Sustainable and Right Sized Intelligence

Pabitra Mitra
Professor, Computer Science and Engineering
IIT Kharagpur
pabitra@cse.iitkgp.ac.in

Content

- Sustainable Intelligence
- Environmental Impact of AI
- Ethical AI
- Responsible AI
- Sustainable AI – path forward
- Preserving Ecosystems
- Right Sized Intelligence in Practice


8:07
Good Morning
“Life is a journey, not a destination. So don’t let the bumps in the road keep you from enjoying the ride.”
FAVORITE
SAVE
SHARE
SUGGEST MESSAGE FOR THIS IMAGE

Costs of AI

Maine Woods Editing

Writing one 100-word email with AI =

14 hours

16.9 oz.

If 10% of working Americans do that once/week for a year =

20 days for all of D.C.

1.5 days for all of Rhode Island

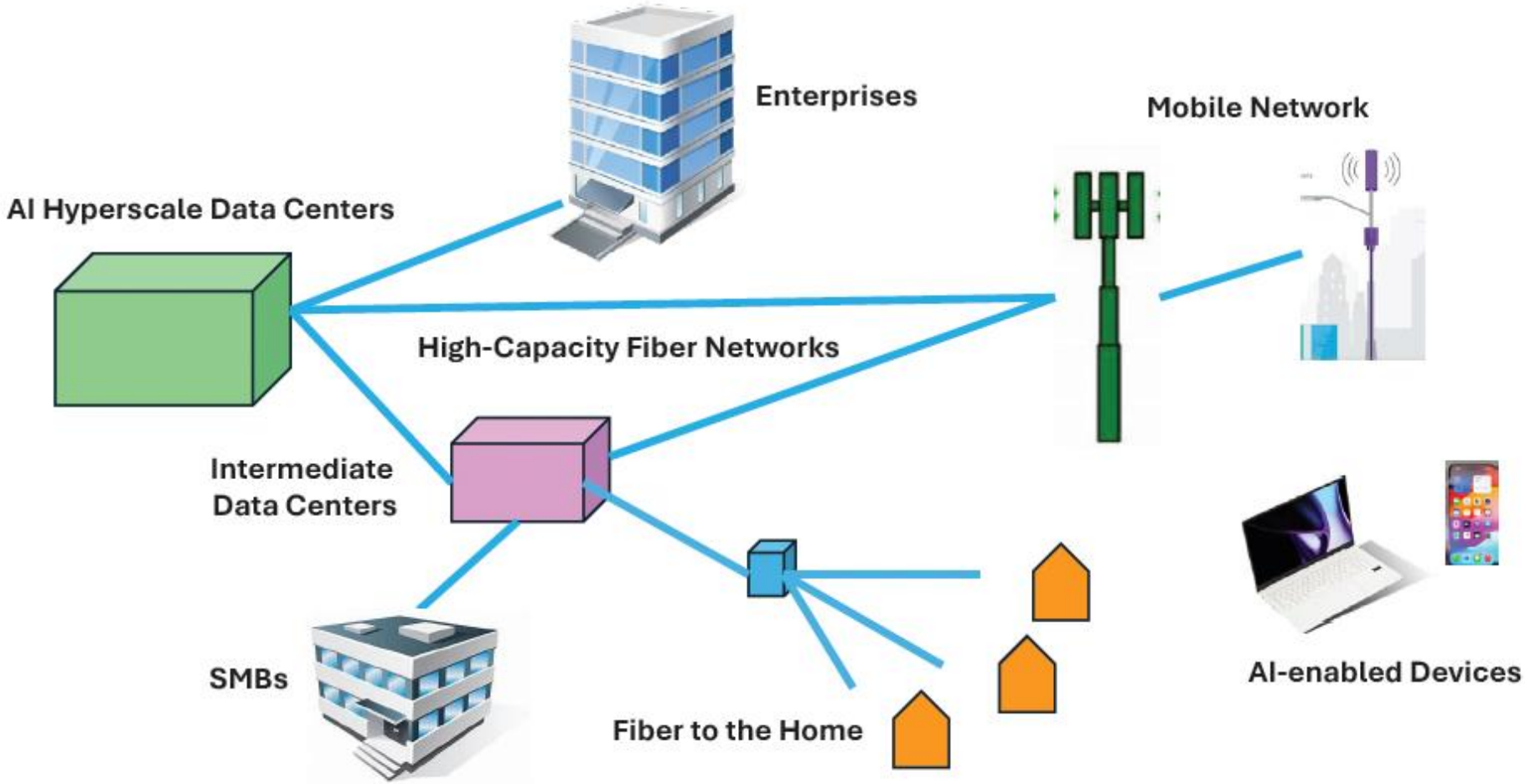
One AI-generated image =

50-100%

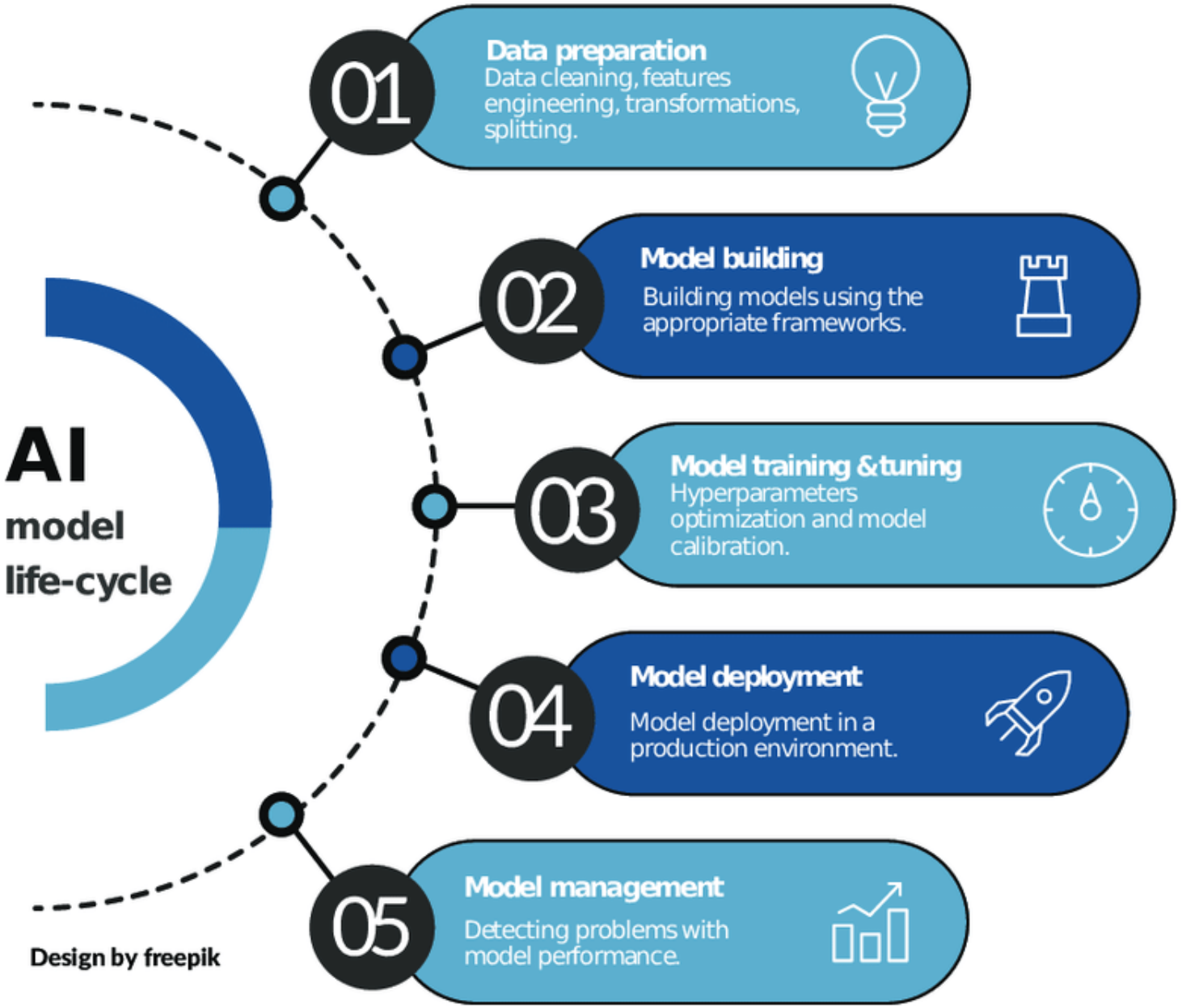
(upward of 34M/day)

