Small is Sufficient: Reducing the AI Energy Consumption Through Model Selection

T. D. S. Barros, F. Giroire, R. Aparicio-Pardo, J. Moulierac

GreenDays, Rennes, 2025

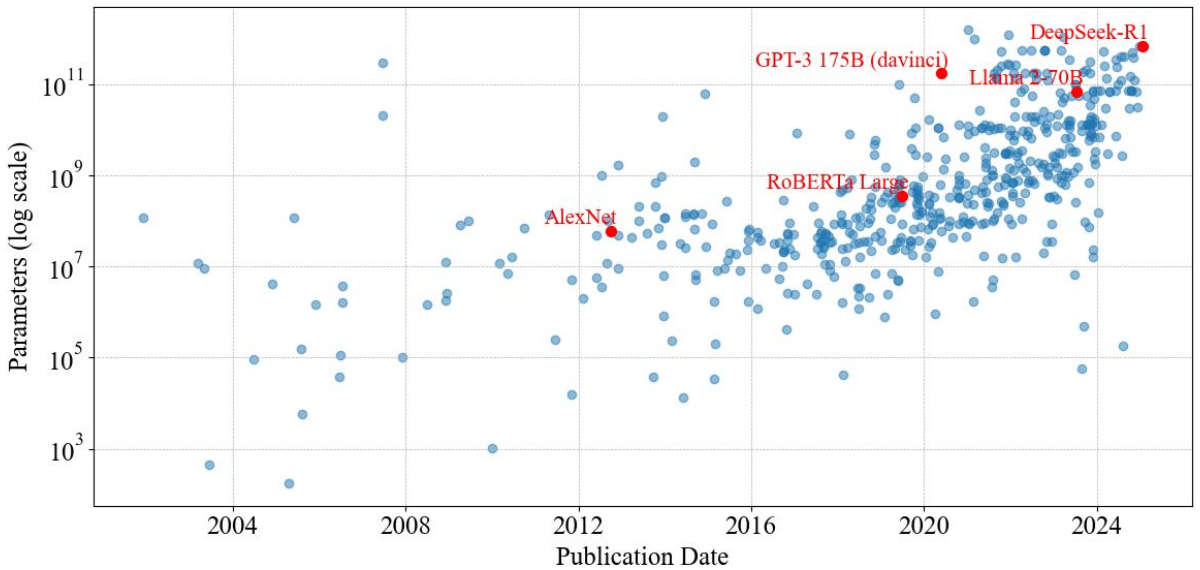
Artificial Intelligence (AI)

AI is present in many applications, e.g. medicine, robotics.

Artificial Intelligence (AI)

AI is present in many applications, e.g. medicine, robotics.

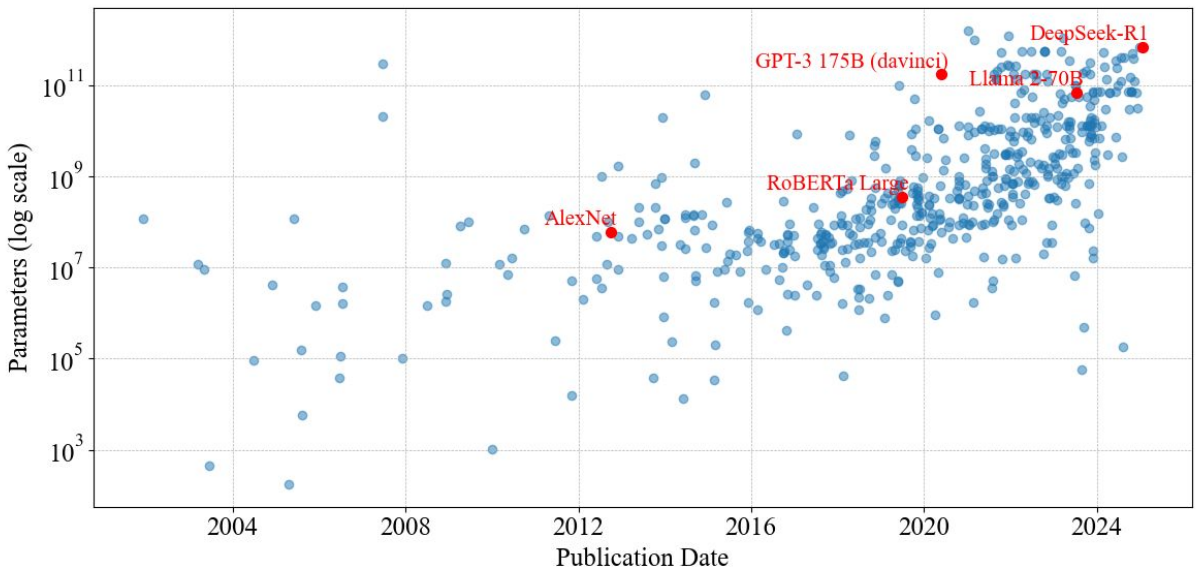
Models present an exponential growth of computational capacity



Artificial Intelligence (AI)

AI is present in many applications, e.g. medicine, robotics.

