

Measuring and understanding energy use in Large Language Model inference

Erik Johannes Husom

2024-03-21

Section 1

Introduction

Carbon footprint of machine learning (ML)


The New York Times

SUBSCRIBER-ONLY NEWSLETTER

Climate Forward

A.I. Could Soon Need as Much Electricity as an Entire Country

Behind the scenes, the technology relies on thousands of specialized computer chips.

 Share full article



Machine learning and computational resources

Three eras of machine learning¹:

¹Sevilla et al. (2022): [*Compute Trends Across Three Eras of Machine Learning*](#)

Machine learning and computational resources

Three eras of machine learning¹:

- 1 *Pre Deep Learning Era* (1952-2010)
- 2 *Deep Learning Era* (2010-2022)
- 3 *Large-Scale Era* (2015-2022)

¹Sevilla et al. (2022): [Compute Trends Across Three Eras of Machine Learning](#)

Machine learning and computational resources

Training compute (FLOPs) of milestone Machine Learning systems over time
n = 121



Figure 1: Size of ML models over time²

Machine learning and computational resources

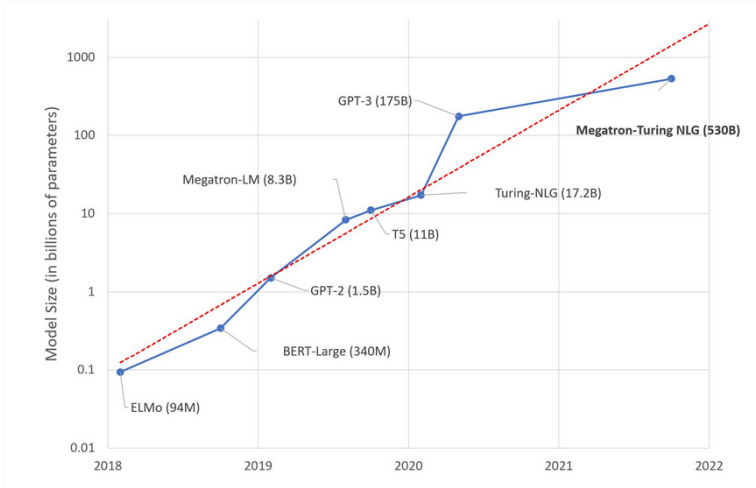


Figure 2: Size of LLMs over time

Machine learning and computational resources

- The last decade of research in machine learning³:
 - The carbon footprint of ML is *increasing*
 - Ca. 70% of the ML-models was trained on high carbon energy sources
 - *Transformer* models are very popular – and very carbon intensive
 - **Larger energy consumption does not necessarily mean better performance**
- LLMs are widely deployed
 - ChatGPT:
 - Reached 1 million users in 5 days
 - Currently over 100 million active users
 - Difficult to estimate resource consumption

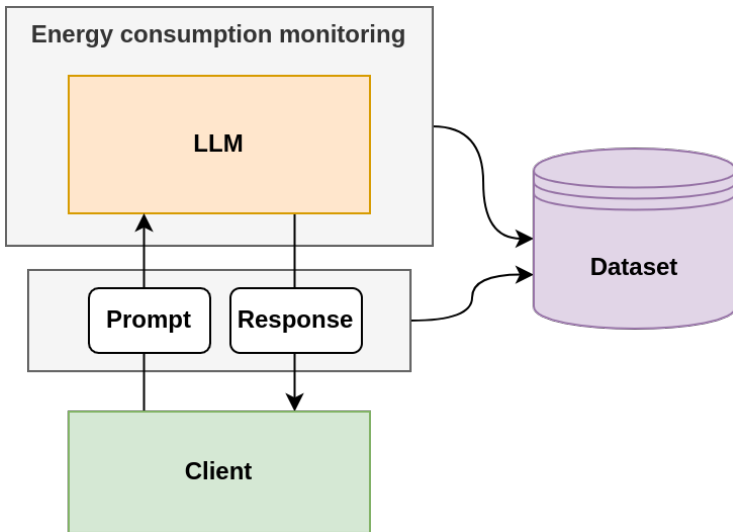
³Luccionu et al. (2023): [Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning](#)

Section 2

Our work: The price of prompting

What is the energy consumption of LLMs?

