

AI FOR GOOD GLOBAL SUMMIT 2026 · GENEVA

Session: Designing Sustainable AI: Models, Infrastructure and Lifecycle

Designing Sustainable AI

Our approach to reducing impact across the life of a system

The biggest footprint is rarely where you think

Prof Ivana Drobnyak

University College London · Member of the UNESCO Chair in AI

Research with Maria Pérez-Ortiz and Zekun Wu (UCL) and Leona Verdadero (UNESCO)

AI's footprint is physical, and measurable

Across its life an AI system has a footprint in:

- greenhouse gases (gCO₂e)
- water, for cooling and power (litres)
- minerals and metals in the hardware
- electronic waste at end-of-life

Climate is measured best; water and nature, far less so.

MACRO

~415 → 945 TWh

Global data-centre electricity, 2024 → 2030 (all data centres; AI is the fastest-growing part).

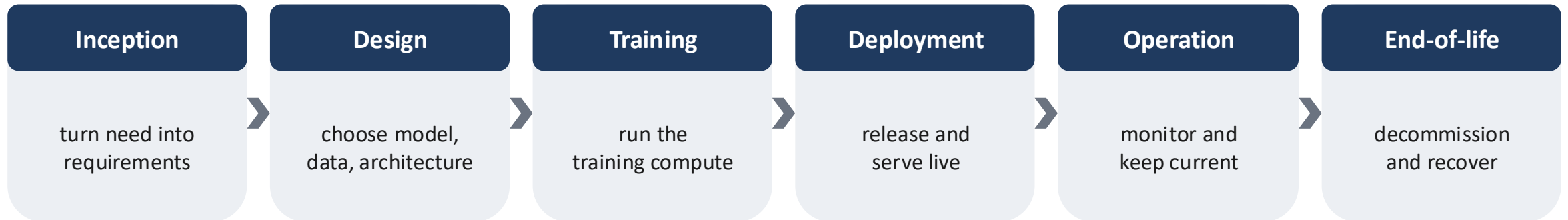
MICRO

~0.24 Wh / query

A median chatbot prompt: tiny on its own (~0.26 mL water), but multiplied by billions.

Every AI system has a life cycle

Impact is not one number at the launch. It accrues at every stage, and each stage leaves a different trace.



Training and inference are not stages but processes: training is concentrated early; inference runs for as long as the system is used.

At each stage we ask: what impact is locked in here, and what can still change it?

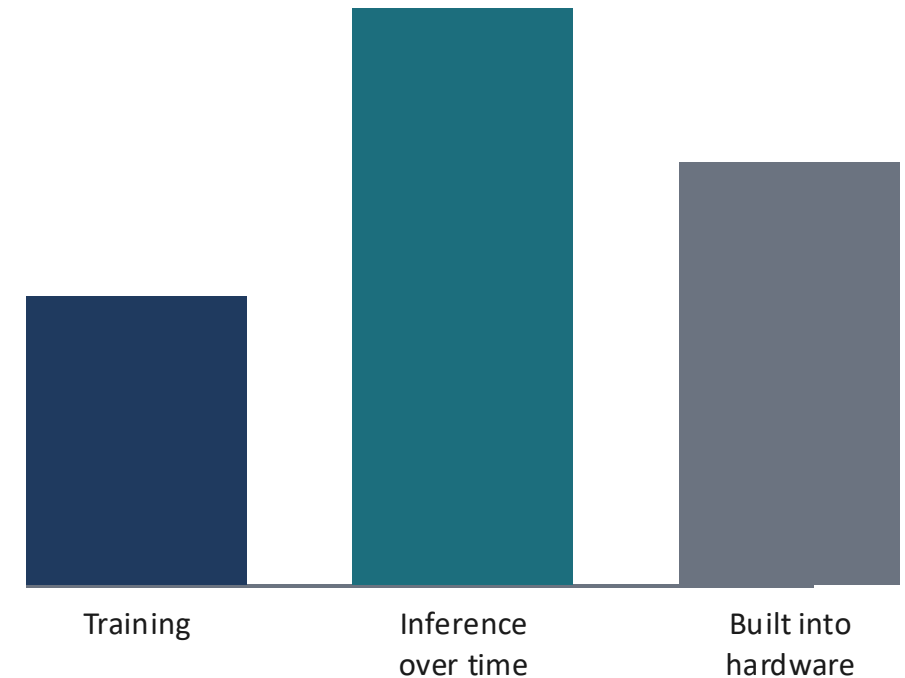
The training headline misses the story

Training makes the headlines. The whole life tells a fuller story:

- inference can outweigh training over a system's life
- the hardware itself can be a tenth to a third of lifecycle emissions
- cooling, power and water add up every hour it runs

Optimise the life, not just the launch.