Models present an exponential growth of computational capacity



Given the scaling computing and climate changes, we need urgent solutions.

Artificial Intelligence (AI)

Possible solution: selection of energy-efficient models
(Asperti et. al. 2021, Yu et. al. 2022)

Artificial Intelligence (AI)

Possible solution: selection of energy-efficient models
(Asperti et. al. 2021, Yu et. al. 2022)

Two main paradigms:

Bigger-is-Better:
We use the best hardware for running the best (and large) model

Small-Is-Sufficient:
We select smaller models with a similar performance to state-of-art models

Artificial Intelligence (AI)

Possible solution: selection of energy-efficient models
(Asperti et. al. 2021, Yu et. al. 2022)

Two main paradigms:

Bigger-is-Better:
We use the best hardware for running the best (and large) model

Small-Is-Sufficient:
We select smaller models with a similar performance to state-of-art models

Question:

How much energy we can save by applying model selection in AI?

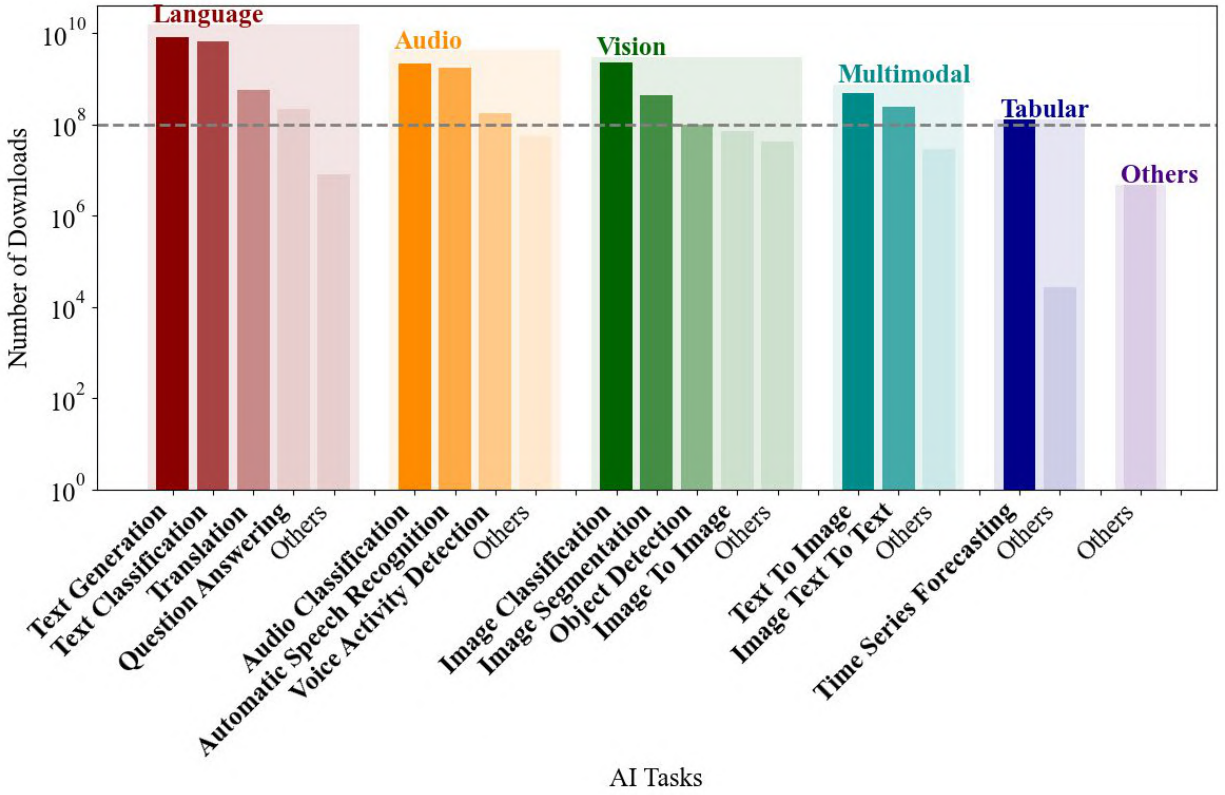
What are the most frequent AI tasks in data centres?

What are the most frequent AI tasks in data centres?

Literature
Papers With Code: Platform with papers with code implementation
Hugging Face: platform for deploying open-source models

What are the most frequent AI tasks in data centres?

Hugging Face: platform for deploying open-source models



What are the most frequent AI tasks in data centres?

