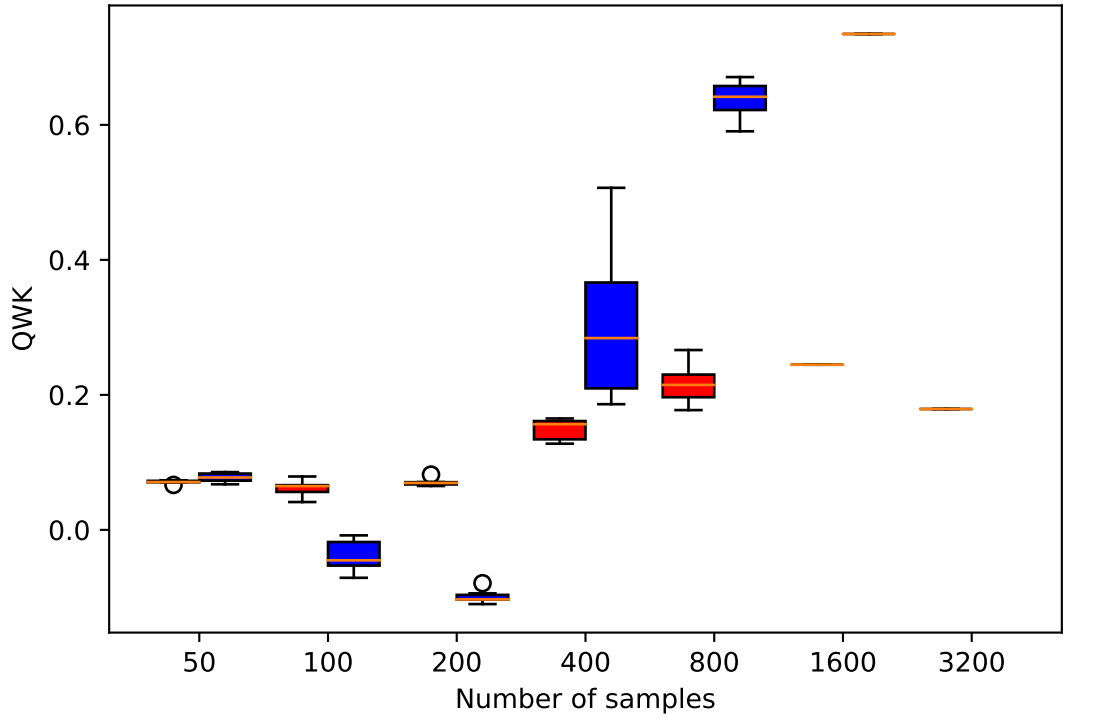
QWK of two BERT Models with the source of the data. One trained with samples created by gpt4 One trained with samples created by human experts.
Both rated answers created by human experts used for testing



Each box represents exactly 6 kappa values
gpt4 is represented by red
human experts is represented by blue