Illustrative, not to scale

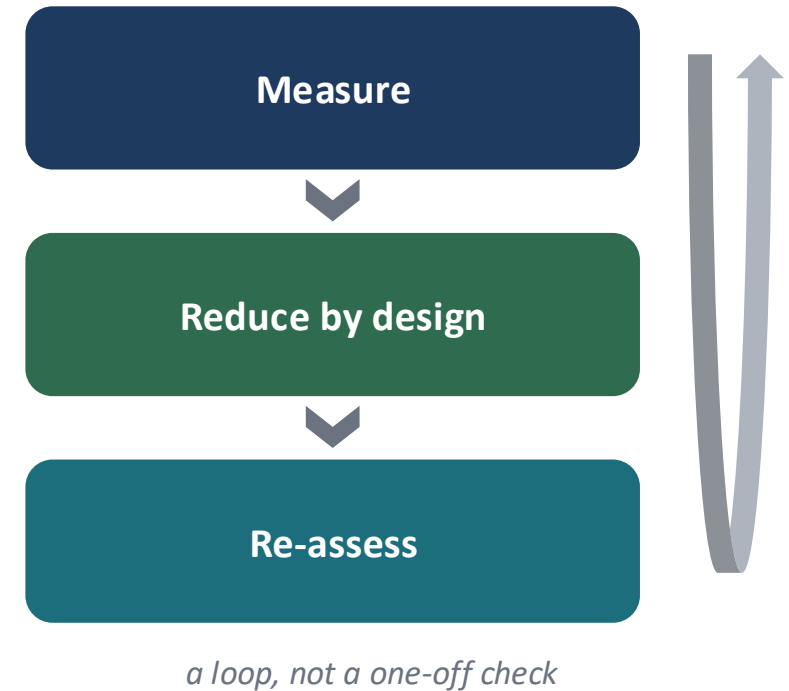


Mitigation by design: measure, reduce, re-assess

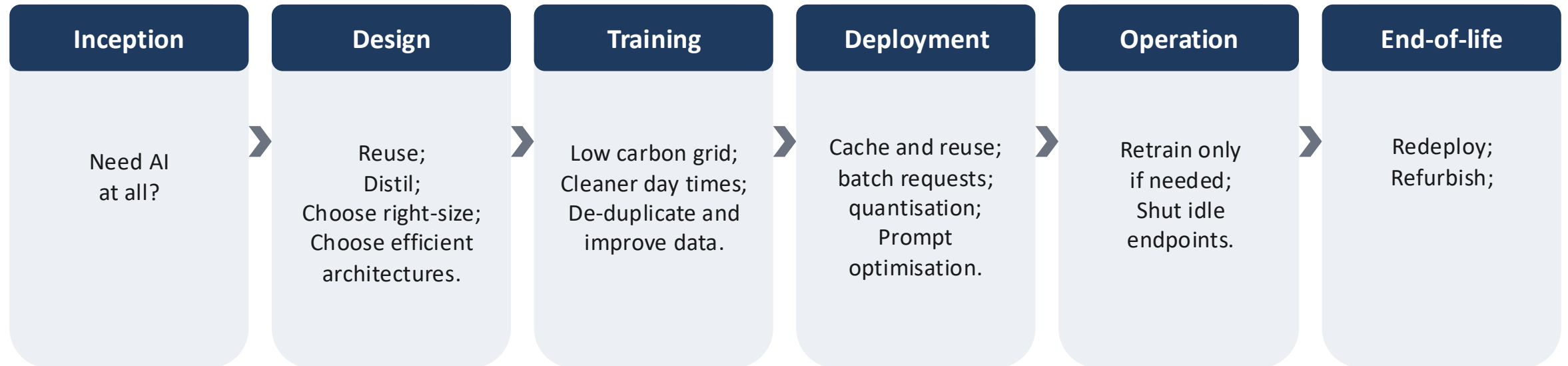
How we at UCL and the UNESCO Chair approach it, three moves:

1. Measure the impact across the life cycle with a life-cycle assessment (functional unit, system boundary, baseline), per ITU-T L.1801.
2. Reduce it by design, before the system is built, where the leverage is largest.
3. Re-assess to confirm the saving is real, and screen for rebound (per ITU-T L.1480).

The cheapest impact is the one you design out.



