

University of Minho
School of Engineering

Master's Dissertation in Informatics Engineering

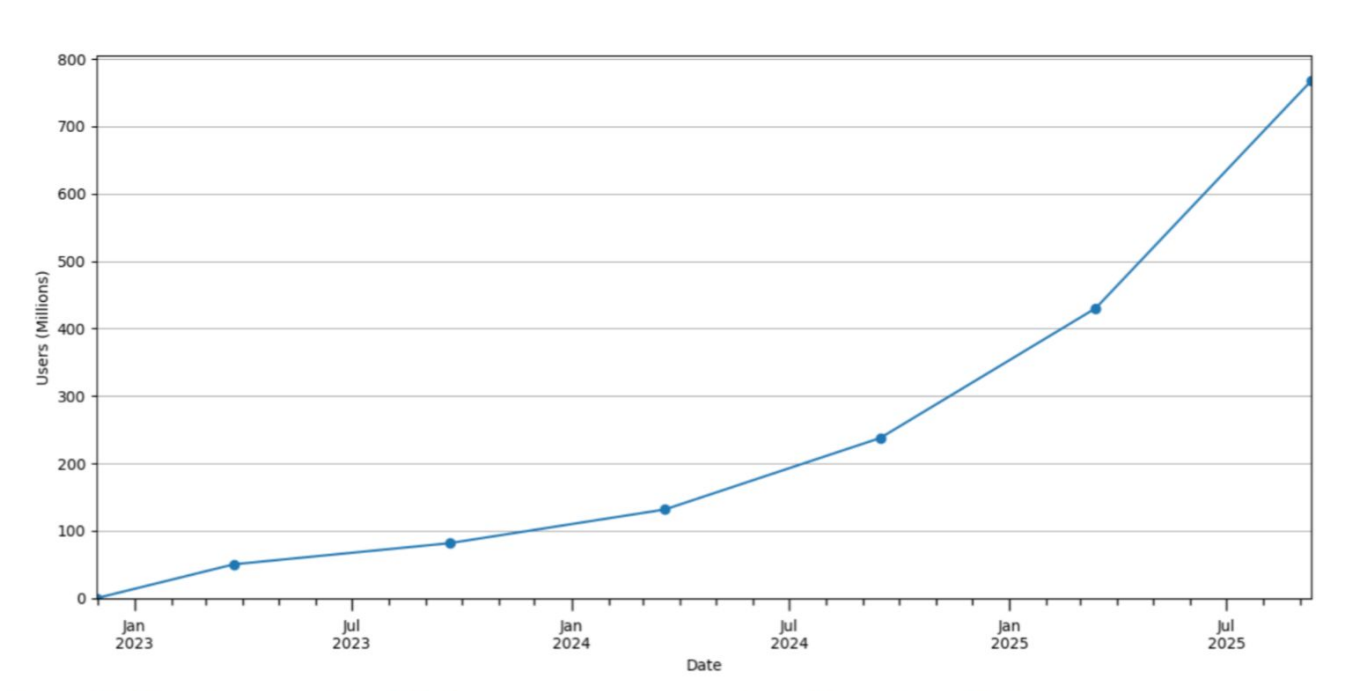
The Evolution of Intelligence Efficiency in Local, Open Large Language Models: A Controlled Empirical Study

Diogo Gomes Matos

Supervised by

João Saraiva [a]
Roberto Verdecchia [b]

[a] University of Minho, Portugal · [b] University of Florence, Italy



Weekly active ChatGPT users on consumer plans (Free, Plus, Pro), shown as point-in-time snapshots every six months, November 2022–September 2025

[1] A. Chatterji et al., "How People Use ChatGPT," Working Paper 34255, Sep. 2025. doi: 10.3386/w34255

700M+

Weekly active users



Of the global adult population

Training GPT-3 took

1,287 MWh

≈ the yearly energy use of **420 people**

[2] D. Patterson et al., “Carbon Emissions and Large Neural Network Training,” arXiv preprint, 2021, doi: 10.48550/arXiv.2104.10350
[3] M. F. Argerich and M. Patiño-Martínez, “Measuring and Improving the Energy Efficiency of Large Language Models Inference,” IEEE Access, vol. 12, pp. 80194–80207, 2024, doi: 10.1109/ACCESS.2024.3409745

1 **year** of **inference** requires

25×

The compute of **training** GPT-3

[4] A. A. Chien, L. Lin, H. Nguyen, V. Rao, T. Sharma, and R. Wijayawardana, "Reducing the Carbon Impact of Generative AI Inference (today and in 2035)," in Proceedings of the 2nd Workshop on Sustainable Computer Systems, in HotCarbon '23. Boston, MA, USA: Association for Computing Machinery, 2023. doi: 10.1145/3604930.3605705

Existing studies...

