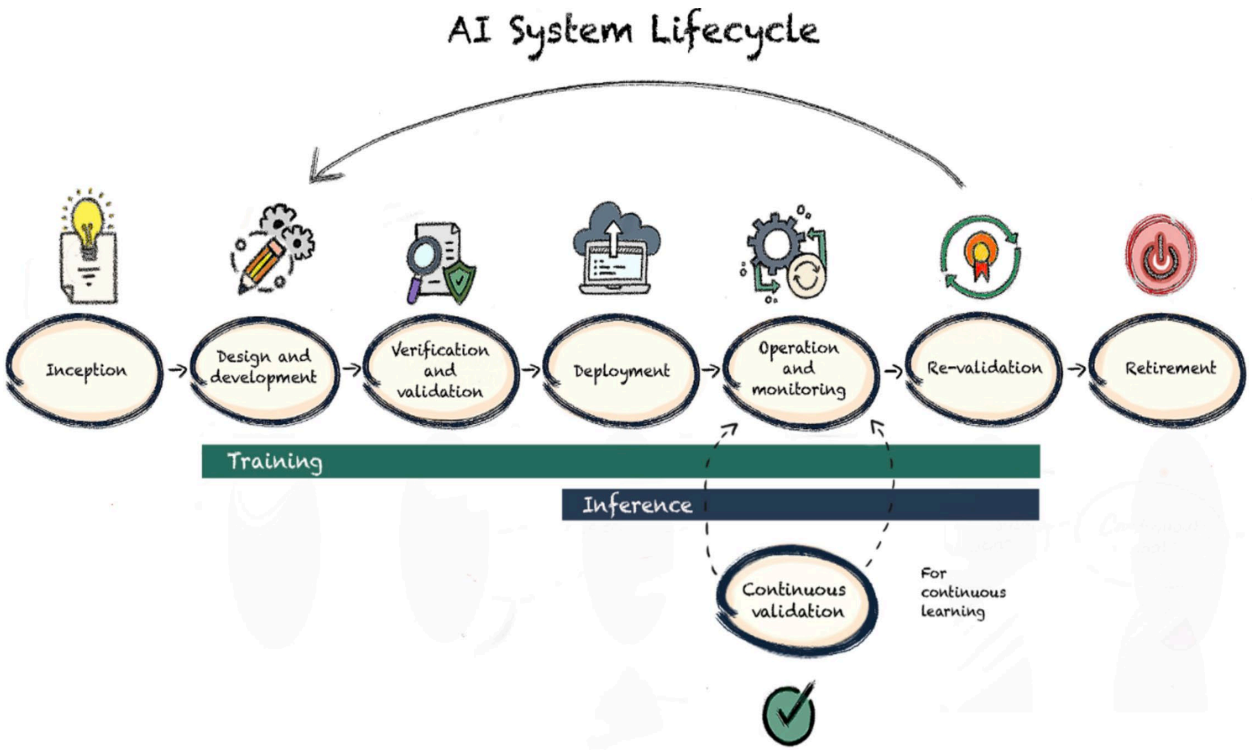
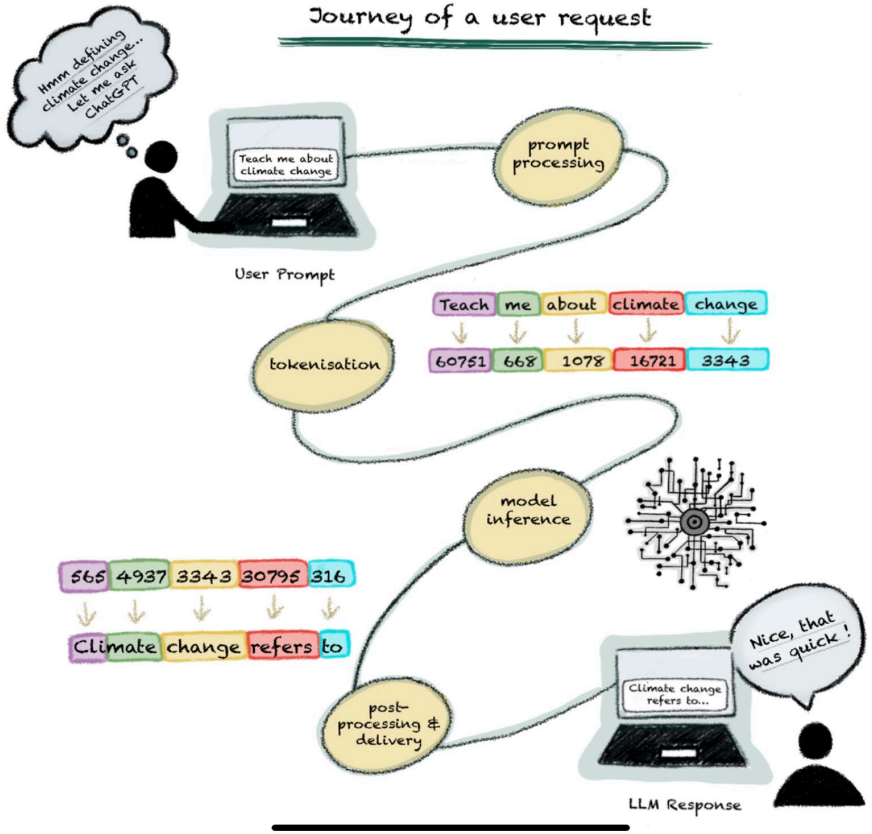
Smarter, Smaller, Stronger: Resource-Efficient Generative AI & the Future of Digital Transformation