Actions at each stage of the life cycle

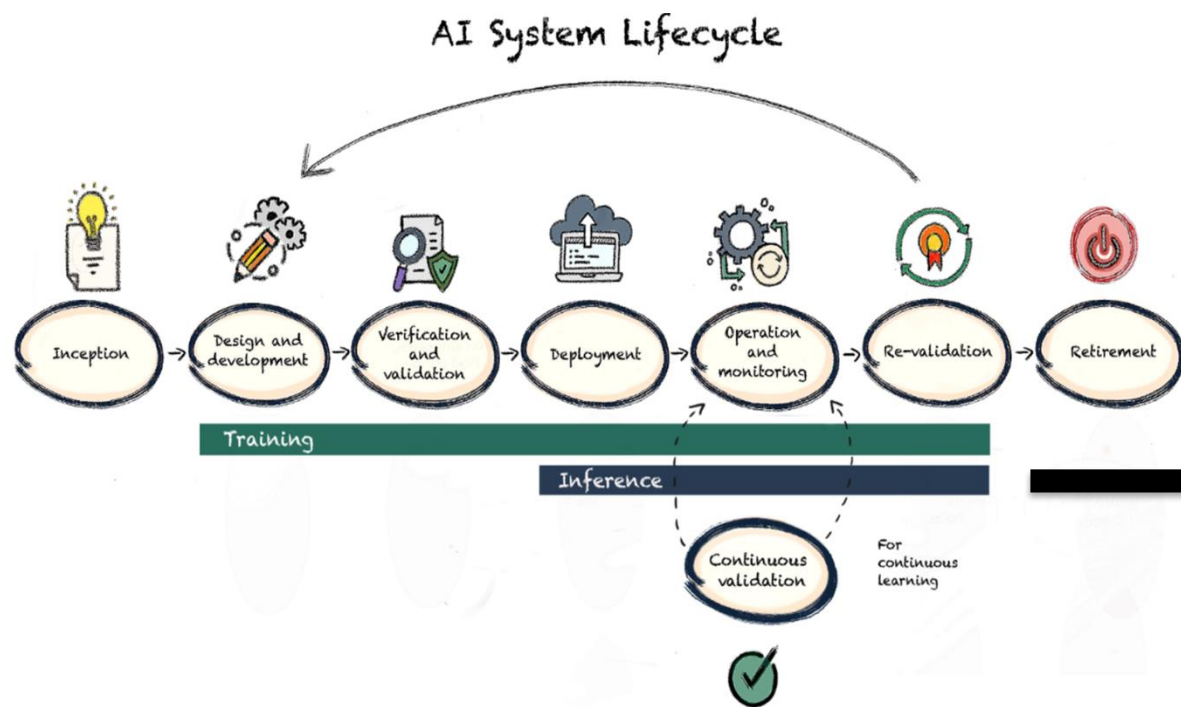


Concrete levers, not one fix: distillation, quantization, caching, carbon-aware siting, hardware reuse. Each is cheaper and more effective the earlier it is taken.

Smarter, Smaller, Stronger:

Resource-Efficient Generative AI &

the Future of Digital Transformation

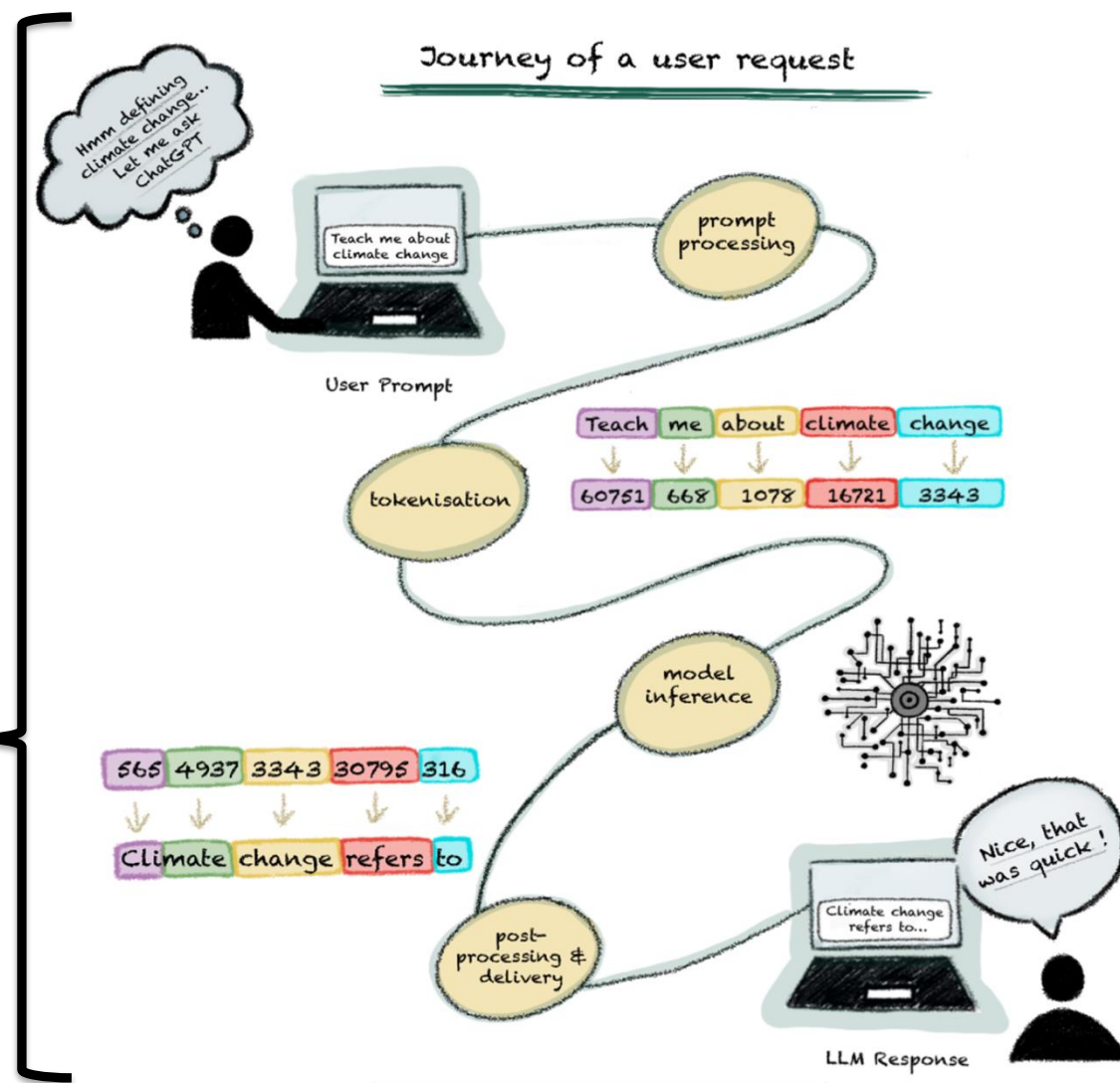


Authored by: Leona Verdadero^{1*}, Ivana Drobnyak^{2*}, Hristijan Bosilkovski^{2*}, Zekun Wu²³, Emma Fischer, and María Pérez-Ortiz².

