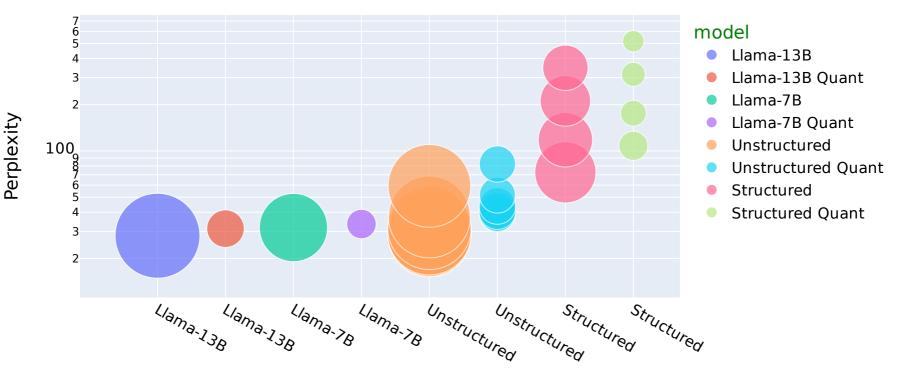Llama Model Performance and Inference time ptb