Compare models in a single moment in time

Change the models along with the hardware

Rely on figures reported by the model providers

Leave accuracy out of the picture

None measures the **energy** and **accuracy** of **successive** releases on the **same hardware**.

Is energy efficiency improving with each release or is it being systematically compromised in favor of raw performance?

Intelligence Efficiency

Accuracy / Energy

[5] J. Saad-Falcon et al., “Intelligence per Watt: Measuring Intelligence Efficiency of Local AI,” arXiv preprint, 2025, doi: 10.48550/arXiv.2511.07885

RQ1 - Evolution

How has the energy efficiency of open LLMs **evolved** across successive **releases**?

RQ2 - Cost of Accuracy

What is the energy **cost** of the **accuracy** gained by newer models?

RQ3 - Design Choices

How do **size**, **architecture** and **operating mode** influence energy efficiency?

RQ4 - Model Families

Do the model **families** follow a common **trajectory**, or do they **diverge**?

Experiment Design

Models

47 models

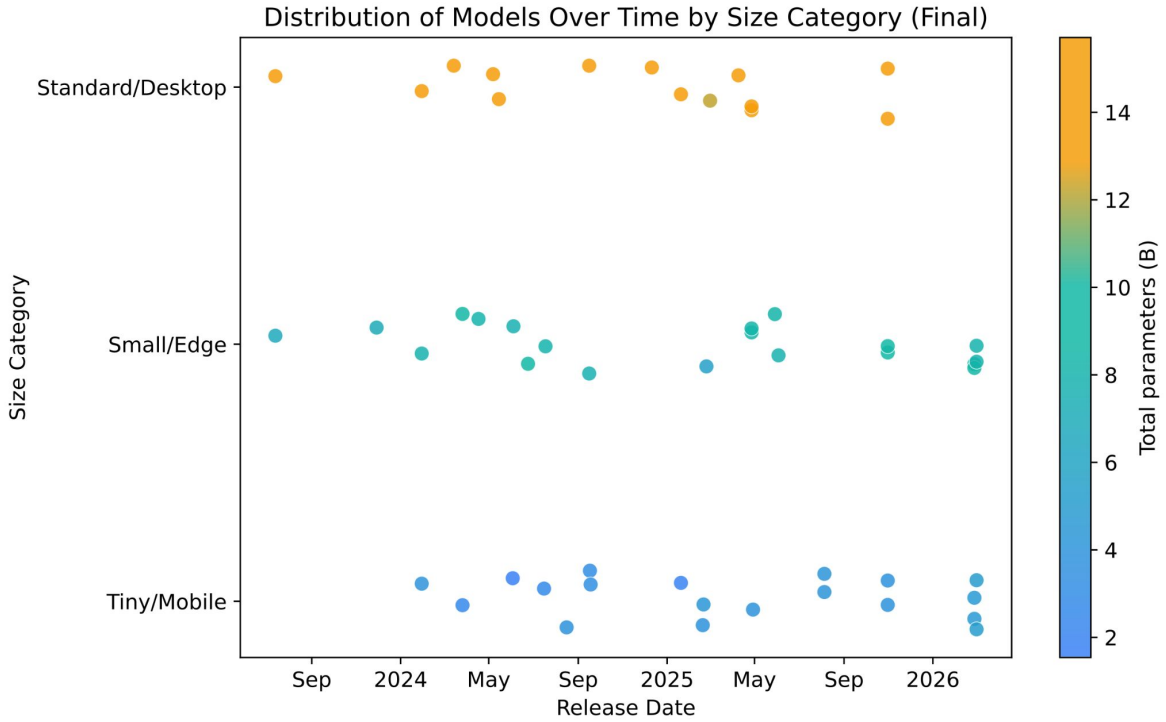
20 releases

6 families:

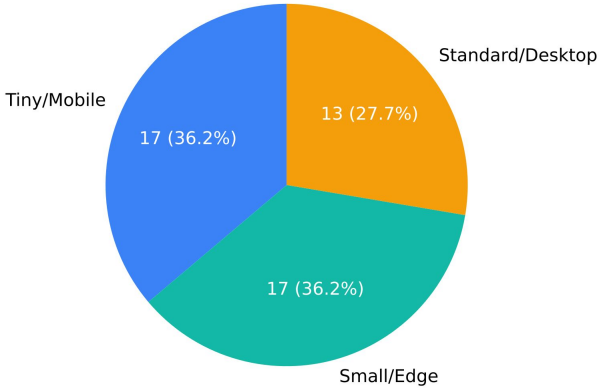
- Mistral
- Gemma
- Llama
- Phi
- DeepSeek
- Qwen

