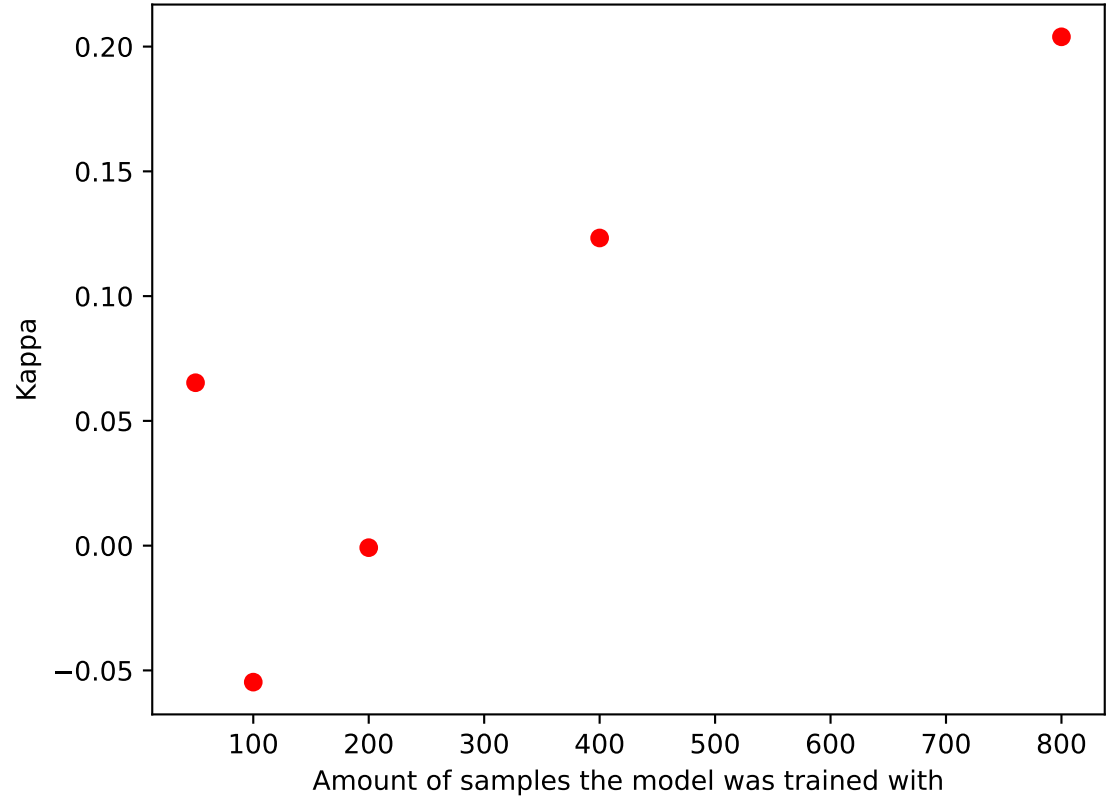XGB Model trained with samples created by human experts rating answers created by gpt4 used for testing

Kappa

Amount of samples the model was trained with

Each value represents the agreement of exactly 1600 ratings