Beyond Test-Time Compute Strategies: Advocating Energy-per-Token in LLM Inference

Is it worth to use reasoning strategies on Small Language Models ?

High energy cost for accuracy improvements in certain categories!

If possible (enough memory) bigger sized models offer a better energy - accuracy trade-off!

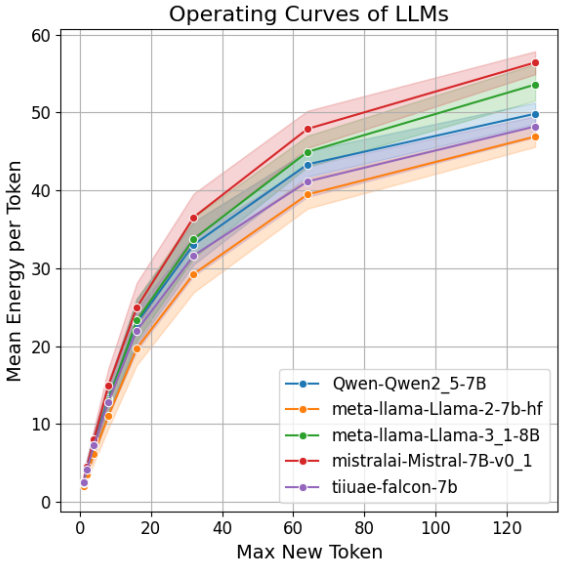
Category	Zero-Shot				Reasoning					
MMLU	Llama 1B		Llama 8B		Llama 1B MV		Llama 1B CoT		DeepSeek 1.5B	
	Acc.	Energy	Acc.	Δ%Acc, Δ%E	Acc.	Δ%Acc, Δ%E	Acc.	Δ%Acc, Δ%E	Acc.	Δ%Acc, Δ%E
Computer Science	0.38	78,556 kJ	0.56	(+47%, +42%)	0.39	(+3%, +118%)	0.43	(+13%, +13,858%)	0.42	(+11%, +39%)
Economics	0.40	80,437 kJ	0.62	(+51%, +65%)	0.42	(+5%, +177%)	0.42	(+5%, +13,211%)	0.41	(+3%, +40%)
Engineering	0.37	74,805 kJ	0.75	(+99%, +36%)	0.44	(+19%, +72%)	0.43	(+17%, +12,233%)	0.55	(+50%, +39%)
Health	0.50	78,484 kJ	0.78	(+57%, +44%)	0.54	(+8%, +108%)	0.55	(+10%, +14,339%)	0.43	(-14%, +42%)
Humanities	0.44	79,029 kJ	0.72	(+61%, +63%)	0.46	(+5%, +174%)	0.45	(+2%, +13,334%)	0.30	(-32%, +40%)
Math	0.11	83,532 kJ	0.39	(+350%, +37%)	0.11	(+0%, +88%)	0.31	(+281%, +15,132%)	0.19	(+69%, +33%)
Natural Sciences	0.26	76,172 kJ	0.54	(+100%, +41%)	0.27	(+4%, +102%)	0.29	(+11%, +15,483%)	0.38	(+45%, +40%)
Sociology	0.47	76,673 kJ	0.82	(+72%, +40%)	0.48	(+2%, +97%)	0.60	(+26%, +13,198%)	0.49	(+4%, +41%)

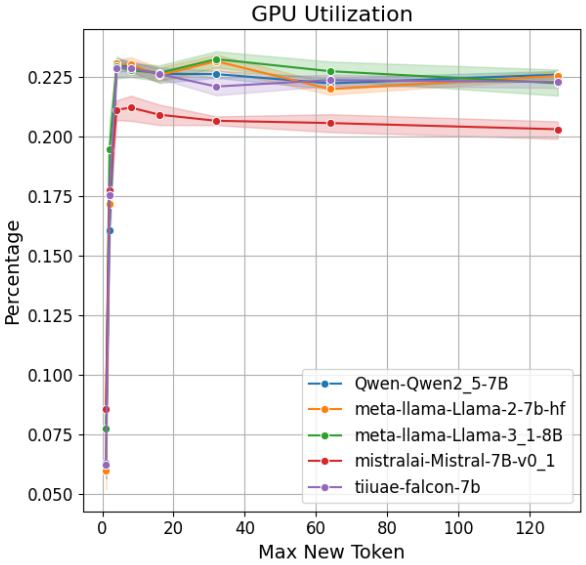
BASELINE: Llama 1B

Beyond Test-Time Compute Strategies: Advocating Energy-per-Token in LLM Inference

$$\text{Energy per Token [Joule]} = \frac{W_{\text{consumed}} * Time(s)}{T_{\text{processed}}}$$

Same Size LLM differ in their energy efficiency!





Takeaway:
Query-Routing for more efficient AI deployment between reasoning based and non-reasoning based LLMs!