with the source of the data. One trained with samples created by text-davinci-003 annotated by gpt-3.5-turbo One trained with samples created by human experts.
Both rated answers created by gpt4 used for testing



Each box represents exactly 6 kappa values
text-davinci-003 annotated by gpt-3.5-turbo is represented by red
human experts is represented by blue