AI Infrastructure Ecosystem

Photo Source: Industry reports, Inside Towers Intelligence™ Estimates



AI Lifecycle



Sustainable AI

- AI utility measured in terms of performance and accuracy
- Sustainable AI shifts the lens to efficiency and responsibility
- Challenges: design, train, deploy and retire systems with the explicit goal of minimizing environmental impact
- Across the full lifecycle – from raw material extraction and chip fabrication to data processing, training, inference and end of life

- World Economic Forum, AI Report, February 2026

The Hidden Environmental Cost of AI

Carbon Emissions:

- Training one large AI model can emit up to 500 tons of CO_2, equivalent to the lifetime emissions of 50 cars.

Energy Consumption:

- Data centers powering AI require as much electricity as thousands of homes each day.

Water Usage:

- Cooling AI servers consumes millions of liters of water annually.

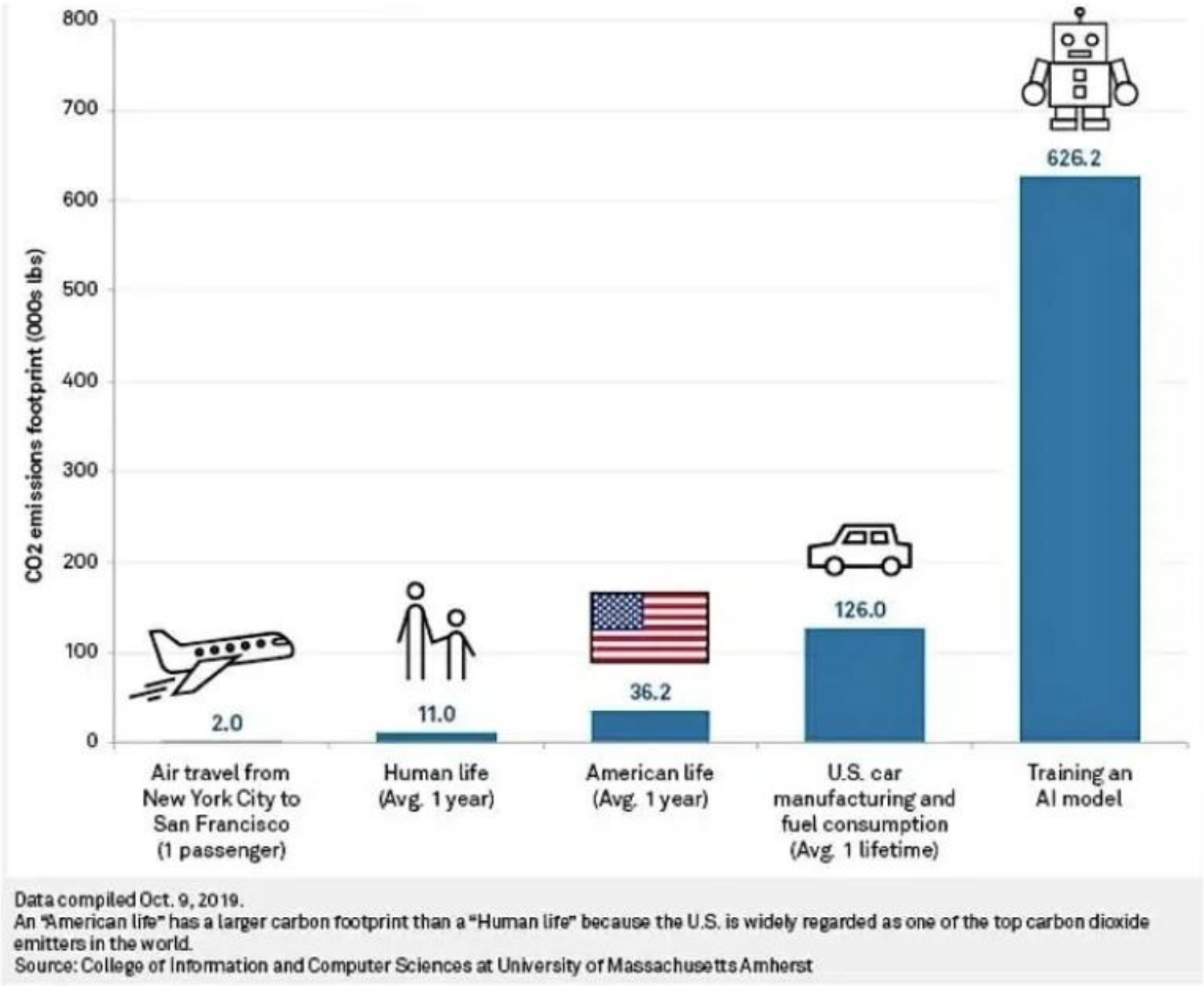
Electronic Waste:

- Rapid hardware upgrades generate tons of e-waste every year, including rare metals and toxic components.

