Node Energy Efficiency (tok / kJ) (bfloat16)

	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
Llama 3.1 8B	15449	10594	7426	19043	16287	13689	27244	24565	17163
Qwen3 8B	14973	11119	8692	18101	14983	12094	25411	22529	15433
Ministral 8B	14238	9739	6870	13623	11335	9835	26073	22897	16130
Qwen3 14B	10176	7880	6483	11115	10185	8445	16121	15124	12419
Mistral 3.1 24B	7928	5977	4211	8423	7943	6982	12638	11239	9420
Qwen3 32B	5031	3826	2721	4088	5059	4516	6089	7561	6669
Llama 3.3 49B	3987	3064	2333		3805	3484		5574	4646
Mixtral 8x7B	5965	5533	3892		7141	6562		9370	7609
Llama 3.3 70B	2848	2280	1737	1295		2479		2160	3381
Qwen2 72B	2739	2217	1658			2425			3264
Mixtral 8x22B			1773						

Tok per kJ (in+out)