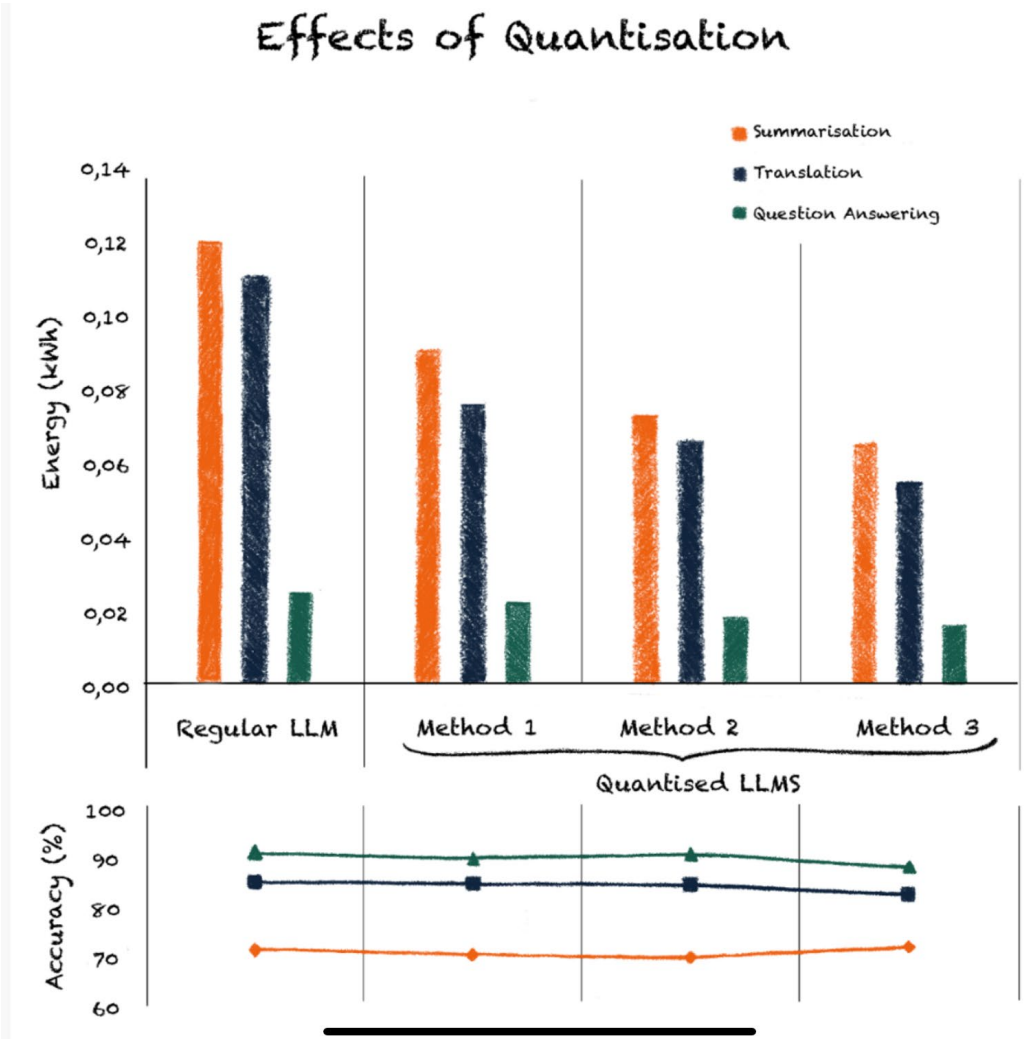
Authored by: Leona Verdadero[1] *, Ivana Drobnjak[2] *, Hristijan Bosilkovski[2]*, Zekun Wu[23], Emma Fischer, and María Pérez-Ortiz[2].