Climate Impact:

- The energy-intensive AI lifecycle contributes to global warming and resource depletion.

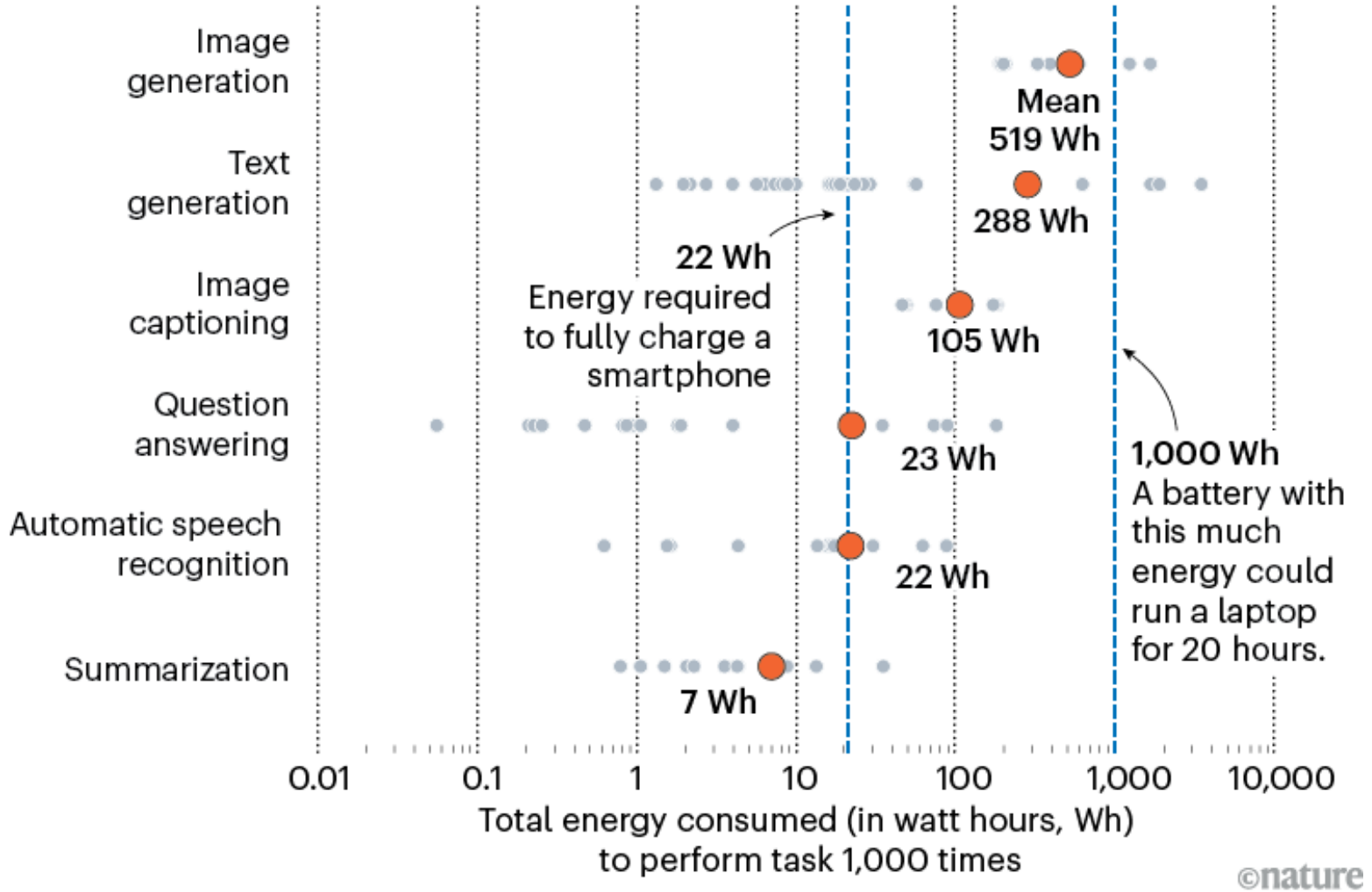
Carbon Footprint



Data compiled Oct. 9, 2019.
An "American life" has a larger carbon footprint than a "Human life" because the U.S. is widely regarded as one of the top carbon dioxide emitters in the world.
Source: College of Information and Computer Sciences at University of Massachusetts Amherst

HOW MUCH ENERGY DOES AI USE?

