Can AI be a Classic Example of Jevon's Paradox

Barış Sanlı, barissanli.com

Code and Data: https://github.com/barissanli/powerconsumption-ai

One bold claim about AI's electricity use of AI is its comparison to traditional internet search. In this brief, by experimenting with offline AI models, detailed energy consumption of AI queries will be compared. The discussion part argues that "for completing a task", AI may be significantly more energy efficient than "search-read-devise" approach to completing tasks. However the easiness and completeness may lead to increased usage of AI and higher energy consumption.

Introduction

Currently AI models are becoming more energy and compute efficient. By selectively activating certain parts or dividing the task into experts, the whole process is getting much leaner and efficient. As a result, better models can now run offline on laptops or desktop computers.

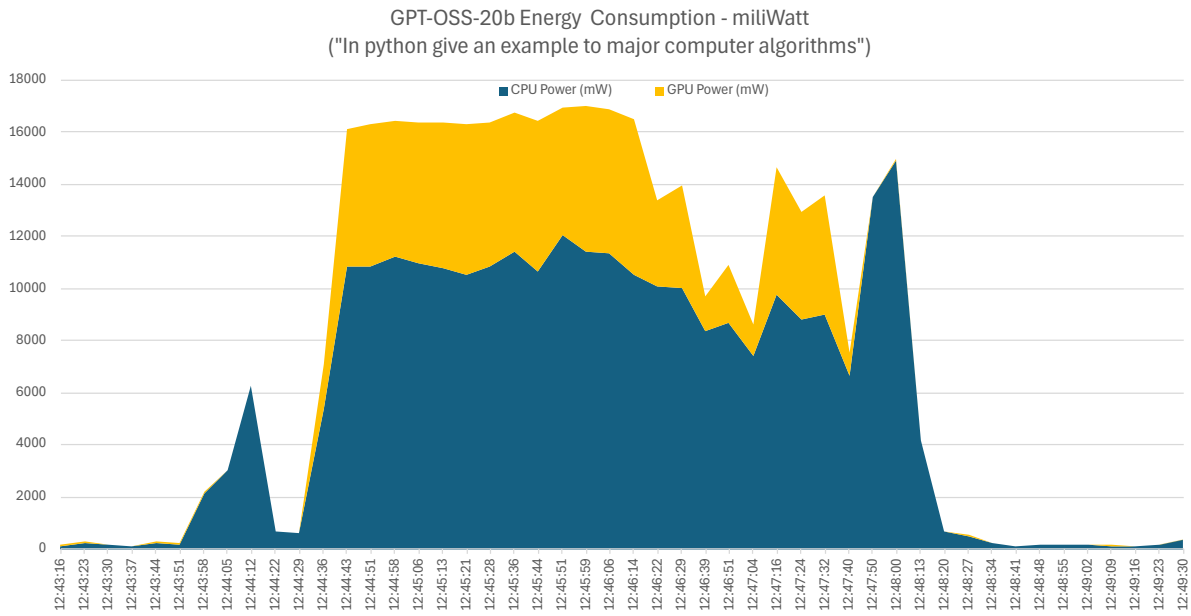
The latest open source ChatGPT 20 and 120 billion parameter models are a great example to this advancement. Therefore we start from this latest gpt-oss:20b model. The computer setup details are intentionally omitted, but the energy consumption is notably lower for its class.

The coding and data files can be reached from
https://github.com/barissanli/powerconsumption-ai

Local GPT

The offline gpt-oss:20b was tasked with generating an example for every major computer algorithm in python. The task took 4 minutes to complete with a token generation rate of 12 tokens/seconds. The power consumption has reached 16 Watts. The total energy cost in watt hours depends on the preparation and disposal of model from the memory. According to prompt statistics, the generation lasts 193seconds.

