Same Size, Different Costs

Phase-Level Energy Variations in Transformer Models

Lola Solovyeva

UNIVERSITY OF TWENTE.

TOYOTA GR86
$30,800

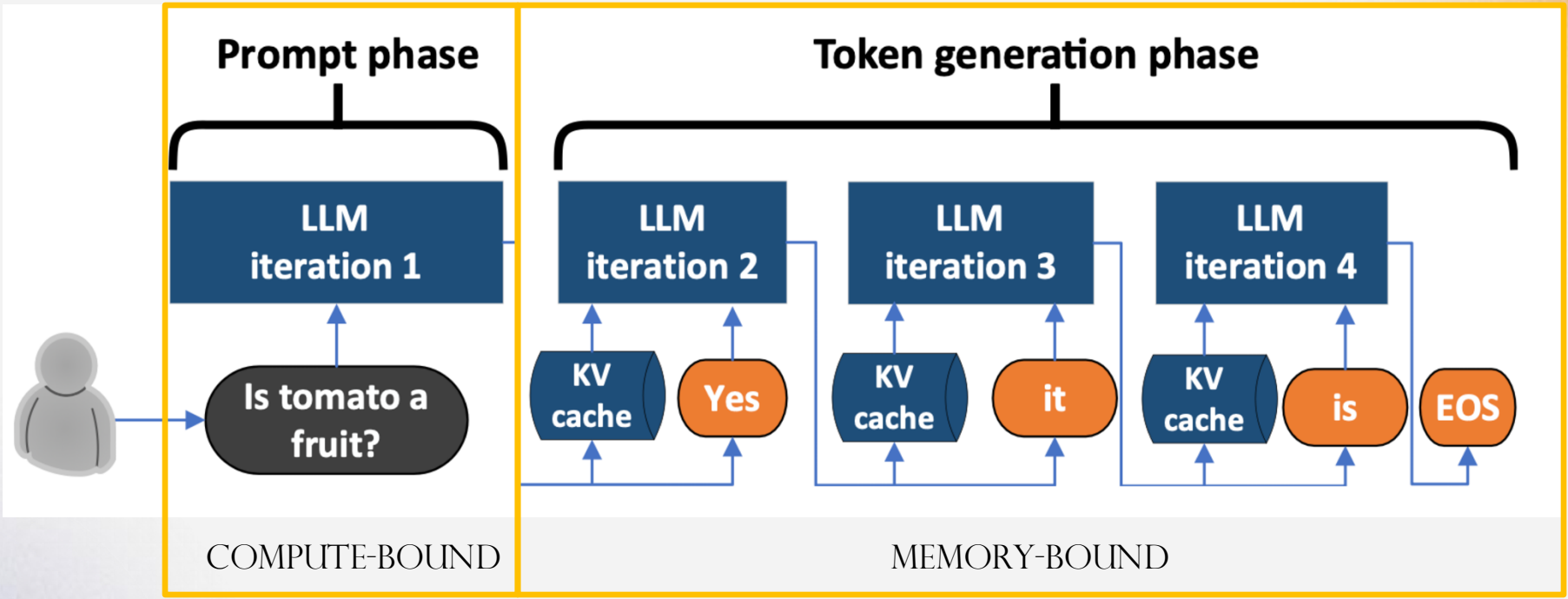
Length: 4,265mm
Width: 1,775mm
Height: 1,310mm

Engine: 2.4L naturally aspirated flat-4

PORSCHE 718 CAYMAN
$74,900

Length: 4,379mm
Width: 1,801mm
Height: 1,295mm

Engine: 2.0L turbo flat-4



[3]: P. Patel et al., "Splitwise: Efficient Generative LLM Inference Using Phase Splitting," 2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA), Buenos Aires, Argentina, 2024, pp. 118-132, doi: 10.1109/ISCA59077.2024.00019.

How much do the two phases differ in their contribution to the total energy consumption of LLM inference?