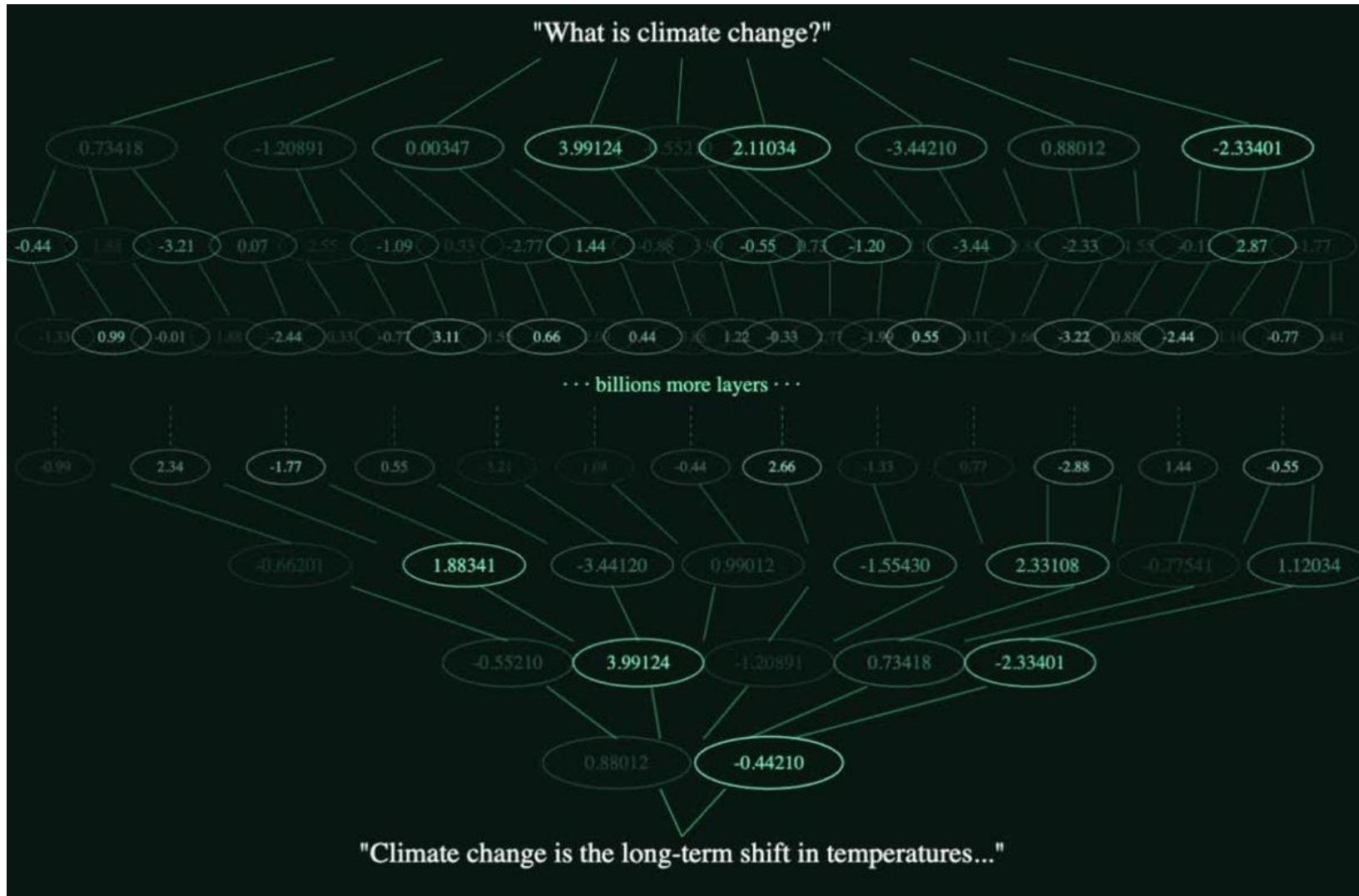
¹ UNESCO

² University College London

³ Holistic AI



AI Brain in action



- Large LLMs: trillions of nodes
- Trillions of calculations
- Each click $\sim 0.24\text{Wh}$
- ~ 4 Billion clicks/day
- $\sim 1\text{GWh}$ electricity
- $\sim 140,000$ UK Houses / day

How can we save an AI brain's thinking energy?

2.088990 X 3.06492725 X 1.000556 =

How can we save an AI brain's thinking energy?

**2.088990 X 3.06492725 X 1.000556 = 6.406162 HIGH
ENERGY COST**

How can we save an AI brain's thinking energy?

**2.088990 X 3.06492725 X 1.000556 = 6.406162 HIGH
ENERGY COST**

Round the numbers

How can we save an AI brain's thinking energy?

**2.088990 X 3.06492725 X 1.000556 = 6.406162 HIGH
ENERGY COST**

2 X 3 X 1 = 6 LOW ENERGY COST

How can we save an AI brain's thinking energy?

2.088990 X 3.06492725 X 1.000556 = 6.406162 HIGH ENERGY COST

2 X 3 X 1 = 6 LOW ENERGY COST

6.406162 $\neq$ 6. We lost some accuracy. Is that a problem?

Depends: Are we measuring curry powder or ingredients for nuclear fusion?

