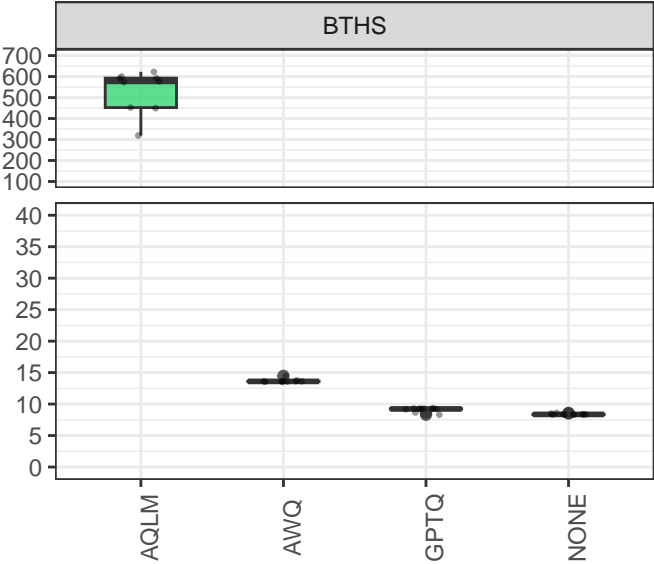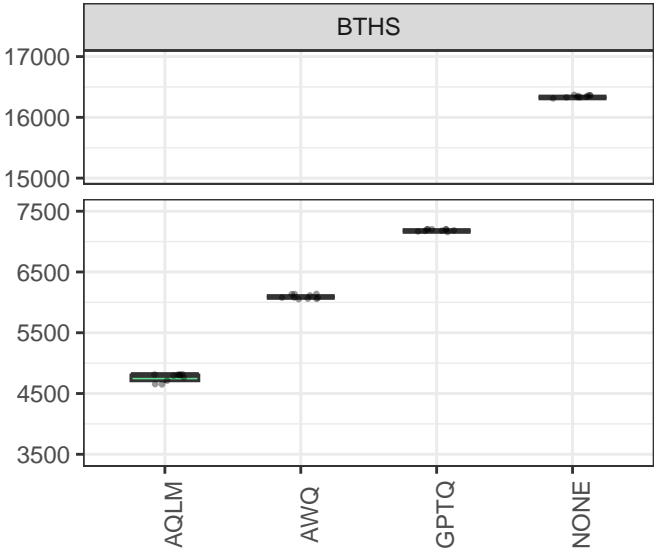# Inference Time in Seconds

VRAM Max Usage MiB

Quantization:

AQLM   AWQ   GPTQ   NONE