Computational Capability Comparison for RAG Encoding

Processor	Peak Performance (TOPS)	Power (W)	Power Efficiency (GOPS/W)	Cost ($)	Cost Efficiency (GOPS/$)	Encoding Throughput (Q/s)	Encoding Latency (ms)	Energy per Query (J)	Memory Bound
A100 GPU	19.5	300	65.0	11,000	1.77	1	70980.7	21294.208	No
V100 GPU	14.0	250	56.0	8,000	1.75	1	98866.0	24716.492	No
RTX 4090	83.0	450	184.4	1,600	51.88	4	16676.2	7504.284	No
ARM Cortex-A78	0.4	5	84.0	200	2.10	0	3295532.2	16477.661	No
ARM Cortex-A76	0.2	3	66.7	150	1.33	0	6920617.7	20761.853	No
Apple M2	0.8	15	53.3	400	2.00	0	1730154.4	25952.316	No
Intel Xeon	1.2	270	4.4	8,000	0.15	0	1153436.3	311427.797	No
AMD EPYC	2.0	280	7.1	7,000	0.29	0	692061.8	193777.296	No