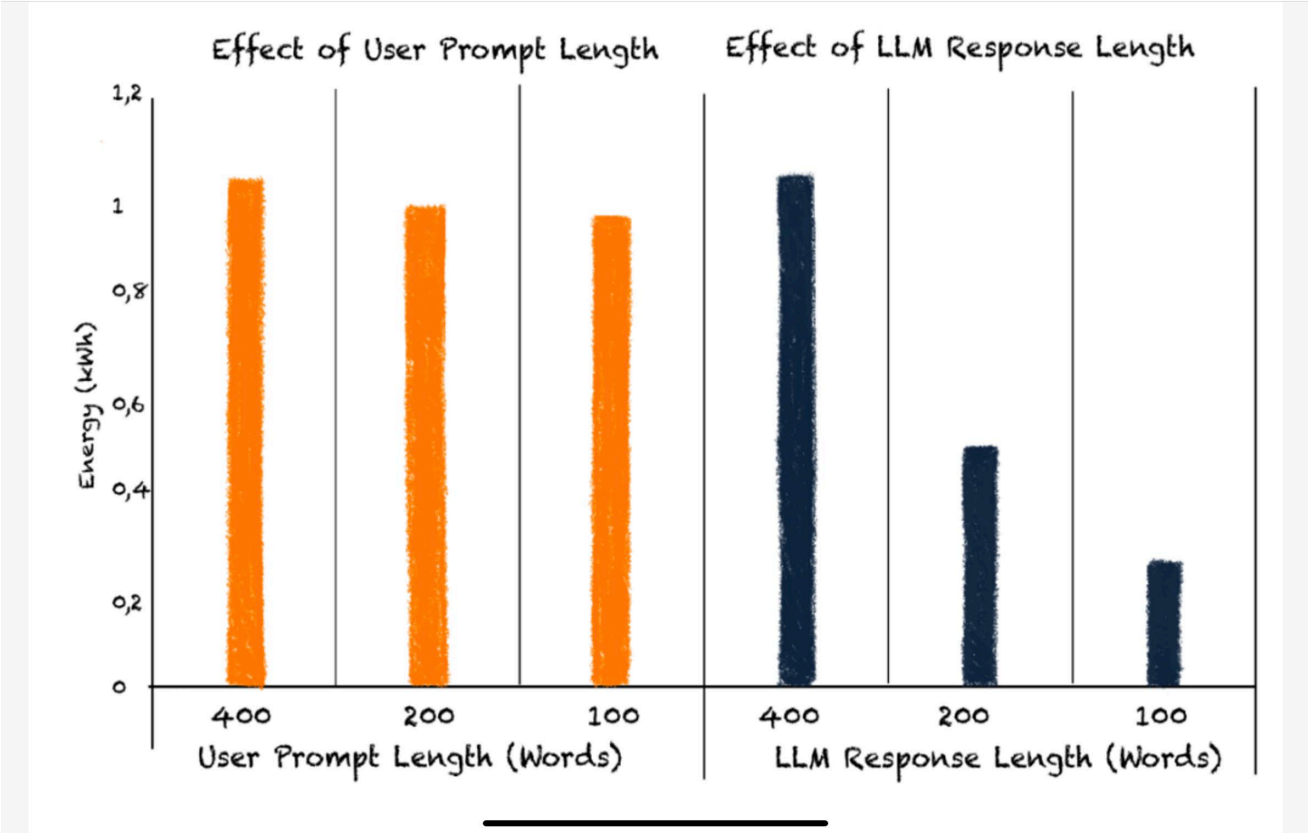
[1] UNESCO
[2] University College London
[3] Holistic AI





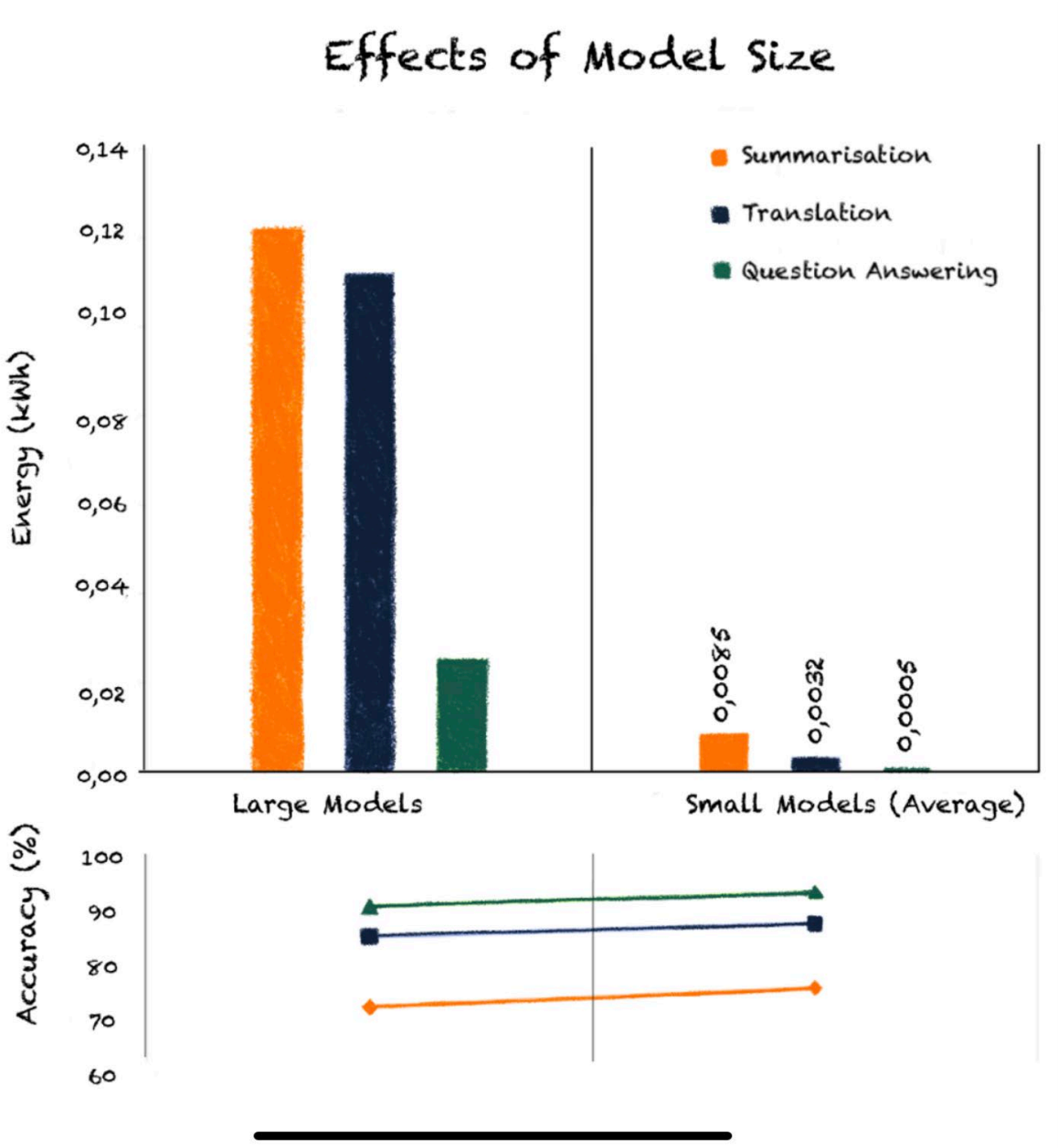
General models: 75% energy reduction

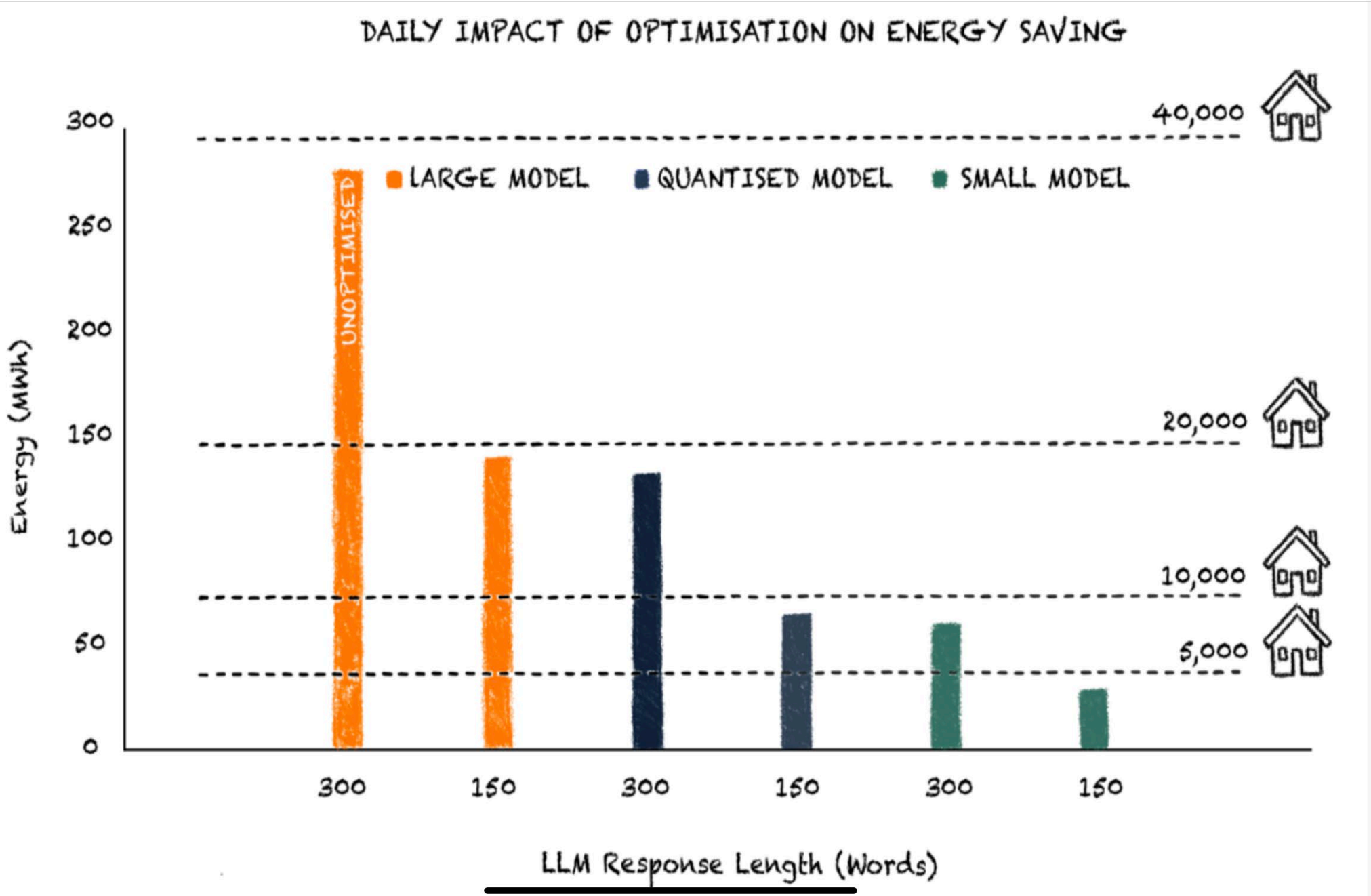




For specialised tasks: Small models