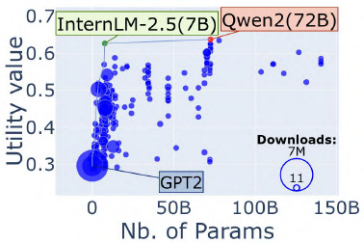
Analyze 14 AI tasks

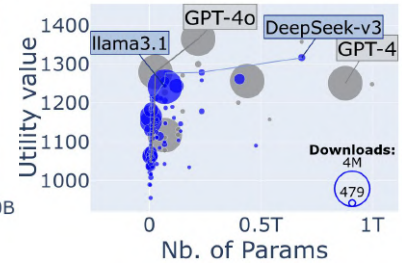
Text generation
Text classification
Translation
Image classification
Object detection
Semantic segmentation
Image generation
Text Clustering
Image-Text to Text
Mathematical reasoning
Code generation
Time series forecasting
Speech recognition
Audio classification

For each task, we select a benchmark, which evaluates and compares the models

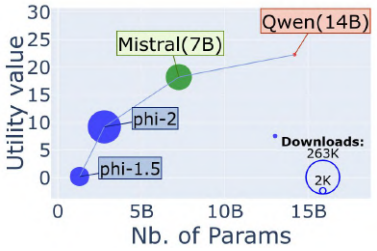
Analysis of AI benchmarks to evaluate the sobriety potential

Text Generation

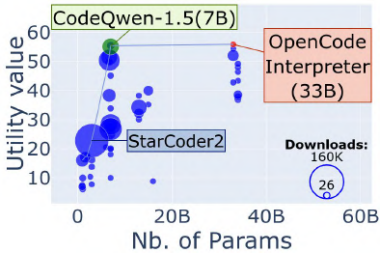




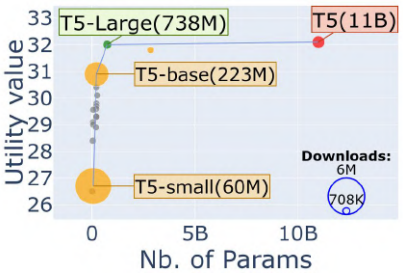
Math Reasoning



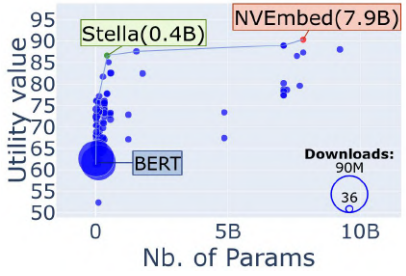
Code Generation



Translation



Text Classification



Text Clustering

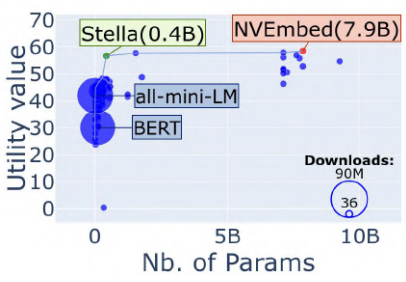
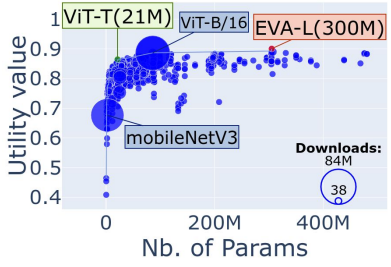


Image Classification



Object Detection

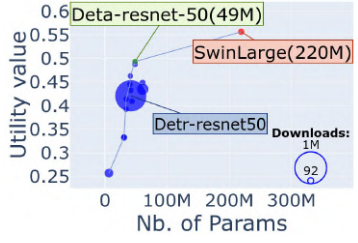
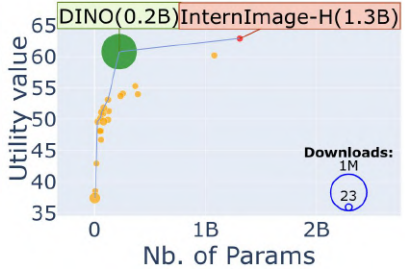
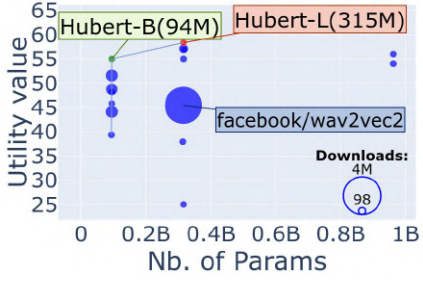


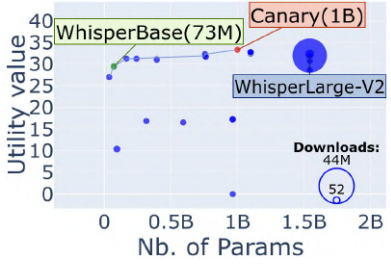
Image Segmentation



Audio Classification



Speech Recognition



Text to Image

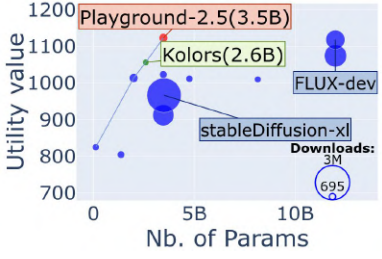
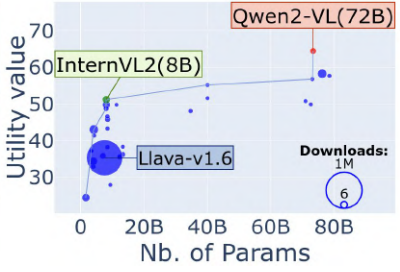
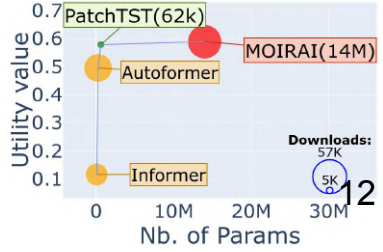