SOMETIMES WE NEED 6

SOMETIMES WE NEED 6.4

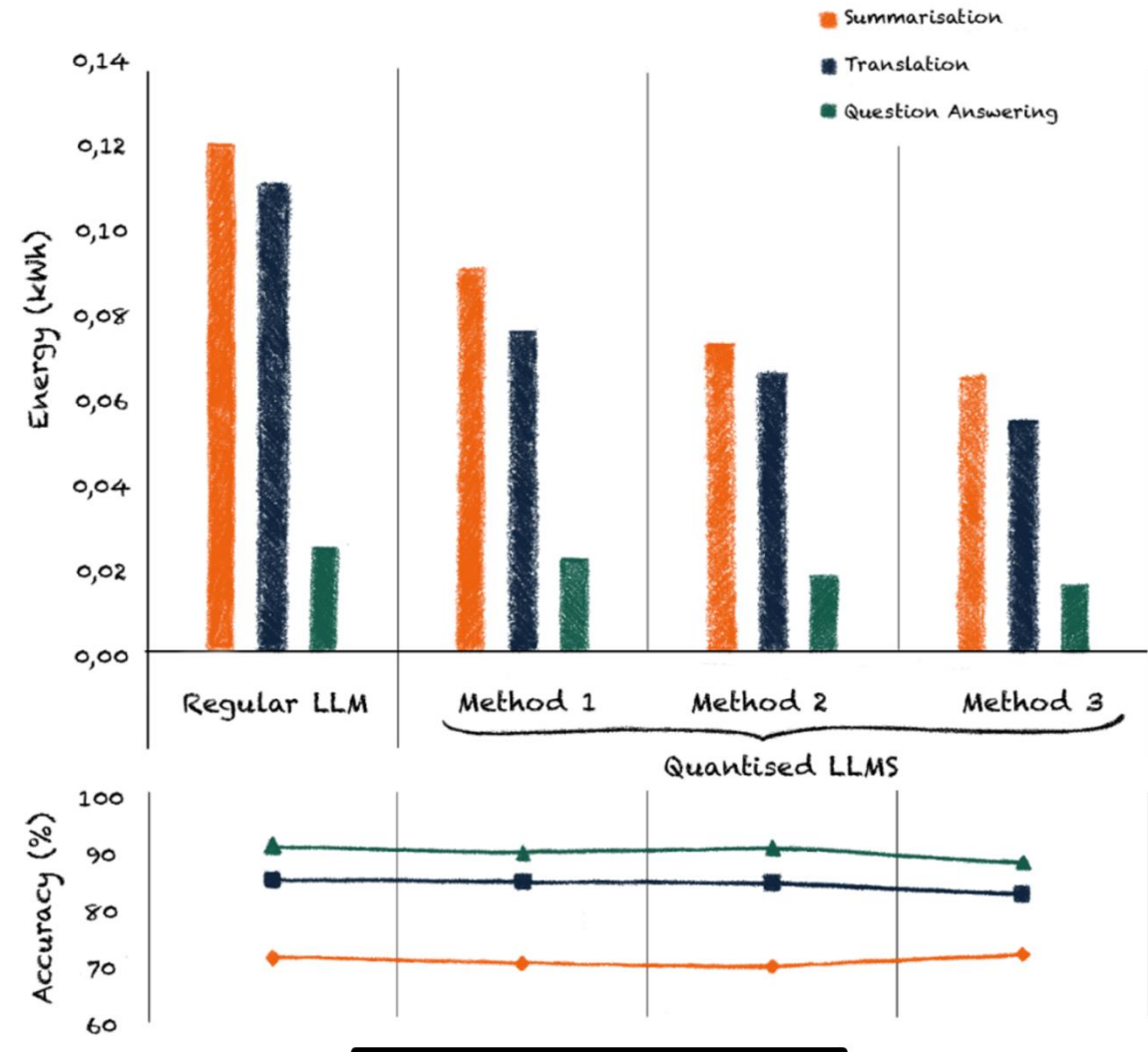
SOMETIMES WE NEED 6.406162

WE CLEVERLY OPTIMISE THE PROCESS, AND WE DON'T LOOSE ANYTHING

Quantisation

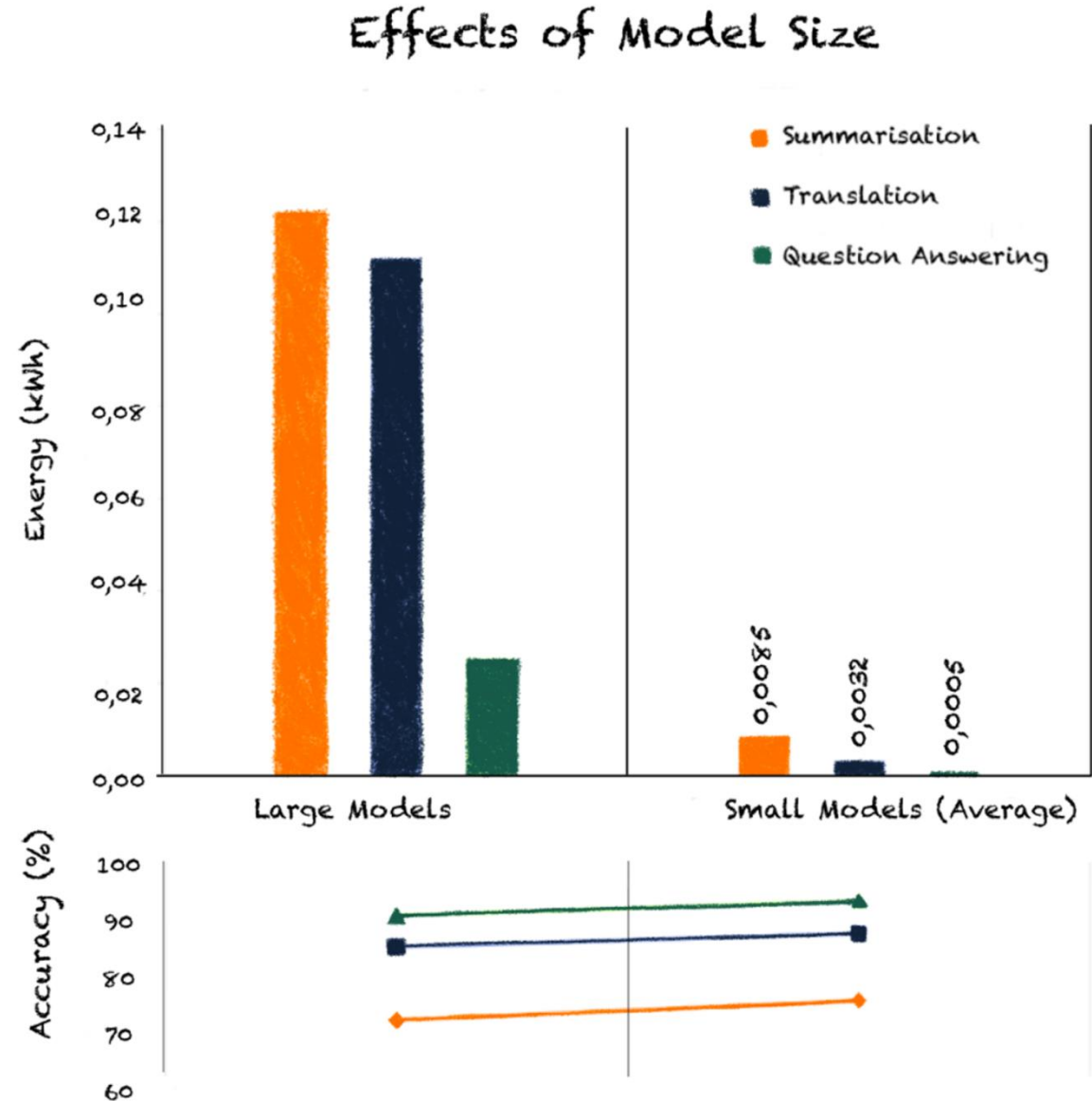
- BNBQ: Achieved a 22% reduction in energy consumption.
- GPTQ: Led to a 35% reduction in energy consumption.
- AWQ: Delivered a 44% reduction in energy consumption and even outperformed the unquantized model on certain tasks.

