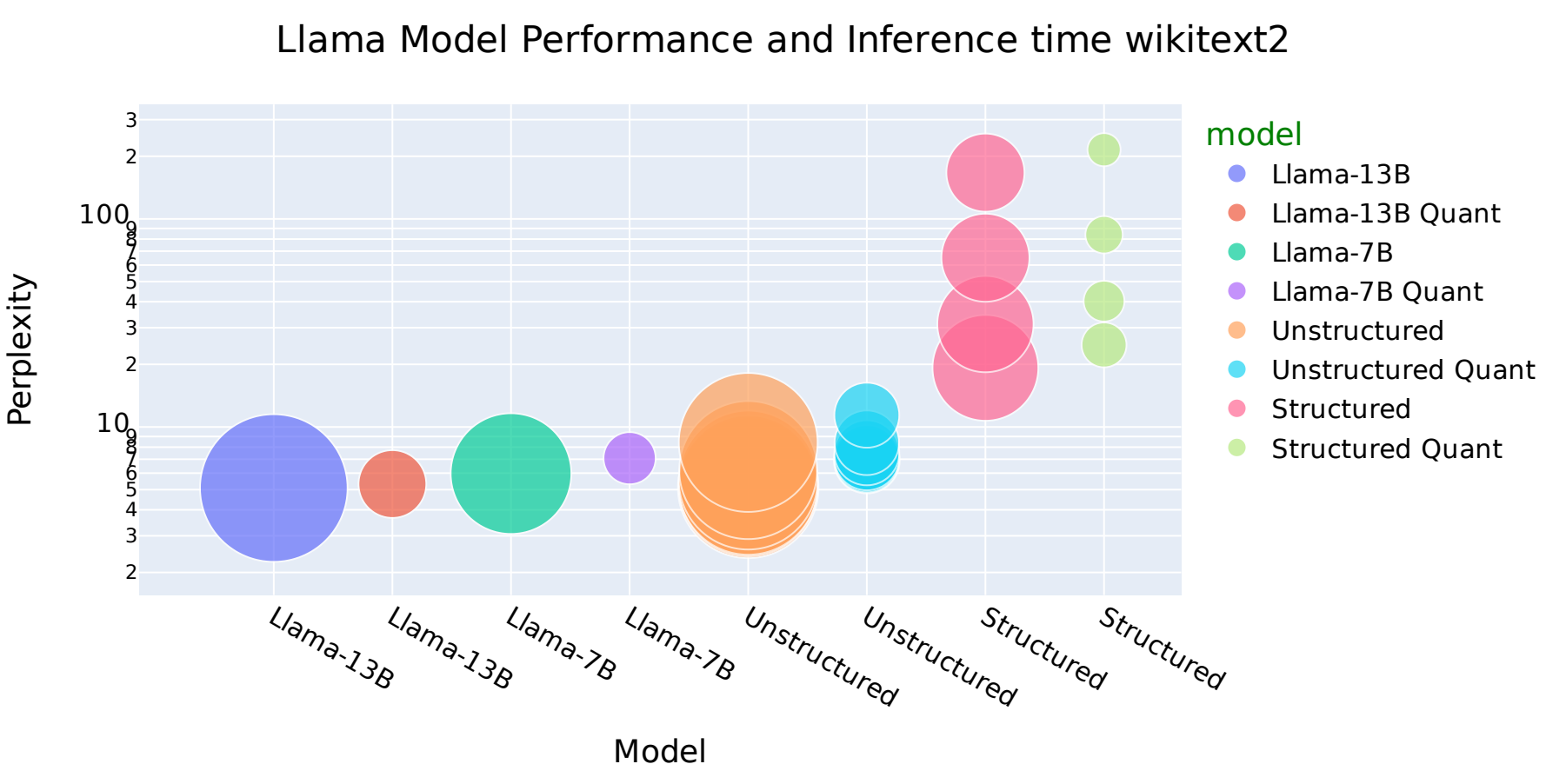
Llama Model Performance and Inference time wikitext2
Perplexity
100
10
Model
Llama-13B
Llama-13B
Llama-7B
Llama-7B
Unstructured
Unstructured
Structured
Structured
model
Llama-13B
Llama-13B Quant
Llama-7B
Llama-7B Quant
Unstructured
Unstructured Quant
Structured
Structured Quant