Image-Text to Text

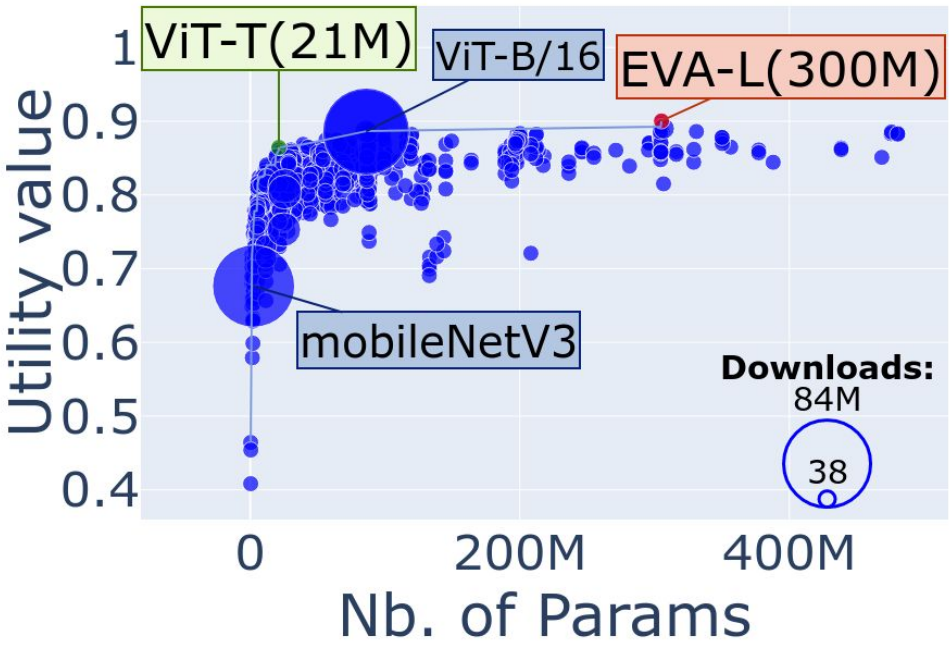


Time series Forecasting



Analysis of AI benchmarks to evaluate the sobriety potential

Image Classification



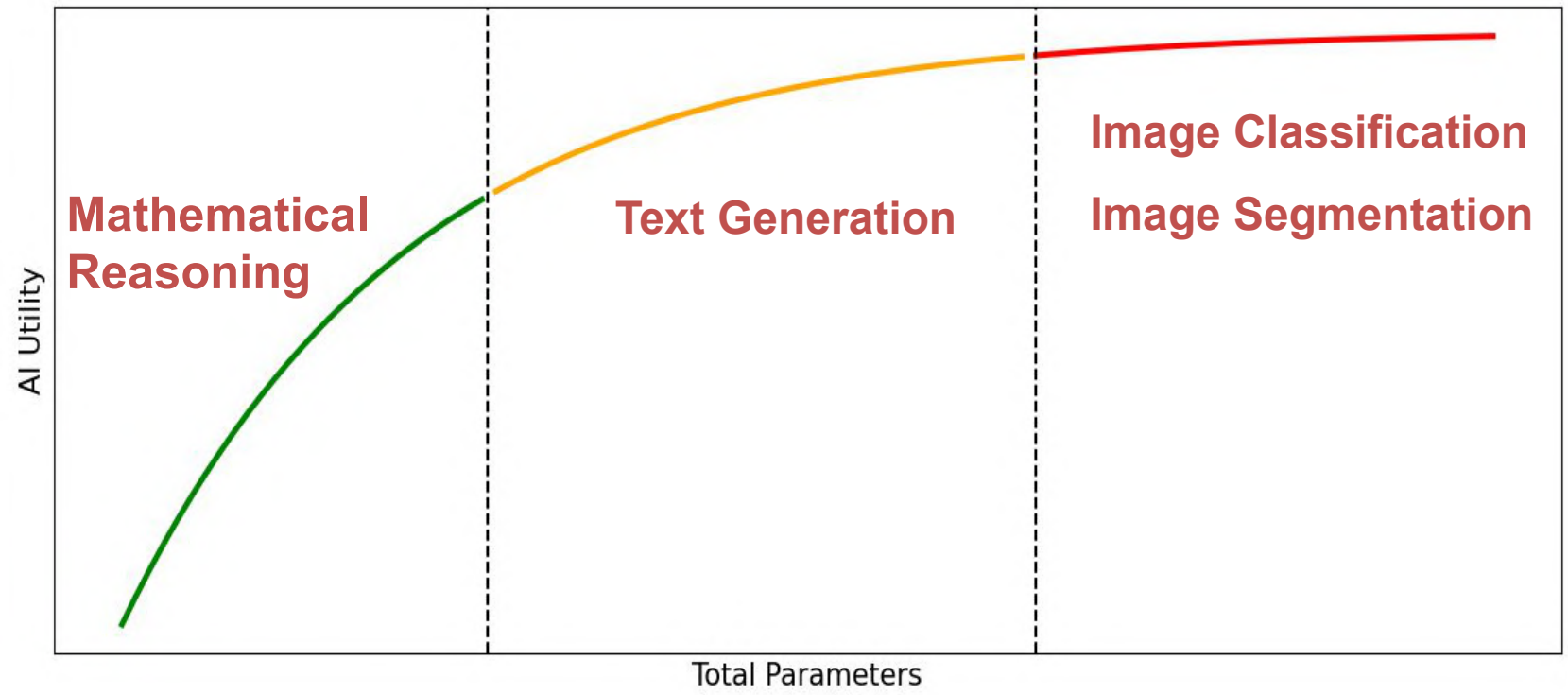
Energy-efficient: ViT-T (Google)