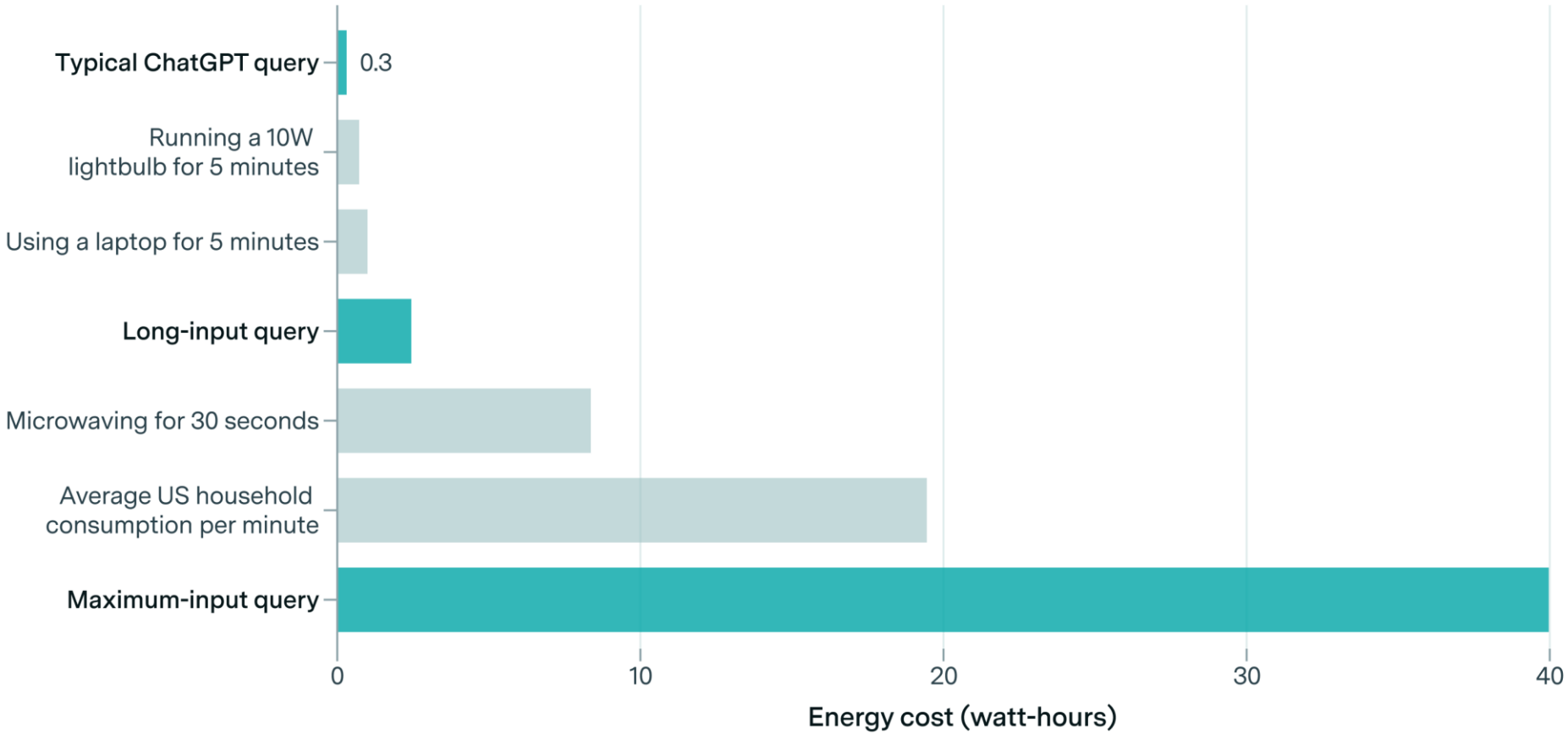
The AI Energy Score project tested dozens of artificial-intelligence models to estimate how much energy they consume when performing various tasks. Plotting the energy required to perform a task 1,000 times shows that energy use varies greatly depending on the task and the model.