What is the energy consumption of LLMs?



Our framework – Technical details

- Developed in Python
- External tools:
 - Monitoring of power consumption: [Scaphandre](#)
 - LLM service with OpenAI-compatible API

Scaphandre

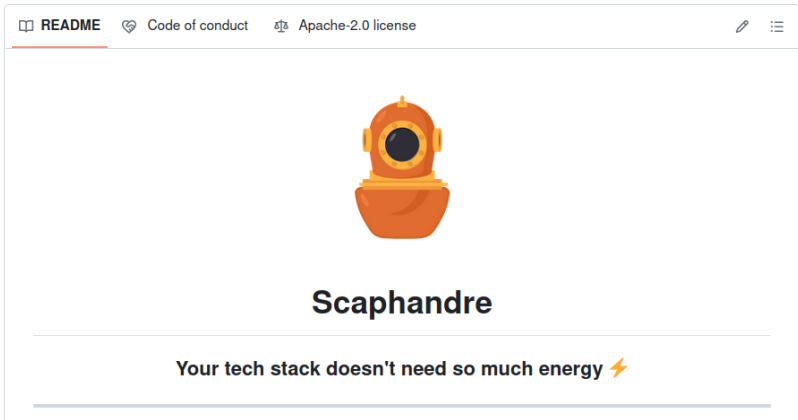


Figure 3: [Scaphandre](#) – open source tool for energy consumption metrics.

Scaphandre – Example of monitoring

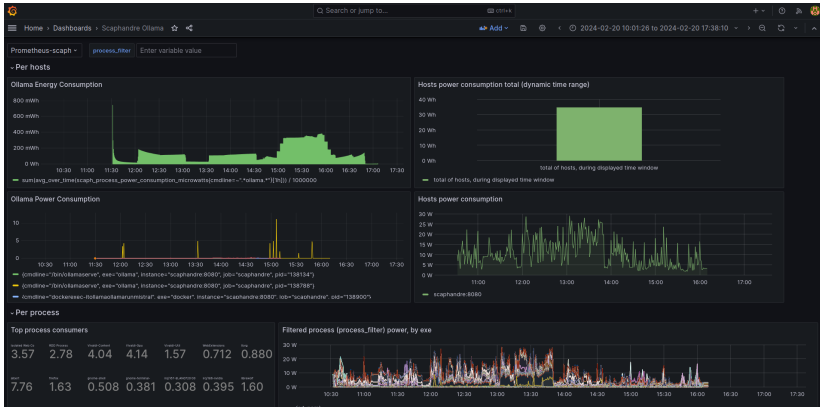


Figure 4: Grafana visualization of Scaphandre monitoring an LLM service.

LLM services



Get up and running with large language models, locally.

Run [Llama 2](#), [Code Llama](#), and other models. Customize and create your own.

Download ↓

Available for macOS, Linux, and Windows (ARM64)

GPT4All

A free-to-use, locally running, privacy-aware chatbot. **No GPU or internet required.**



Easy, fast, and cheap LLM serving for everyone

llamafile



Ollama

- Easy to install
- CPU and GPU support
- Wide range of available models
- REST API
- ... and more



Get up and running with large language models, locally.

Run [Llama 2](#), [Code Llama](#), and other models.
Customize and create your own.

Download ↓

Available for macOS, Linux,
and Windows (preview)

Open source LLMs



Models

Featured



gemma

Gemma is a family of lightweight, state-of-the-art open models built by Google DeepMind.

↓ 210.5K Pulls ↻ 69 Tags ⌚ Updated 13 days ago

llama2

Llama 2 is a collection of foundation language models ranging from 7B to 70B parameters.