EFFECT OF INPUT AND OUTPUT

Code generation with HumanEval:

- 0-shot
- 2-shot
- 0-shot Chain-of-thought

Q&A with LongBench

- Multiple choice
- Open questions

EFFECT OF MODEL CHOICE

Model name	Parameters
CodeLlama	7B
Qwen2.5-Coder	7B
DeepseekCoder	6.7B
CodeGemma	7B
CodeQwen1.5	7B
NextCoder	7B
Phi3.5	4B
Phi4	4B
Qwen3	4B
Qwen2.5-Coder	3B

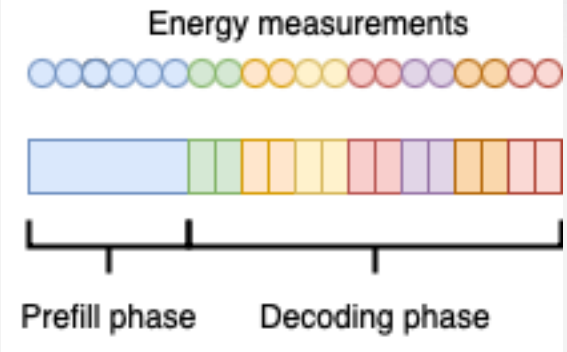
Energy profiling [4]

ENERGY MEASUREMENT

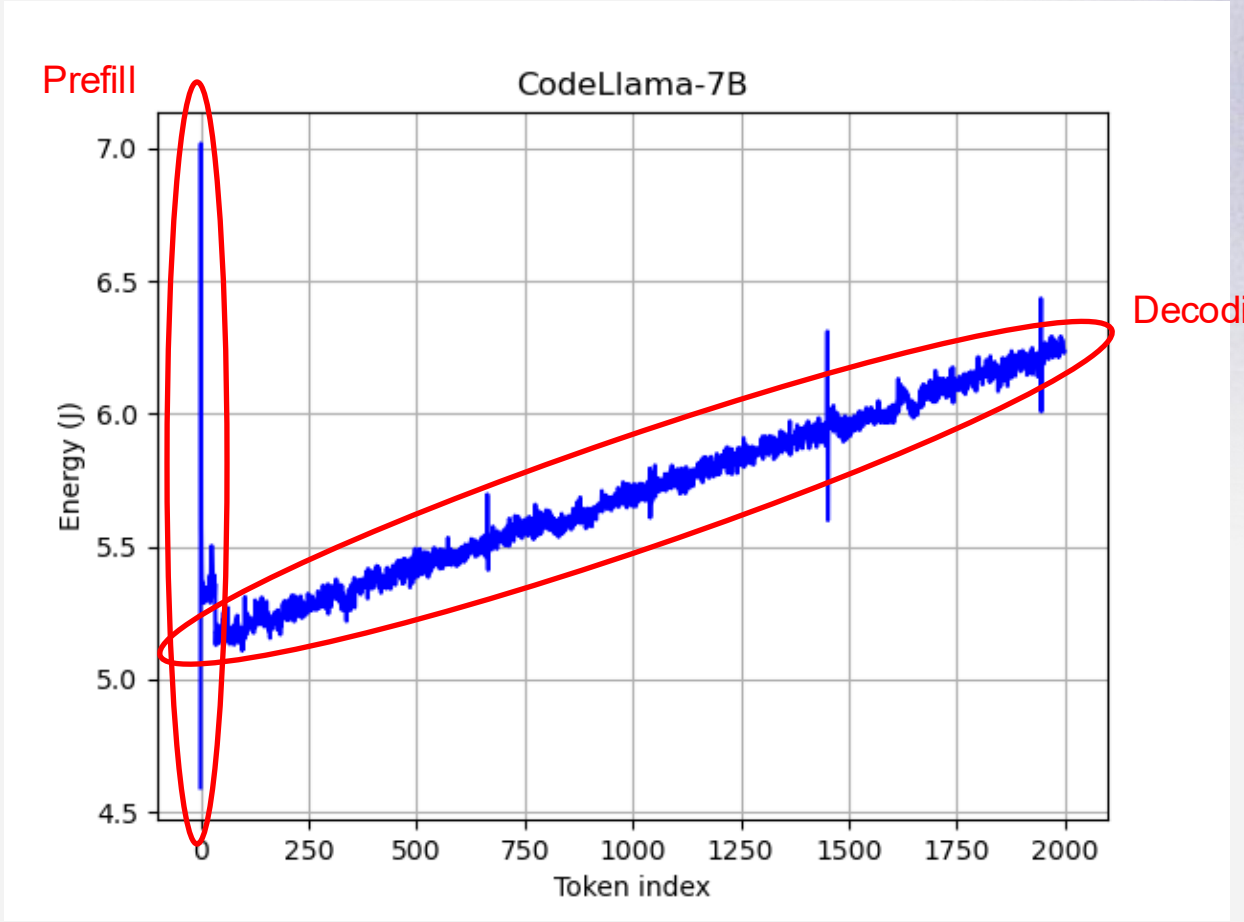
- Parallel process monitoring power consumption
- 100Hz
- pyNVML
- Records: timestamps, power consumption