©nature

Energy consumption per ChatGPT query is small compared to everyday electricity use

EPOCH AI



Pessimistic estimates of the energy usage of ChatGPT with GPT-4o across for different query lengths: typical (<100 words), long (~7,500 words), and maximum context length (~75,000 words), with an average response length of 400 words.

CC-BY

AI Water Usage

Every time you ask an AI
a question, it uses about

500 ml of water

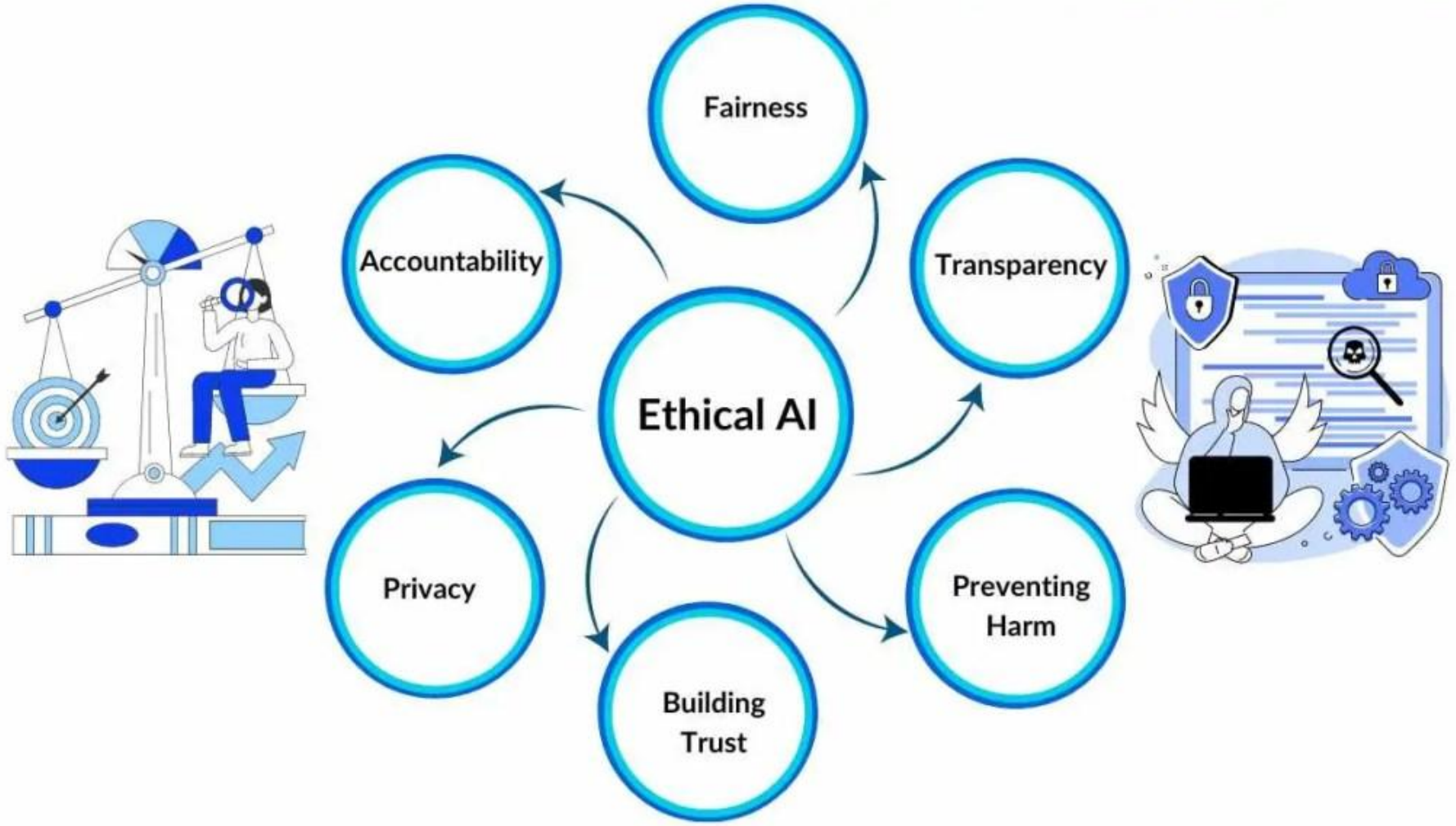
One question = one small bottle of water

A full day of heavy AI use = a kiddie pool

Training a whole AI model = a lake

As we continue to innovate, it’s important
to remember the hidden environmental costs.

