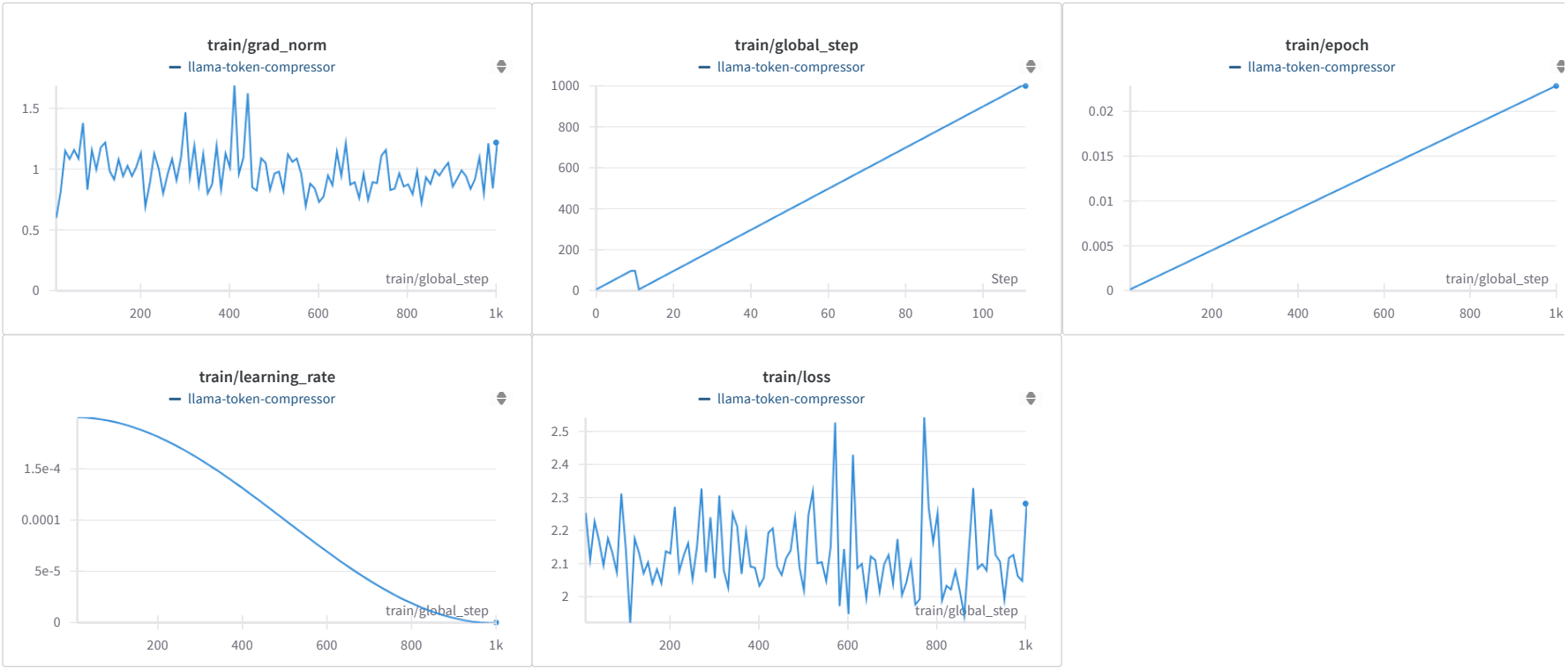
Token Compressor

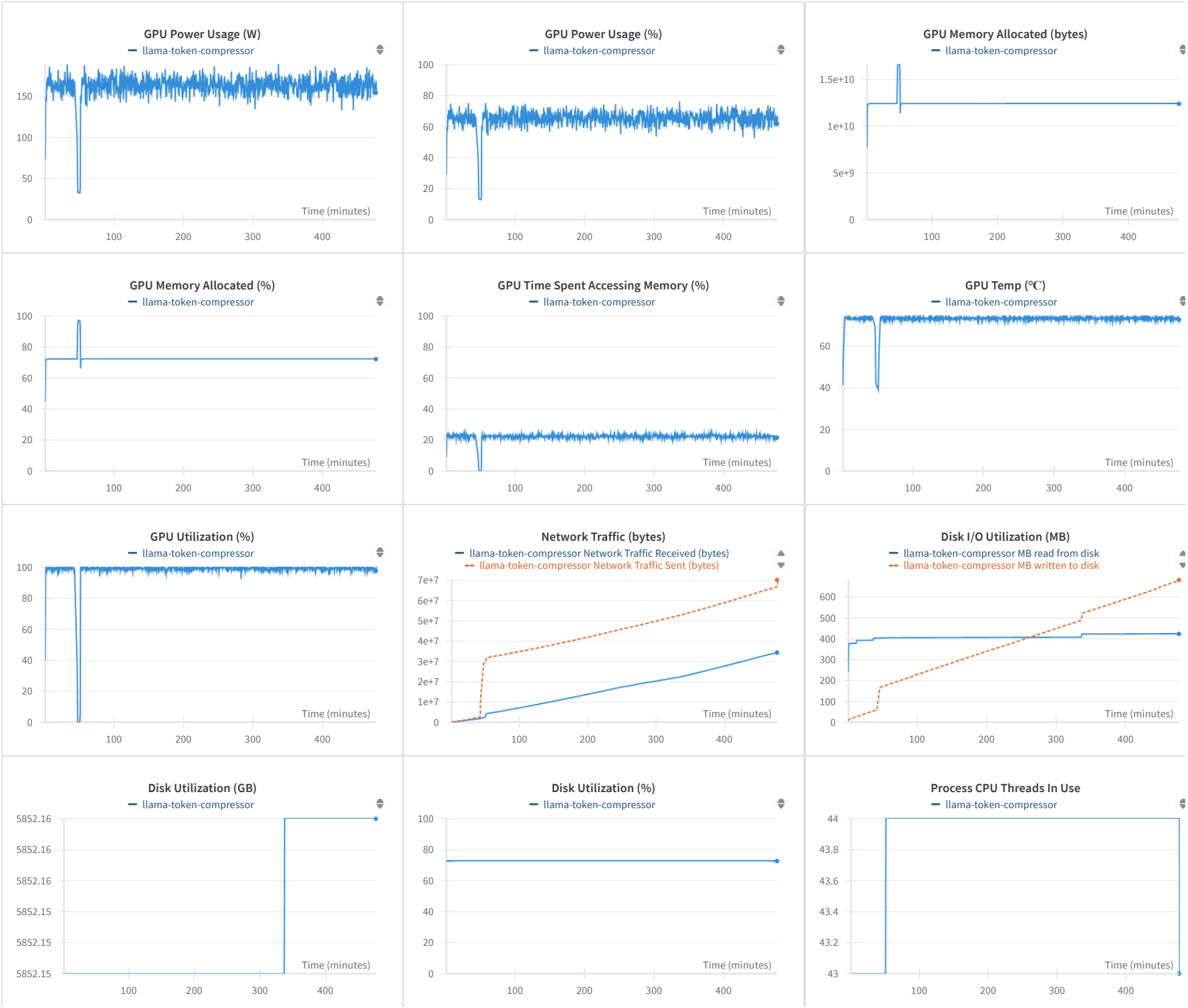
Based on QLoRA finetuning of Llama 3.1 8B in 4-bit

Alosh Denny

Section 1



Import panel Add panel


GPU Power Usage (W)
llama-token-compressor
150
100
50
0
Time (minutes)
100
200
300
400
GPU Power Usage (%)
llama-token-compressor
100
80
60
40
20
0
Time (minutes)
100
200
300
400
GPU Memory Allocated (bytes)
llama-token-compressor
1.5e+10
1e+10
5e+9
0
Time (minutes)
100
200
300
400
GPU Memory Allocated (%)
llama-token-compressor
100
80
60
40
20
0
Time (minutes)
100
200
300
400
GPU Time Spent Accessing Memory (%)
llama-token-compressor
100
80
60
40
20
0
Time (minutes)
100
200
300
400
GPU Temp (℃)
llama-token-compressor
60
40
20
0
Time (minutes)
100
200
300
400
GPU Utilization (%)
llama-token-compressor
100
80
60
40
20
0
Time (minutes)
100
200
300
400
Network Traffic (bytes)
llama-token-compressor Network Traffic Received (bytes)
llama-token-compressor Network Traffic Sent (bytes)
7e+7
6e+7
5e+7
4e+7
3e+7
2e+7
1e+7
0
Time (minutes)
100
200
300
400
Disk I/O Utilization (MB)
llama-token-compressor MB read from disk
llama-token-compressor MB written to disk
600
500
400
300
200
100
0
Time (minutes)
100
200
300
400
Disk Utilization (GB)
llama-token-compressor
5852.16
5852.16
5852.16
5852.15
5852.15
5852.15
Time (minutes)
100
200
300
400
Disk Utilization (%)
llama-token-compressor
100
80
60
40
20
0
Time (minutes)
100
200
300
400
Process CPU Threads In Use
llama-token-compressor
44
43.8
43.6
43.4
43.2
43
Time (minutes)
100
200
300
400

Import panel Add panel