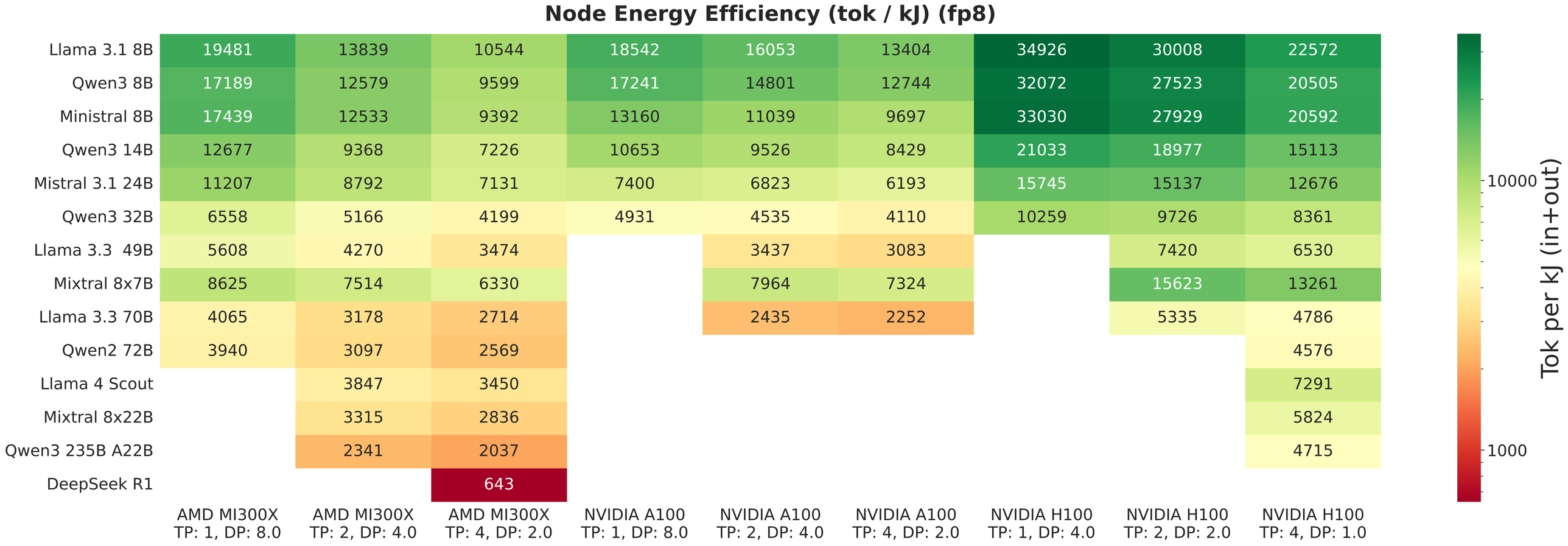
Node Energy Efficiency (tok / kJ) (fp8)
Llama 3.1 8B 19481 13839 10544 18542 16053 13404 34926 30008 22572
Qwen3 8B 17189 12579 9599 17241 14801 12744 32072 27523 20505
Ministral 8B 17439 12533 9392 13160 11039 9697 33030 27929 20592
Qwen3 14B 12677 9368 7226 10653 9526 8429 21033 18977 15113
Mistral 3.1 24B 11207 8792 7131 7400 6823 6193 15745 15137 12676
Qwen3 32B 6558 5166 4199 4931 4535 4110 10259 9726 8361
Llama 3.3 49B 5608 4270 3474 3437 3083 7420 6530
Mixtral 8x7B 8625 7514 6330 7964 7324 15623 13261
Llama 3.3 70B 4065 3178 2714 2435 2252 5335 4786
Qwen2 72B 3940 3097 2569 4576
Llama 4 Scout 3847 3450 7291
Mixtral 8x22B 3315 2836 5824
Qwen3 235B A22B 2341 2037 4715
DeepSeek R1 643
AMD MI300X TP: 1, DP: 8.0
AMD MI300X TP: 2, DP: 4.0
AMD MI300X TP: 4, DP: 2.0
NVIDIA A100 TP: 1, DP: 8.0
NVIDIA A100 TP: 2, DP: 4.0
NVIDIA A100 TP: 4, DP: 2.0
NVIDIA H100 TP: 1, DP: 4.0
NVIDIA H100 TP: 2, DP: 2.0
NVIDIA H100 TP: 4, DP: 1.0
Tok per kJ (in+out)
10000
1000