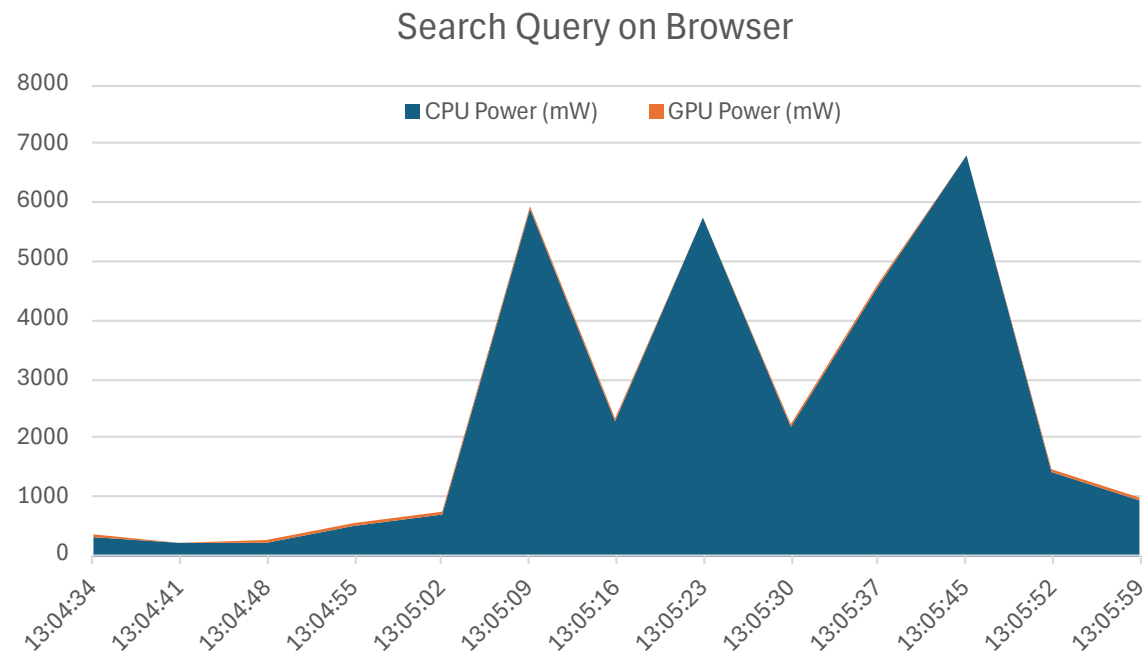
The total energy cost of the generationj is 0.86 Wh(watthour) or depending on the loading and cleanup processes (preparation), maximum consumption is 1.2 Wh.



The query result provided code that executes successfully at the first attempt without re-prompting.

Search Query

How can we compare this to a simple query? The same task to generate Python examples to major algorithms requires research time using a browser. The browser used is one of the most efficient for that laptop. Each search and review consumes 6-6.7 Wh. But this value fluctuates with bursts and dips.



The search was completed quickly, with the early queries yielding the relevant results. This provided insight into the energy cost of queryin. On an energy efficient machine, the process consumes around 4 Watts per unit of time.

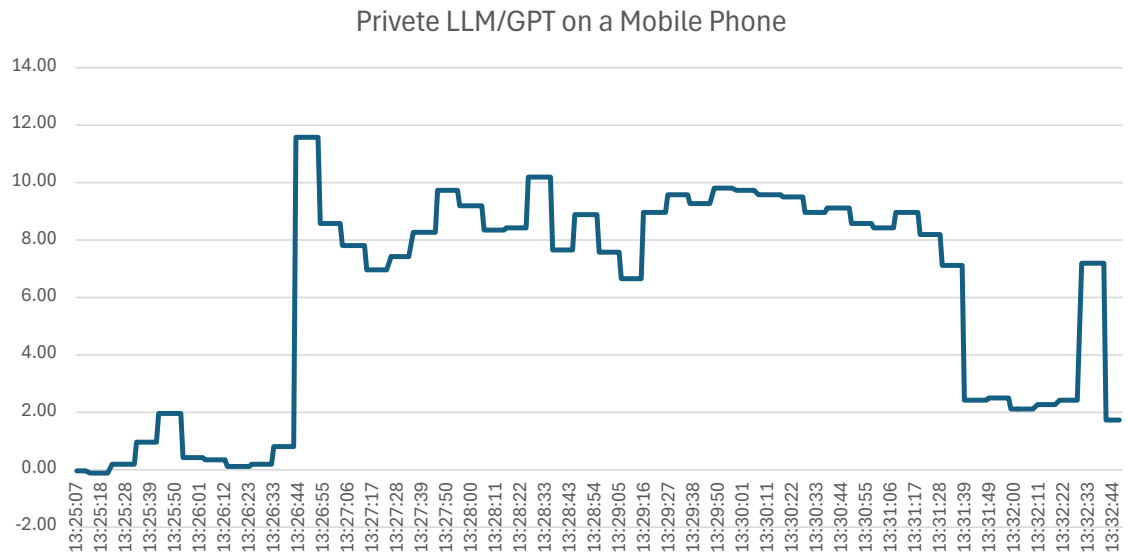
For this specific query, which took 1 minute luckily, the total energy cost was 0.067 Wh. However the details matter.