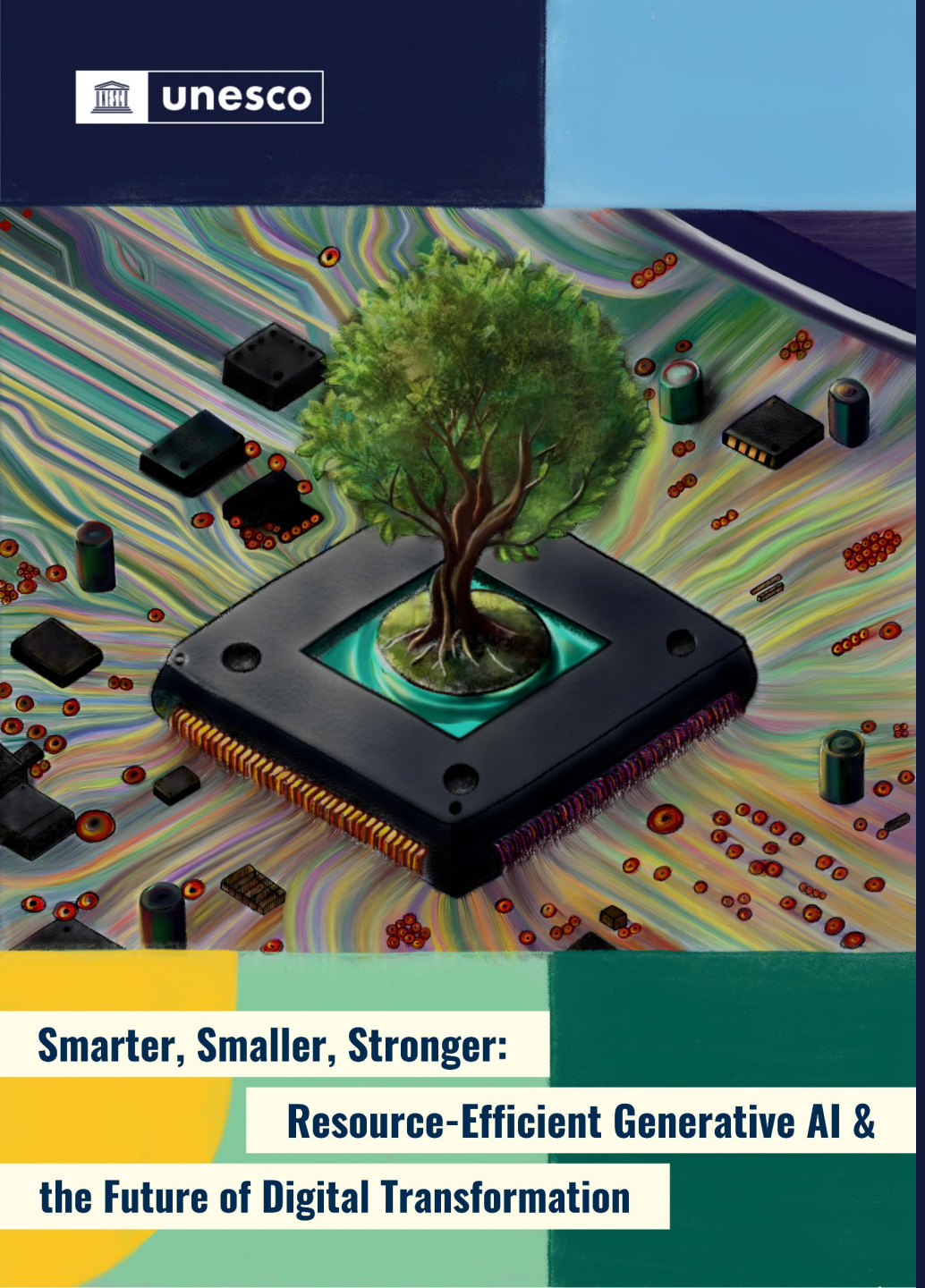
over 90% energy reduction




DAILY IMPACT OF OPTIMISATION ON ENERGY SAVING
Energy (MWh)
300
250
200
150
100
50
0
UNOPTIMISED
LARGE MODEL
QUANTISED MODEL
SMALL MODEL
40,000
20,000
10,000
5,000
300
150
300
150
300
150
LLM Response Length (Words)

The future

- Optimising large models (e.g. quantisation etc.)
- Using small models where appropriate
- More efficient large model architectures (work in progress):
  - Mixture of Experts
  - Multi Agent
  - Sparse and conditional computation
  - Retrieval-augmented generation
  - Neurosymbolic and brain inspired architectures
- More work needs to be done – designing with efficiency in mind



Smarter, Smaller, Stronger: Resource-Efficient Generative AI & the Future of Digital Transformation