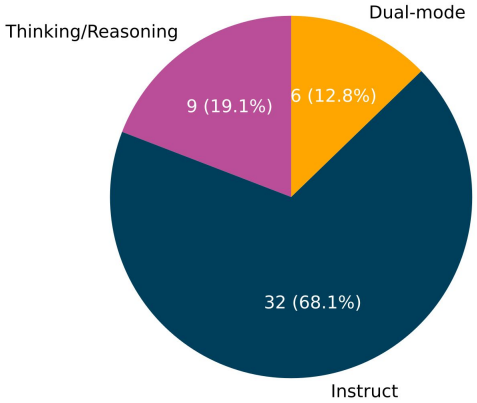
53 model runs

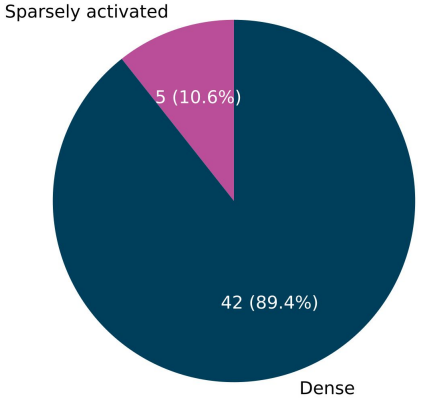
Model Selection



Model Selection







Experiment Design

Models

47 models
20 releases
6 families:

- Mistral
- Gemma
- Llama
- Phi
- DeepSeek
- Qwen

53 model runs

×

Dataset

MMLU-Pro
sampled to
372 questions

×

10 repetitions
(random order)

=

Results

530 runs
197,160 inferences

Testing Environment

OS/Hardware

Ubuntu 22.04.5 LTS
2 × Intel Xeon Gold 5418Y
NVIDIA L40S (48GB VRAM)

Measurement Stack

vLLM
Intel RAPL
NVML

≈ 50 days

Of runtime

≈ 342 kWh

Of energy consumption (GPU)