INFERENCE EVENT TRACE

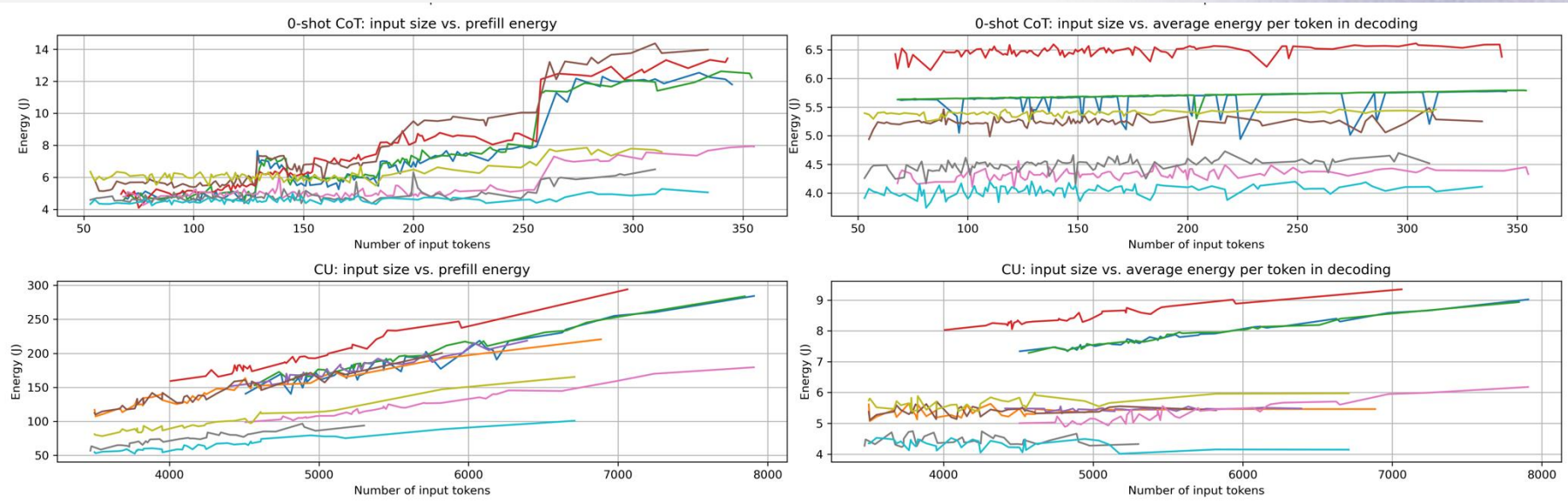
- Records: timestamps for start of prefill and each generated token