Best-performing: EVA-L (BAAI)

Largely adopted models: ViT/B, MobileNetV3 (Google)

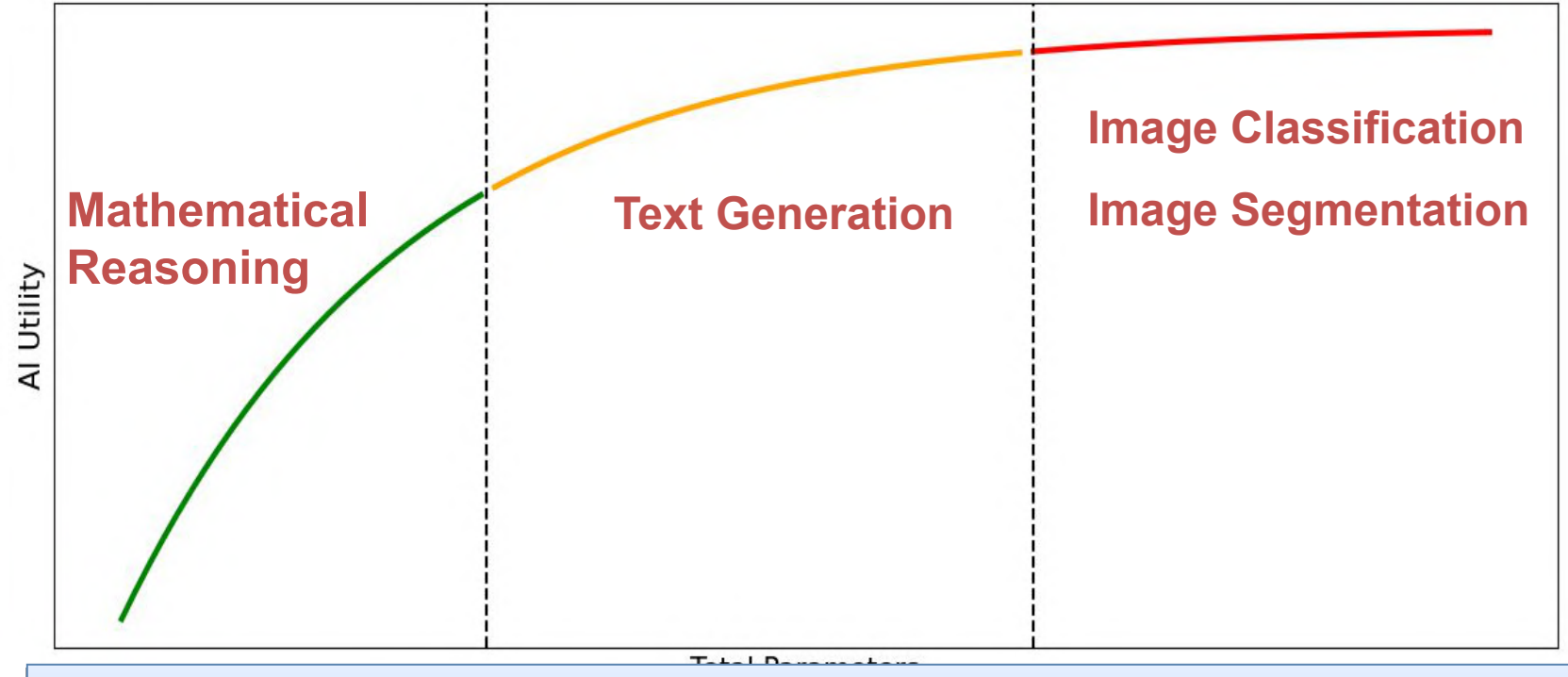
Analysis of AI benchmarks to evaluate the sobriety potential

