

Hardware: a100 Inference: autohf

Throughput (Tokens per seconds)