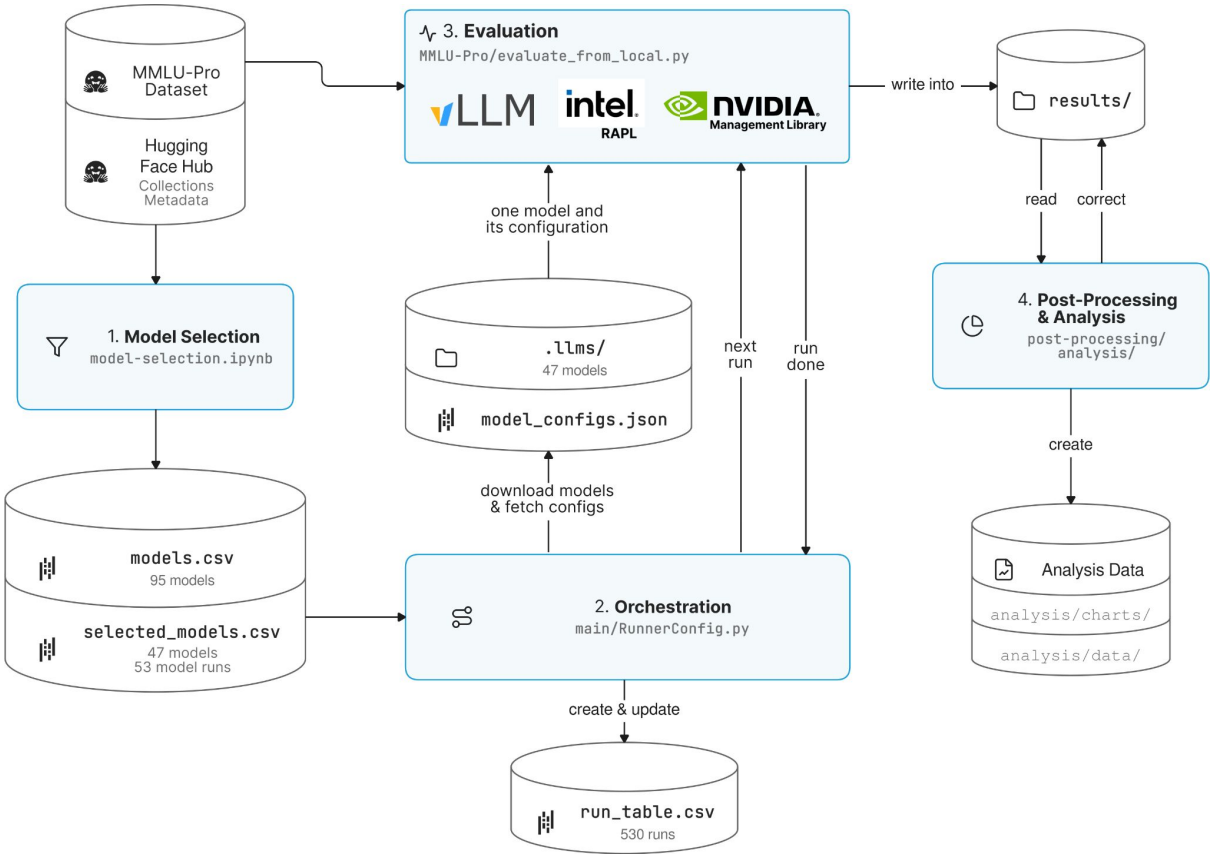
≈ 1,810 km

Travelled in an electric car

(191 Wh/km)

[6] EV Database, “Energy consumption of full electric vehicles.” Accessed: Oct. 07, 2026. [Online]. Available: https://ev-database.org/cheatsheet/energy-consumption-electric-car

Experiment Architecture



Results

Analysis

One ratio, three factors

$$\text{Intelligence Efficiency} = \frac{\text{Accuracy}}{\text{Energy per question}}$$

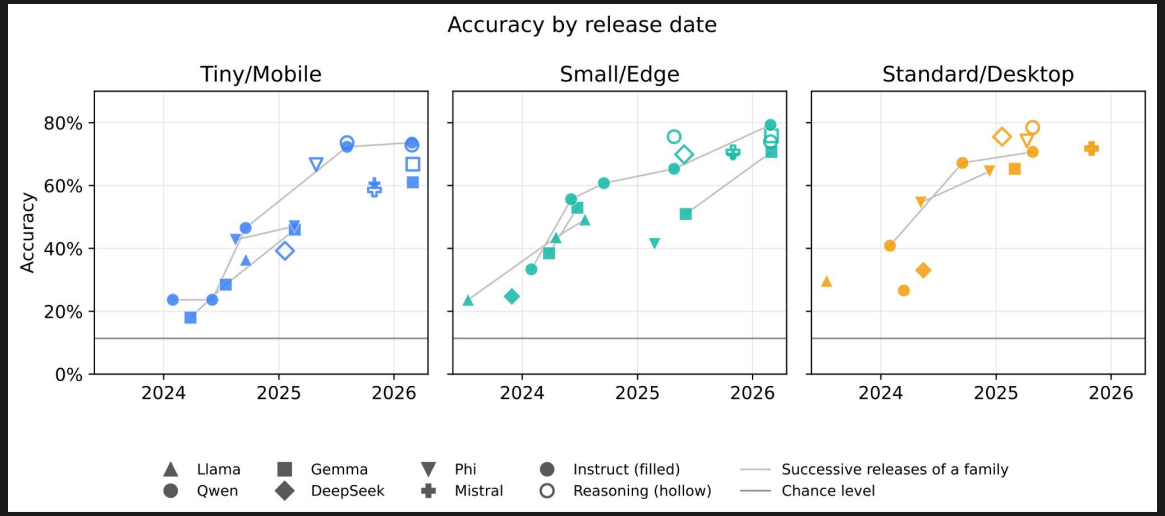
Analysis

One ratio, three factors

$$\text{Intelligence Efficiency} = \frac{\text{Accuracy}}{\text{Tokens per question} \times \text{Energy per token}}$$

RQ1 · EVOLUTION ACROSS RELEASES

Accuracy rose by about 20 points per year



+18–24

Accuracy points per year, depending on the size category.

Median instruct run:

25% in 2023

72% in 2026

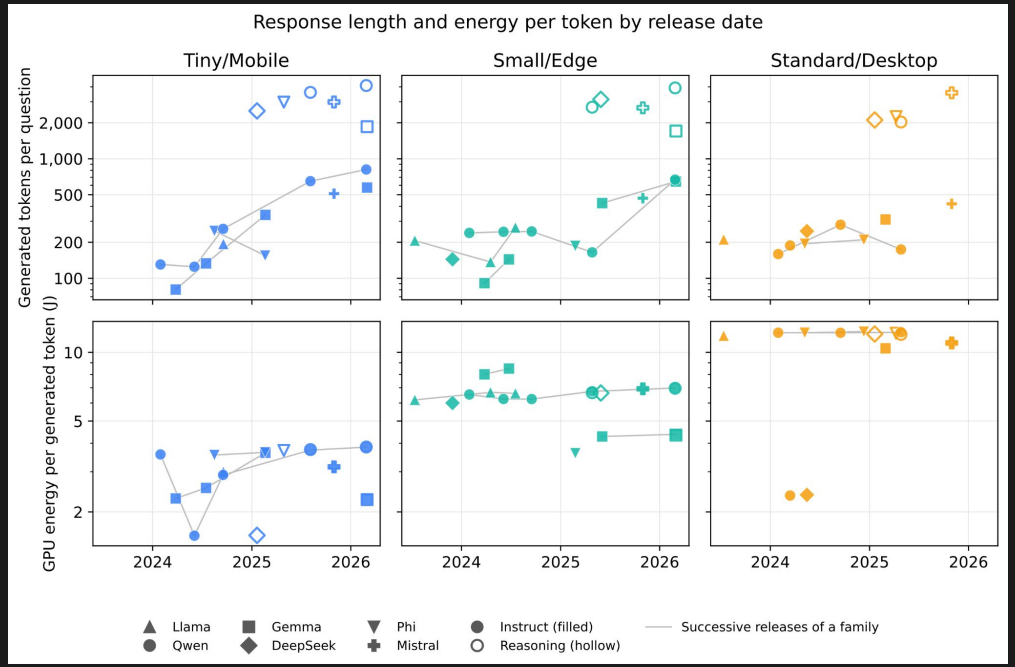
Analysis

One ratio, three factors

$$\text{Intelligence Efficiency} = \frac{\text{Accuracy} \uparrow}{\text{Tokens per question} \times \text{Energy per token}}$$

RQ1 · EVOLUTION ACROSS RELEASES

Responses grew longer, each token cost the same



194 → 657

Generated tokens per question of the median instruct run, from 2024 to 2026

Energy per token shows no trend over time in any size category.

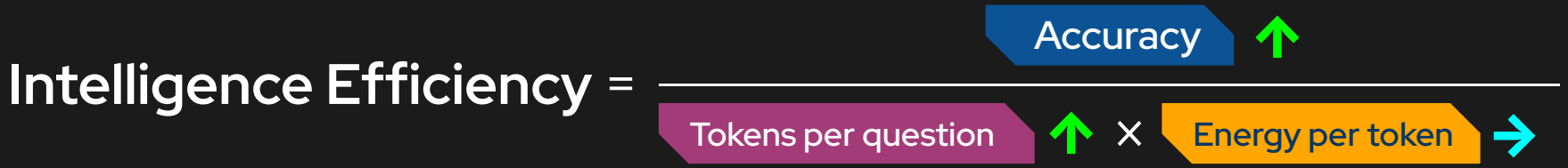
In 18 pairs of successive releases from the same model family:

17 became more accurate

13 with longer responses

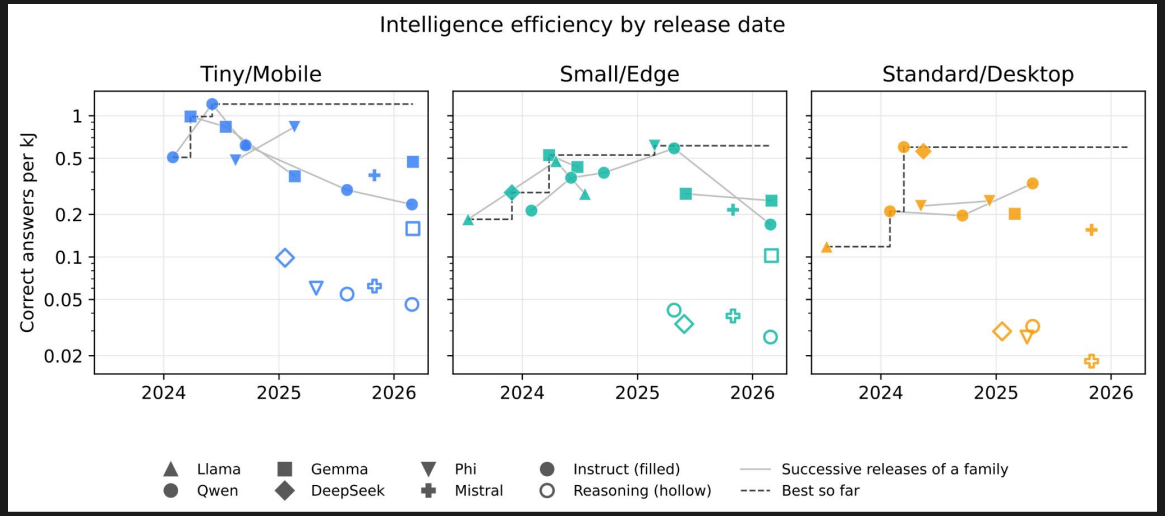
Analysis

One ratio, three factors



RQ1 · EVOLUTION ACROSS RELEASES

Intelligence efficiency did not improve for released models



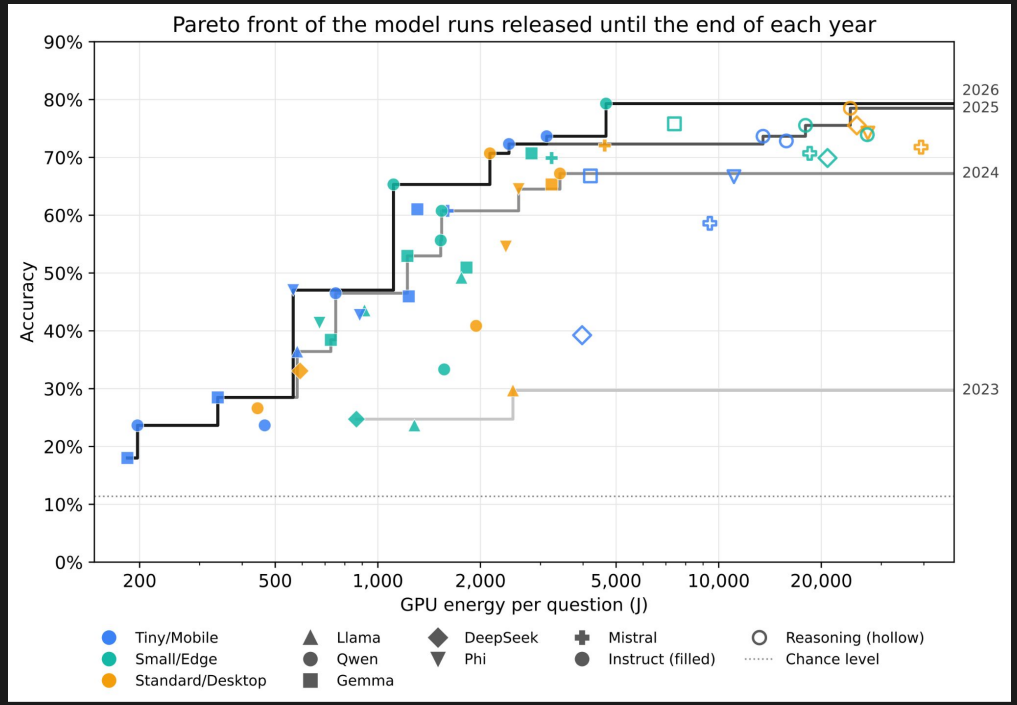
Only 8 of 18

Release pairs improved intelligence efficiency, even though 17 became more accurate.

| The gain in accuracy was
| matched by the growth of
| the responses.

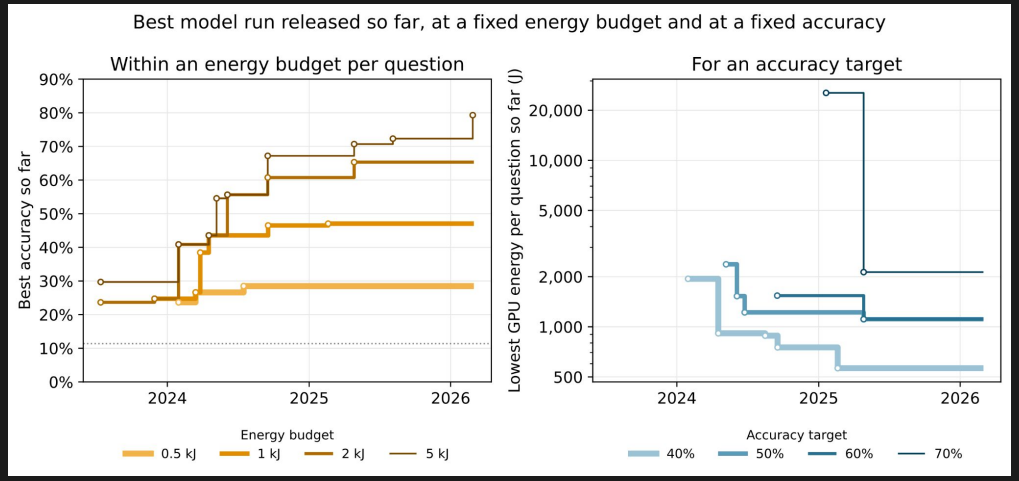
RQ2 · ENERGY COST OF ACCURACY

A fixed energy budget came to buy more accuracy



RQ2 · ENERGY COST OF ACCURACY

A fixed energy budget came to buy more accuracy



25% → 65%

Best accuracy within 2 kJ per question, from 2023 to 2025.

A given accuracy came to need

1.4–3.4× less energy

from early 2024 to April 2025.