Tasks at different maturity levels



Analysis of AI benchmarks to evaluate the sobriety potential

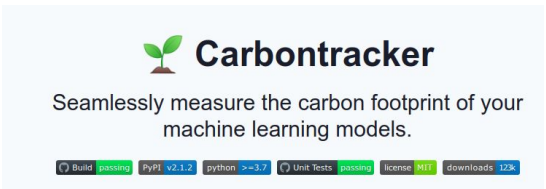
Tasks at different maturity levels



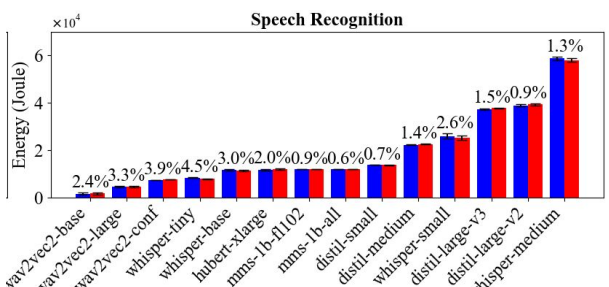
How much energy we can save by selecting energy-efficient model?

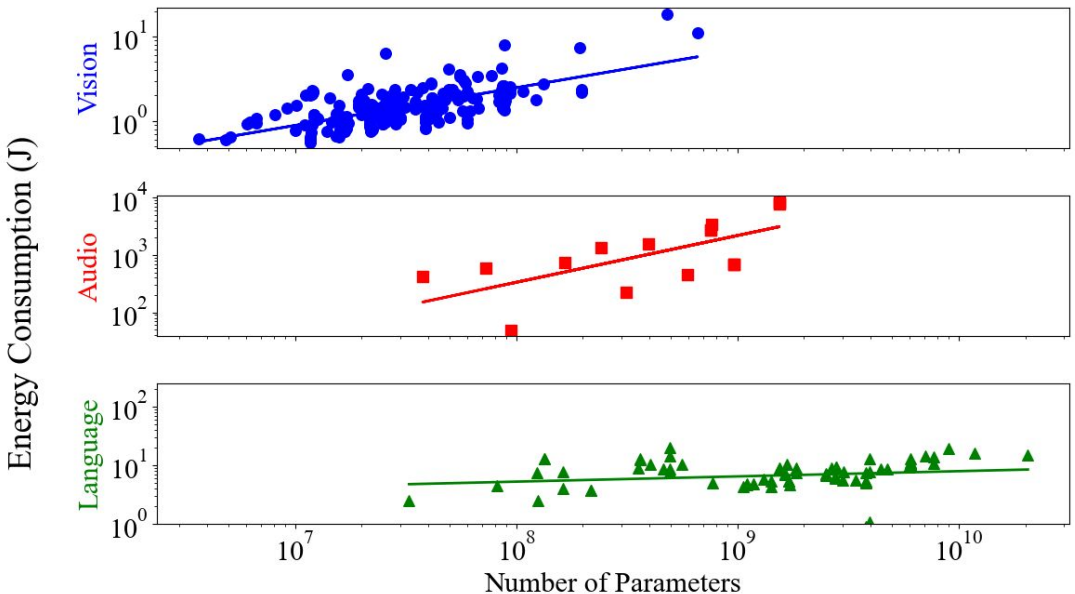
Methodology

Since there are several and large models, it is impractical to measure the energy consumption of all models



Measured energy consumption of key models using *Carbon Tracker* Tool and power meter measurements





Regression method for estimating energy based on the number of parameters

Estimating energy savings

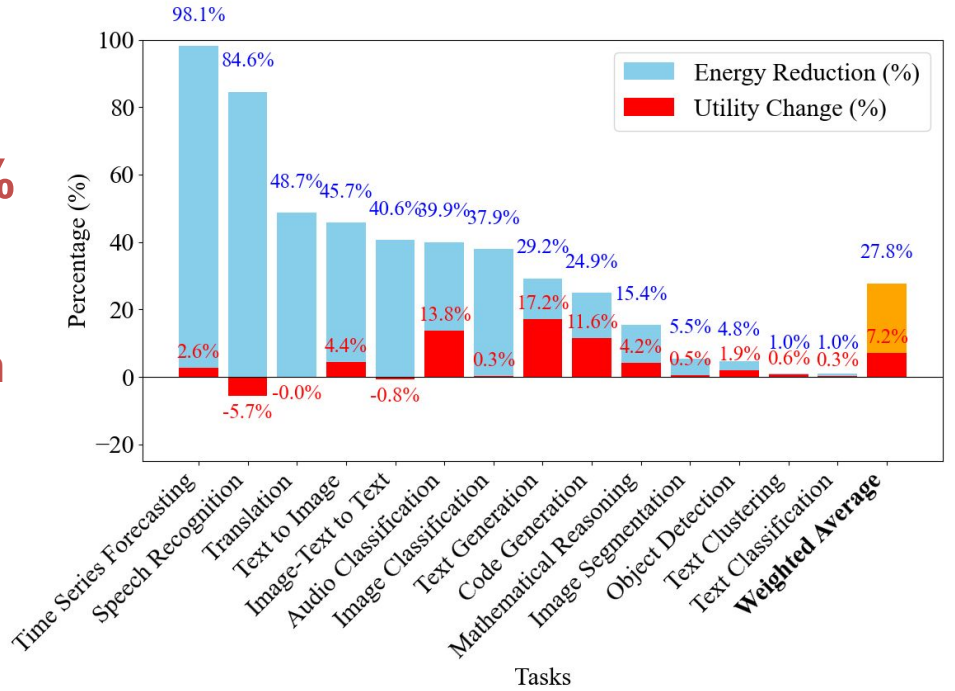
Redirecting inference requests from **models larger than the energy-efficient** to the **energy-efficient** model