Effects of Quantisation



Distilling

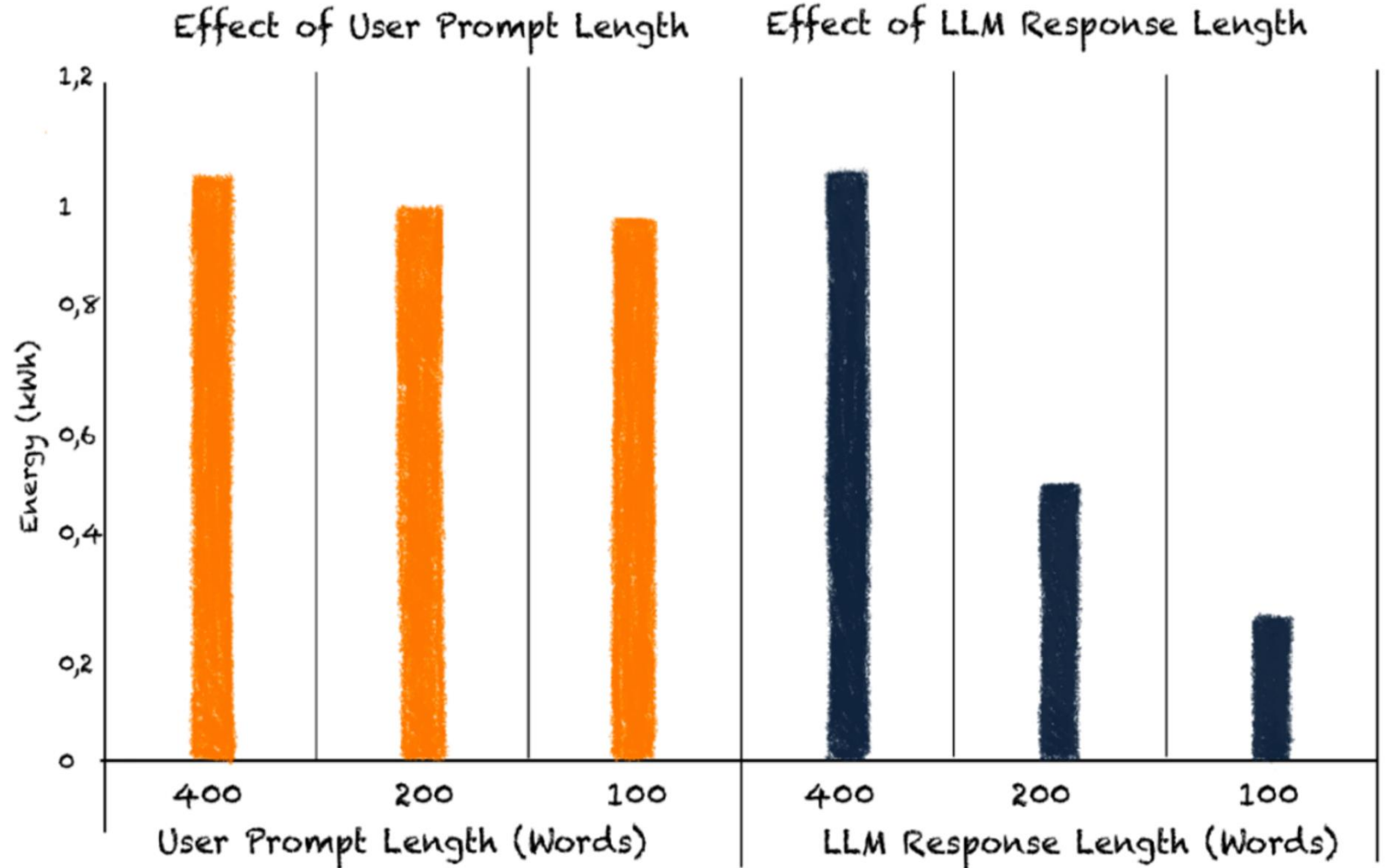
- Removing nodes when possible
- Using small models for specialised tasks



