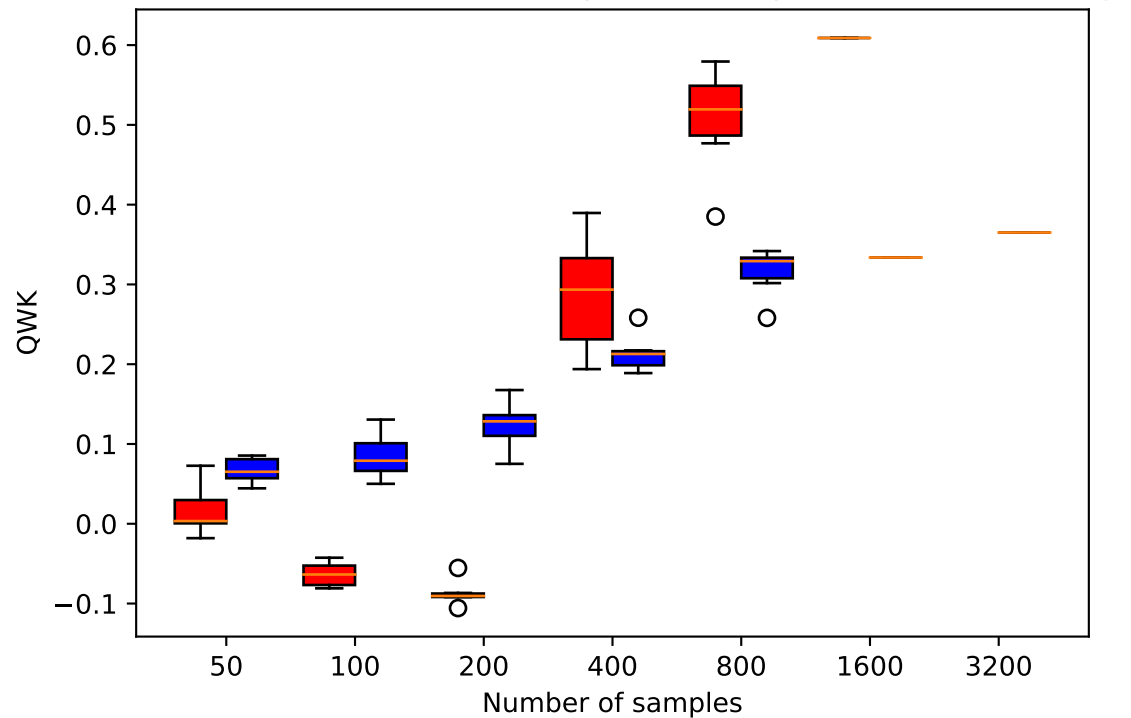QWK of two BERT Models with the source of the data. One trained with samples created by human experts One trained with samples created by gpt4. Both rated answers created by human experts used for testing

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