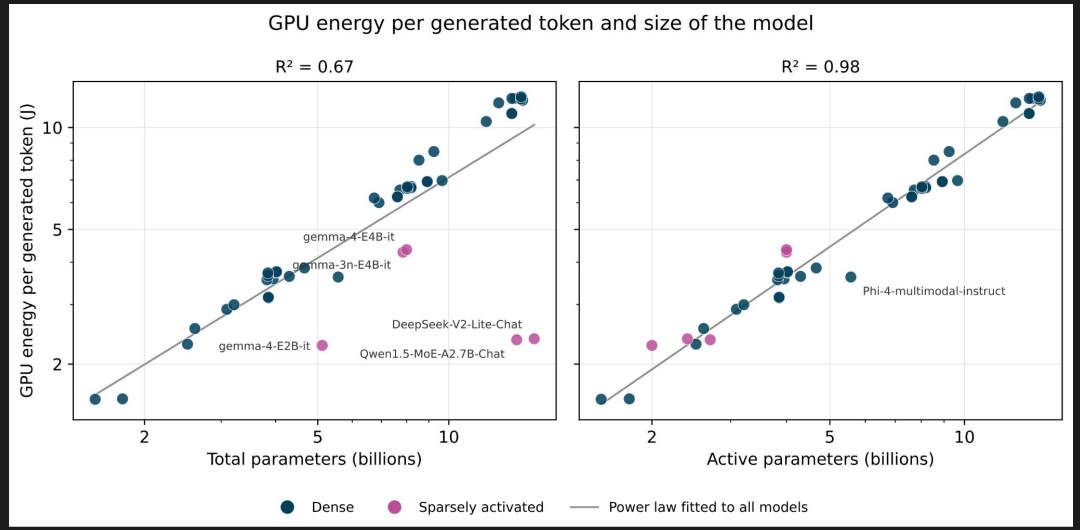
Analysis

One ratio, three factors

$$\text{Intelligence Efficiency} = \frac{\text{Accuracy}}{\text{Tokens per question} \times \text{Energy per token}}$$

RQ3 · DESIGN CHOICES

The energy of a token follows active parameters



98%

Of the variation in energy per token is explained by active parameters (67% by total parameters).

Sparse activation gave the energy of a small model, not the accuracy of a large one.

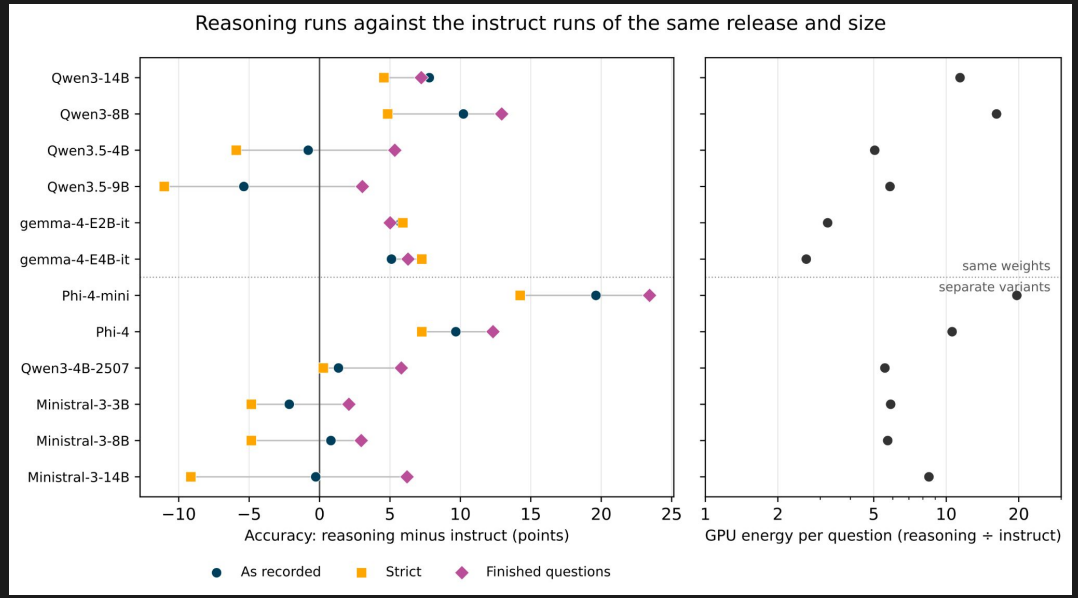
Analysis

One ratio, three factors

$$\text{Intelligence Efficiency} = \frac{\text{Accuracy}}{\text{Tokens per question} \times \text{Energy per token}}$$

RQ3 · DESIGN CHOICES

Reasoning multiplied energy for a modest gain



2.6–19.6×

The energy per question of the instruct counterpart (median 5.9×, 12 pairs).