Estimating energy savings

Redirecting inference requests from **models larger than the energy-efficient** to the **energy-efficient** model

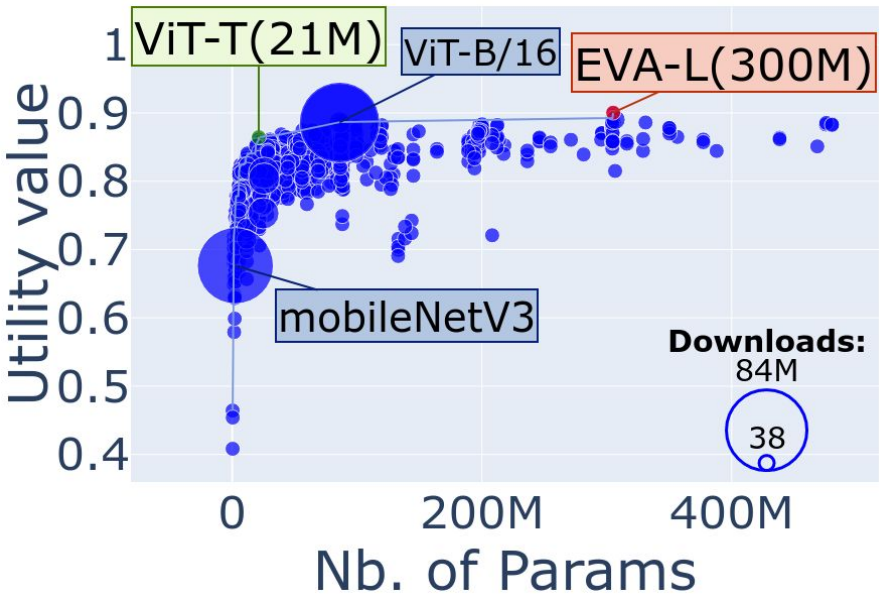
Energy reduction ranges from 1% to 98% for different AI tasks

Influenced by task maturity and model adoption



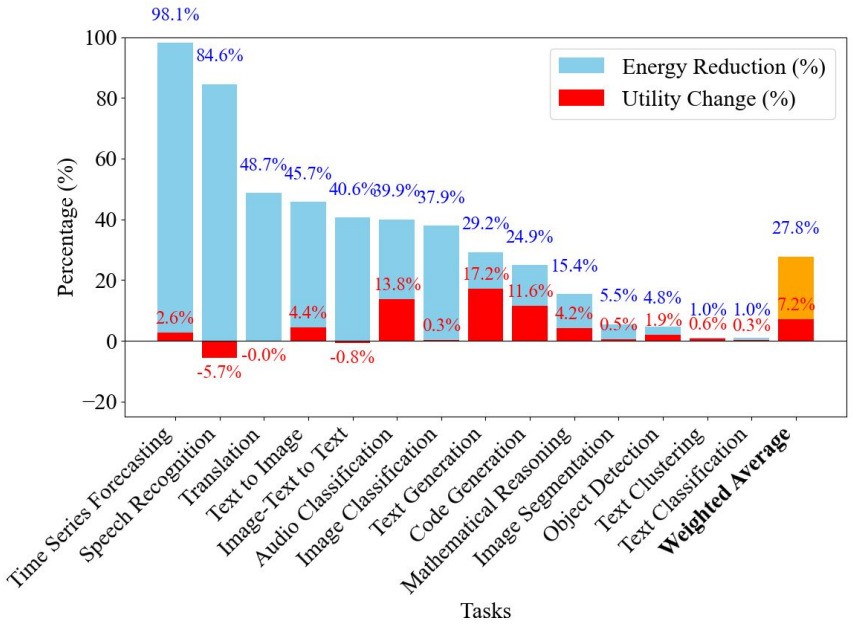
Analysis of AI benchmarks to evaluate the sobriety potential