↓ 784.9K Pulls ↻ 102 Tags ⌚ Updated 5 weeks ago

mistral

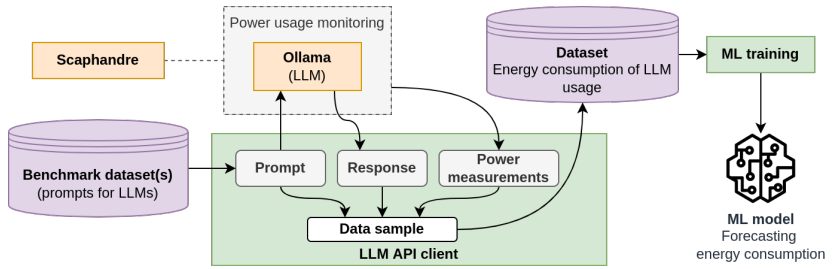
The 7B model released by Mistral AI, updated to version 0.2.

↓ 351.8K Pulls ↻ 53 Tags ⌚ Updated 2 months ago

mixtral

A high-quality Mixture of Experts (MoE) model with open weights by Mistral AI.

Our framework



Our framework – Methodology

- 1 Get prompt from dataset (or write one manually)
- 2 Start power monitoring (Scaphandre)
- 3 Query LLM service with prompt
- 4 Receive response from LLM service
- 5 Stop power monitoring
- 6 Save prompt, response, metadata, and metrics.

Inference power usage

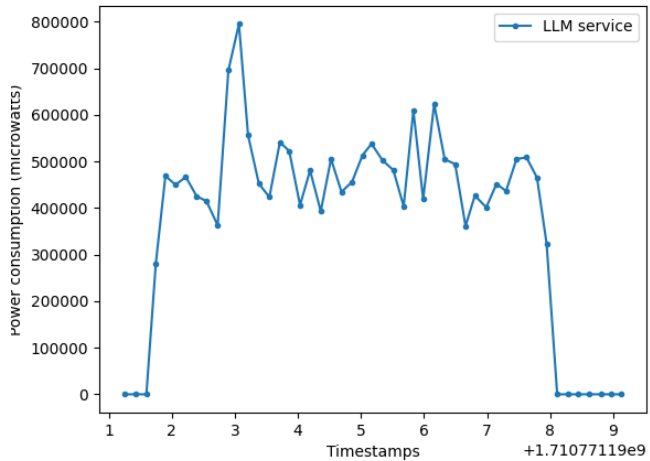


Figure 6: Example of power use during LLM inference.

Correlation matrix

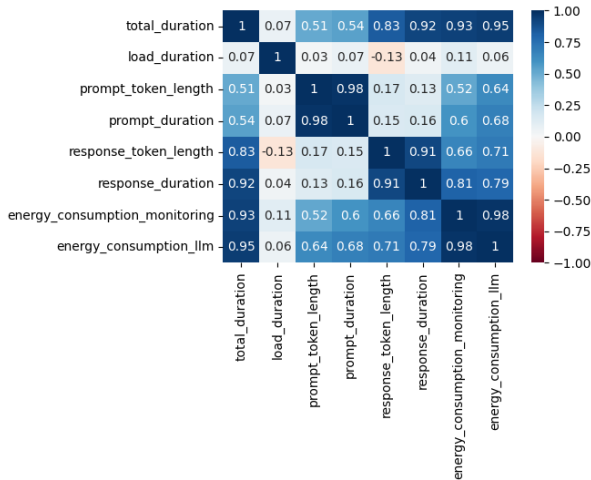


Figure 7: Correlations of data from ~5000 general purpose prompts.

LLM energy consumption

- What to do with the tool and the data?
 - What was the cost of this chatbot conversation (monitoring)?
 - What will this task cost (forecasting)?
 - What model is most efficient (for a given task)?
 - What deployment service is most efficient?
 - Find a balance between efficiency and performance

LLM energy consumption

- How to forecast consumption?
 - Prompt length as input – too simple
 - NLP-based model – may be too complex
- Forecasting based on type of task

Section 3

Summary

Summary

- Our work:
 - Make a framework for monitoring LLM services
 - Create datasets for LLM usage and energy consumption
 - Analyze and create models that can forecast and guide towards more sustainable LLM usage
 - *Paper coming soon*



Figure 8: [Link to slides](#)

- erik.johannes.husom@sintef.no