For a median gain of only

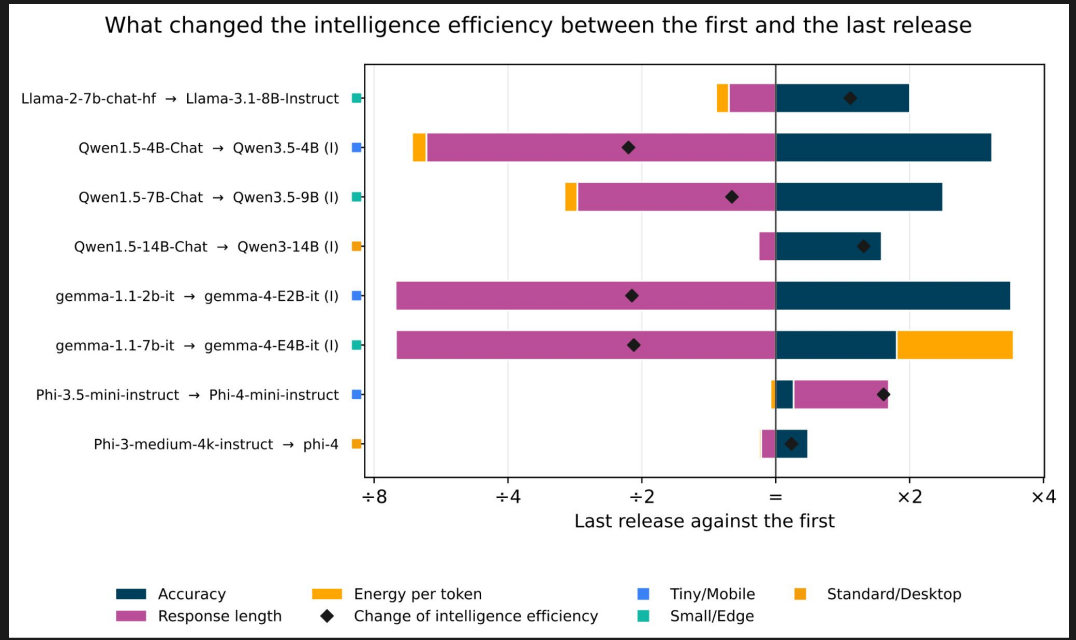
+3.2 accuracy points

Less accurate in

4 of 12 pairs

RQ4 · TRAJECTORIES OF THE FAMILIES

Families: a common trajectory, with divergent means



Common results. All eight trajectories became more accurate, seven of them with longer responses.

No significant family effect at equal release date and size category.

Divergent means. Gemma adopted sparse activation and the least costly reasoning mode; Phi shortened its responses; Qwen lengthened them the most.

Is energy efficiency improving with each release or is it being systematically compromised in favor of raw performance?

The gains in efficiency were reinvested in accuracy.

RECOMMENDATIONS

What this work suggests

Recommendations

Treat response length as a design target,
and report the tokens generated per answer.

Make reasoning proportional to the question.

Develop and select small models:
the smallest one that attains the required accuracy.

Compare models within an energy budget,
not by accuracy alone.

CONTRIBUTIONS

What this work adds

Contributions

1. **Evidence.** Energy and accuracy of successive releases under fixed conditions, which **reconciles studies that appeared to disagree.**

2. **A method.** The **factors** a comparison of releases has to **control**, and a **decomposition** that traces each change to its origin.

3. **Open artefacts.** The **dataset**, with every response, and the **pipeline**, ready for new releases.

LIMITATIONS AND FUTURE WORK

Where the conclusions stop

Limitations

→ **One GPU**, **one engine**, **one request at a time**.

→ **One benchmark**, multiple choice and **English only**.

→ Models of **≤ 16B total parameters**

→ **Reasoning truncated at times**, so its energy is a lower bound.

→ Prompts for instruct models asked for **step-by-step** answers.

LIMITATIONS AND FUTURE WORK

What's next?

Future work

→ Repeat **without 'think step-by-step'** and the **full reasoning budget**.

→ **Larger and smaller models**, local **devices** and **runtimes**.

→ **Batched serving**, **long contexts** and **code generation**

→ **A continuous, efficiency-aware benchmark** built on the pipeline.

Thank you.

Open to questions!



University of Minho
School of Engineering

Master's Dissertation in Informatics Engineering

The Evolution of Intelligence Efficiency in Local, Open Large Language Models: A Controlled Empirical Study

Diogo Gomes Matos

Supervised by

João Saraiva [a]
Roberto Verdecchia [b]

[a] University of Minho, Portugal · [b] University of Florence, Italy