Image Classification



Energy Savings: 37.9%

Utility variation: 0.3%

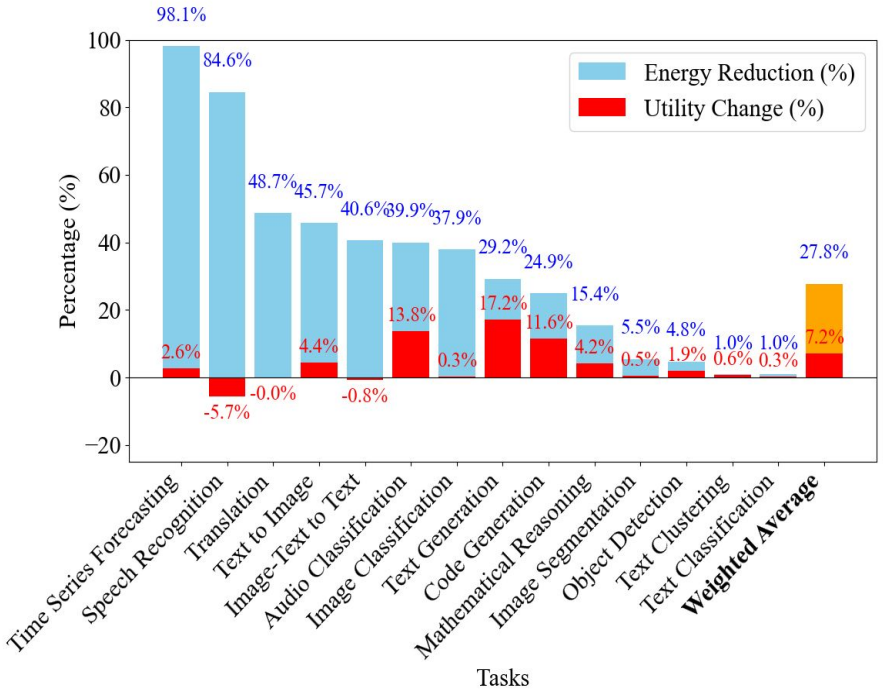


Estimating energy savings

Redirecting inference requests from **models larger than the energy-efficient** to the **energy-efficient** model

Energy reduction ranges from 1% to 98% for different AI tasks

Influenced by task maturity and model adoption

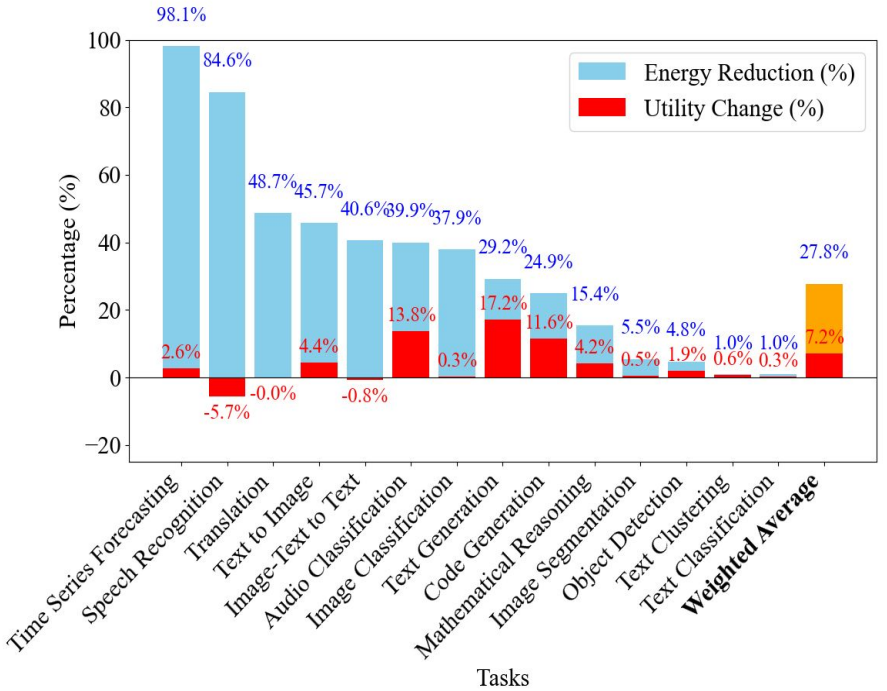


Estimating energy savings

Redirecting inference requests from **models larger than the energy-efficient** to the **energy-efficient** model

Energy reduction ranges from 1% to 98% for different AI tasks

Influenced by task maturity and model adoption



We estimate saving 27.8% of energy consumption with model selection

Conclusion

We investigated how model selection (energy-efficient models) can contribute to energy sobriety
- Present a small size
- Maintain a high utility

We estimated energy consumption for different scenarios:

Bigger-is-Better: Increase of 112% - 128.74 TWh (2025)
Small-is-Sufficient: Savings of 27.8% - 31.9 TWh (2025), equivalent to 5 nuclear power plants production

Future Work: Expand to the entire AI life cycle analysis, from HW manufacturing to deployment

Thank you!