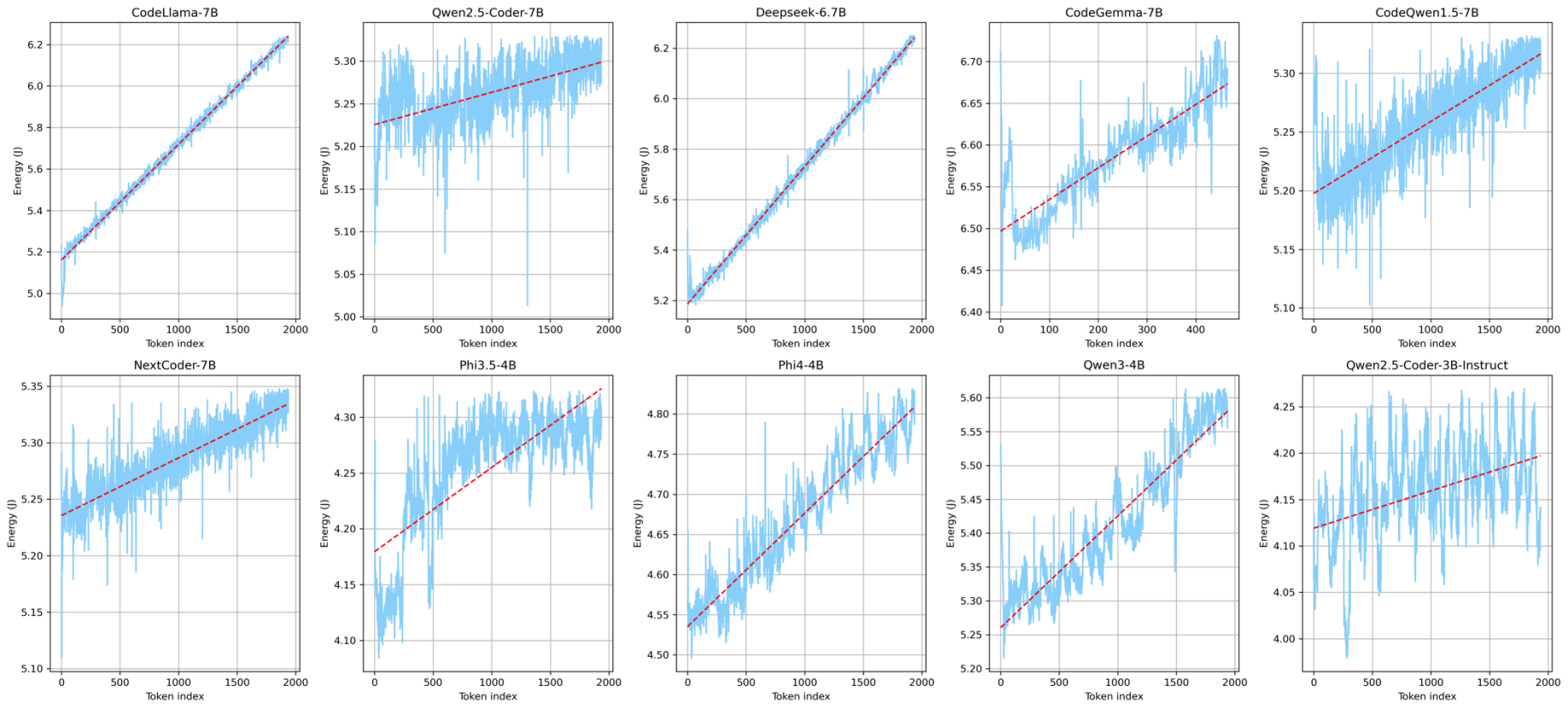
[4]: Timur Babakol and Yu David Liu. 2024. Tensor-Aware Energy Accounting. In Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE '24). Association for Computing Machinery, New York, NY, USA, Article 93, 1–12. https://doi.org/10.1145/3597503.3639156


Prefill
CodeLlama-7B
Decoding
Energy (J)
7.0
6.5
6.0
5.5
5.0
4.5
0
250
500
750
1000
1250
1500
1750
2000
Token index

Prefill


0-shot CoT: input size vs. prefill energy
Energy (J)
Number of input tokens
0-shot CoT: input size vs. average energy per token in decoding
Energy (J)
Number of input tokens
CU: input size vs. prefill energy
Energy (J)
Number of input tokens
CU: input size vs. average energy per token in decoding
Energy (J)
Number of input tokens
CodeLlama-7B
Qwen2.5-Coder-7B
Deepseek-6.7B
CodeGemma-7B
CodeQwen1.5-7B
NextCoder-7B
Phi3.5-4B
Phi4-4B
Qwen3-4B
Qwen2.5-Coder-3B-Instruct

Decoding


CodeLlama-7B
Qwen2.5-Coder-7B
Deepseek-6.7B
CodeGemma-7B
CodeQwen1.5-7B
NextCoder-7B
Phi3.5-4B
Phi4-4B
Qwen3-4B
Qwen2.5-Coder-3B-Instruct
Energy (J)
Token index

MAIN FINDINGS

- Prefill accounts for only a small share of the total energy consumption
- Prefill energy increases with input length (the extent depends on the model)
- Prefill impacts the cost of the first token in the decoding phase
- Each newly generated token becomes more expensive than the previous one (*the difference between the first and last can be up to 20%!*)
- The pattern of increase is dependent on the model family, hence on the implementation differences.
- Models like CodeLlama, Deepseek, Qwen3 *babble,* meaning producing tokens up to the allowed maximum

IMPLICATIONS:

- Input-dominated tasks (e.g., classification): Prefer models less sensitive to increases in input size.
- Output-heavy tasks (e.g., text generation): Prefer models with minimal per-token energy growth and avoid *babbling* models.