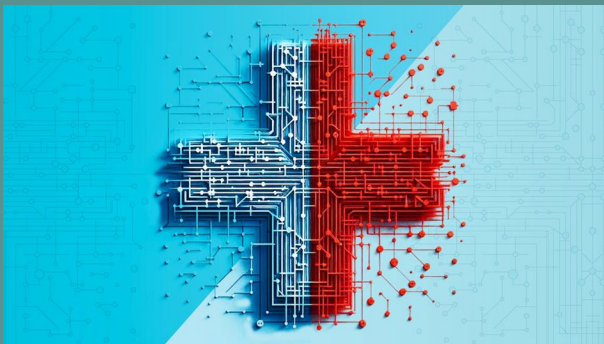




Measure energy use of Apertus

Understanding the energy behaviour of Apertus

Agustín Herrericapicazo
Luis Barros
Stefan Aeschbacher

agustin.it@proton.me
luisantonio1998@gmail.com
stefan@aesbacher.ch



Goals

The goal of the project was to

- measure the energy consumption of inference on Apertus
- understand the influence of prompt and response on the energy consumption
- Bonus: compare to llama



Measuring Infrastructure

Our measuring infrastructure consisted of

- Mac Studio with remote access.
- powermetrics tool
- Begasoft BrandBot





Experiments

Experiment 1: Prompt/Response length

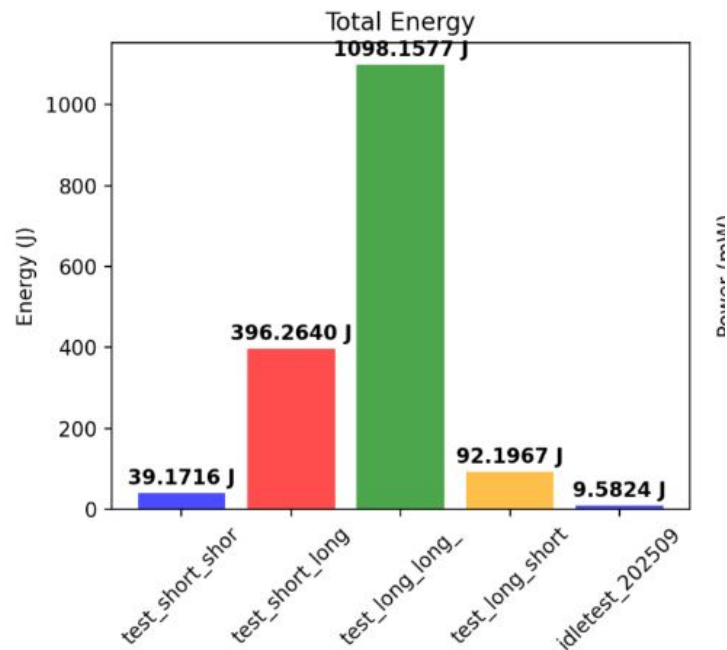
We tested different prompt and response lengths.

Ex: Short-Long: Explain photosynthesis in detail.

Experiment 2: Different subjects

Experiment 4: Different languages

Experiment 3: Compare with Llama





Learnings

We failed successfully on:

- getting a stable setup ✗
- excluding overhead ✗
- proper energy calculations ✗
- getting 2nd system to run ✗
- do physical measurements ✗
- change measurement method ✗
- automate everything ✗
- doing proper statistics ✗

We succeeded successfully on:

- measure idle consumption ✓
- measure different prompts ✓
- Integrated a new team member ✓
- learn a lot ✓
- **having fun!** ✓



Results

Thanks to all!

- Measuring is hard!
- Prompt size:
 - long answers dominate the energy use
 - long prompts have some influence
- Different types of prompts: Basic knowledge, Chemistry, Geography, Math
 - does not seem to influence the energy consumption (other than the length of the response)
- Different languages:
 - apart from the length, the language seems to have some influence (portuguese: +20%)
- Different models: inconclusive but similar