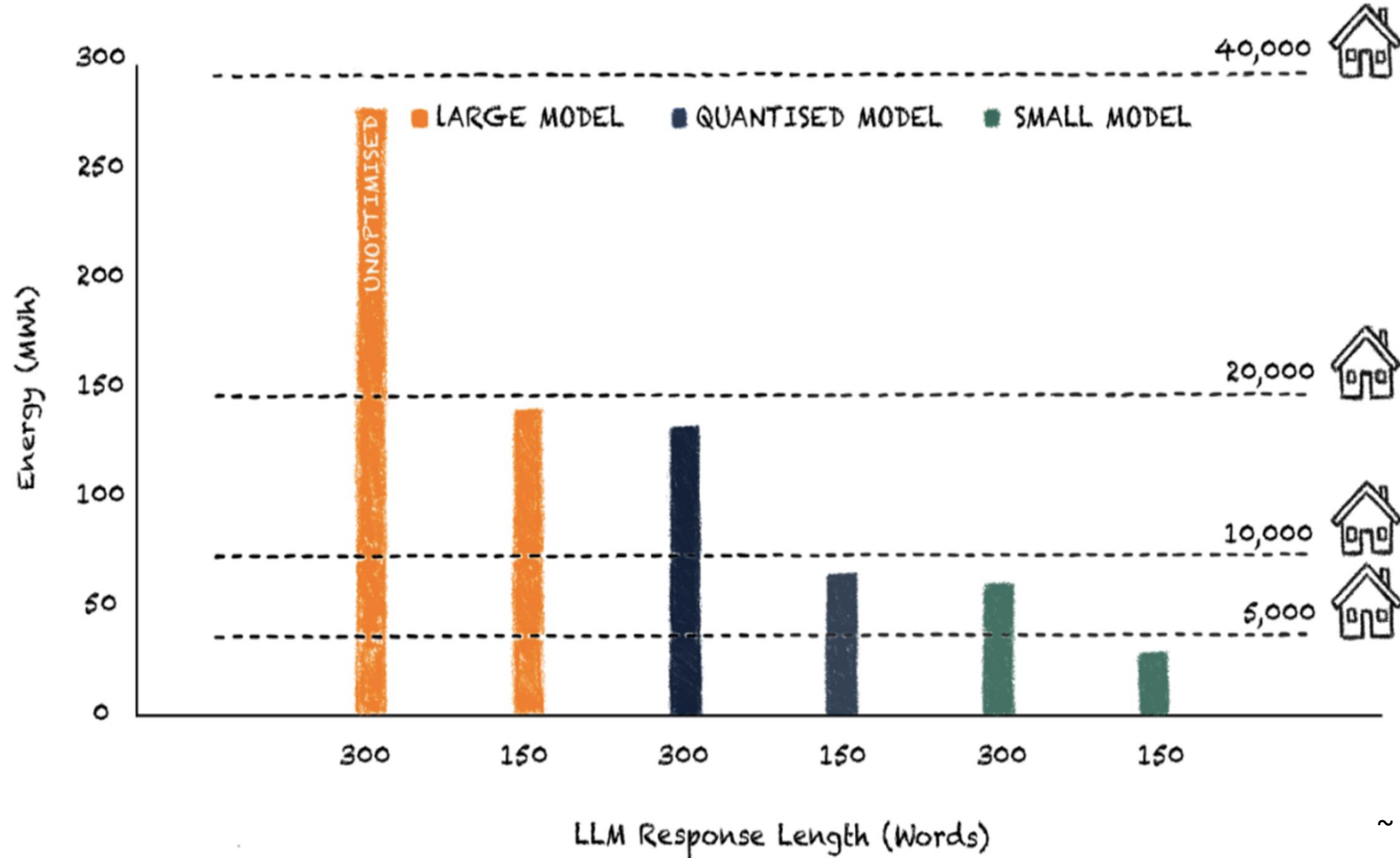
Prompt optimisation

- Reducing prompt text
- Asking the model to cap the answer size

Be Brief



DAILY IMPACT OF OPTIMISATION ON ENERGY SAVING



From our approach to shared guidance

- Design choices can save energy at each stage of an AI lifecycle
- We are turning our approach into shared guidance through an open ITU-T SG5 (Q6/5) standards document [work ongoing – expecting to finish in 2027]
- The new standard will be a counterpart to the recently released ITU-T L.1801 assessment on how to measure the energy footprint of AI.



Here at the AI for Good Summit

Resilient AI Challenge



Awards for best compression methods to the best teams:

Wednesday July 8
Room V from 12:30 to 1pm.

<https://www.sustainableaicoalition.org/resilient-ai-challenge/>

Resource material, educational videos etc.

Sustainable AI is a design choice

Measure it, reduce it by design across the whole life of a system, and re-check the saving is real.

The cheapest impact is the one you design out.

Prof Ivana Drobnyak · UCL · Member of the UNESCO Chair in AI

Thank you.