If “search-read-devise” approach takes four times longer than a GPT query for a given task, the GPT becomes much more efficient.

Mobile

Some models are compact enough to run on modern mobile phones and can generate acceptable answers. One such model belongs to the family of offline/Private GPT or LLMs. Power consumption was measured via the USB interface. Positive values indicated charging, while negative values reflected battery drain. However these negative values account for both USB power input and battery power consumption.

Measurement must be offset to zero, when USB is plugged in. The query is the same as before. Model is Gemma 2B-Q5KM. The end result is acceptable.



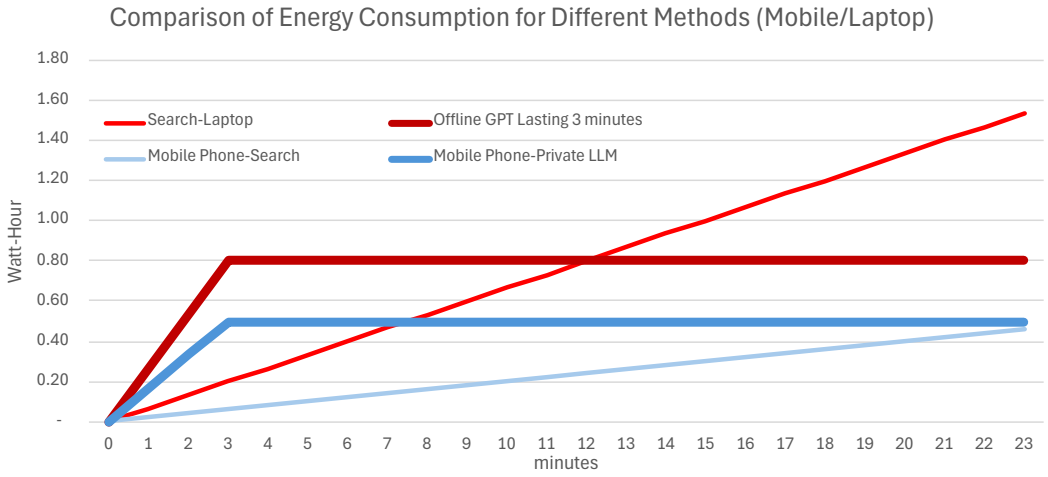
In a phone setup, the power consumption peaks at 12 Watt, but sustained power usage is close to 9-10 Watts.

Discussion

The results are compiled into a single table below. The primary factor in energy consumption is the duration required to complete a task. The server-communication energy costs are ignored. The laptop's energy consumption is 4 Watts for an energy efficient processor.

Method	Peak Power (watt)	Sustained Power (watt)	Duration
Laptop – Offline GPT	17	16	4 min
Laptop -Query	6.7	4	1 min
Mobile – Offline LLM	12	9	3 min
Mobile – Query	2	1.2	1 min

In terms of energy consumption for a complete task, GPT becomes much more efficient requiring more than 12 minutes of "search-read-devise" in this example, or four times the duration required for GPT.



Why Jevon's Paradox

This section reflects the personal experience than a scientific conclusion. LLM's popularity is a testimony to demand for quick answers and cost-benefit preference of users to completing tasks through a single point query (GPT) rather than iterative search and reads.

Since GPT typically provides answers under 3 minutes, it is more satisfactory than search for more users. This encourages users to seek more details and explore additional aspects.

In programming for example, completing a coding solution successfully often requires multiple searches, forum queries and test runs. However as LLMs become more competent, they can deliver better results can be reached on the first attempt. This encourages users to add more features or conduct additional checks.

Conclusion

In this controlled experiment, we analyzed the energy consumption of offline AI models on a laptop and mobile device. Then these results were compared to traditional search based task completions. When a task lasts longer to complete, GPT consistently demonstrated significantly greater efficiency.

Specfically, for this specific experiment, the duration factor was 4. This means tasks taking four times longer via search-read-devise consumes more energy than GPT. Previous studies omitted the energy use of “user” and mostly assumes energy consumption of search or AI in vacuum. This study approaches the issue with a real life example and incorporates total energy cost of user’s device and applications.

Barış Sanlı