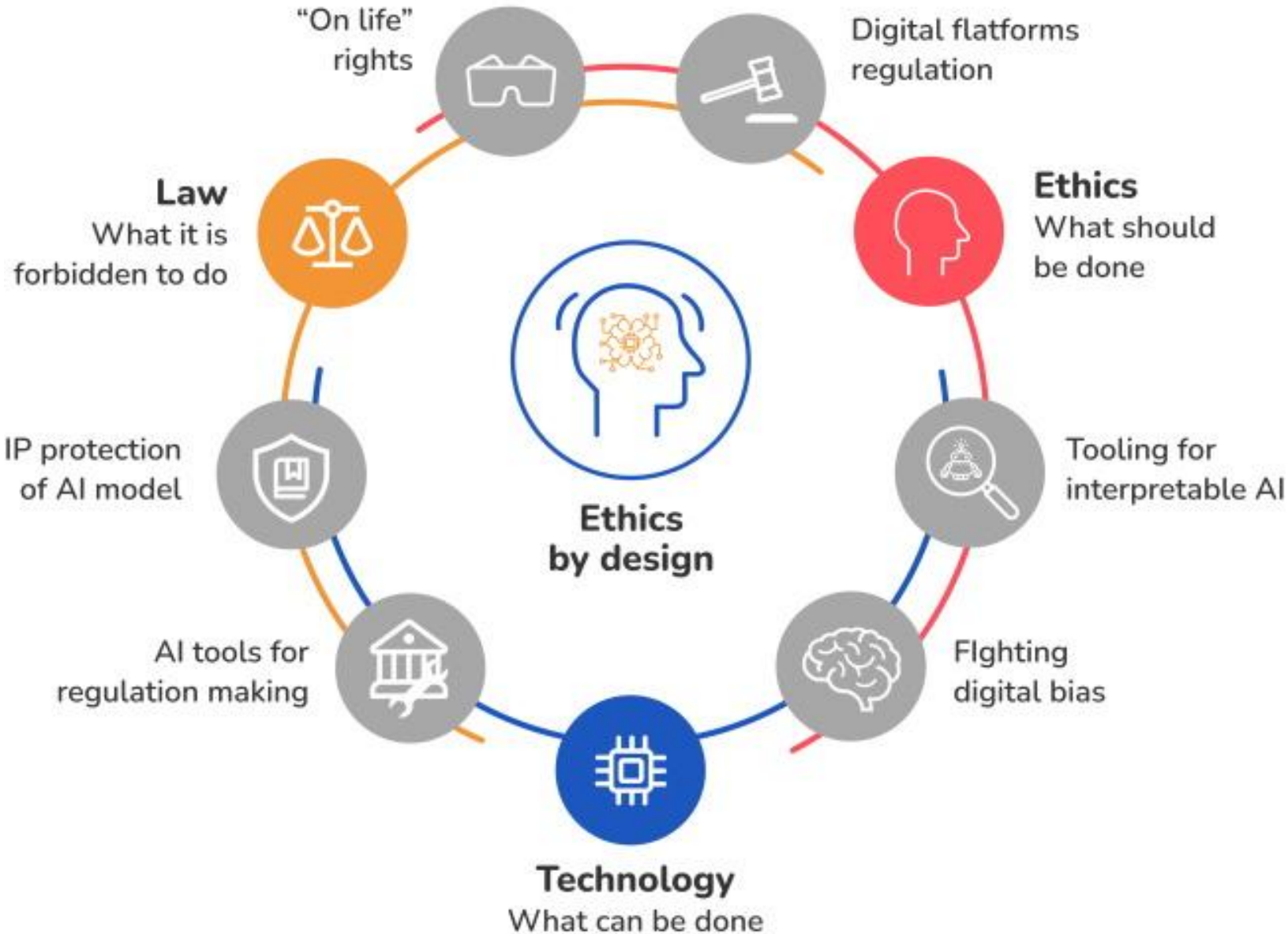
Ethical AI

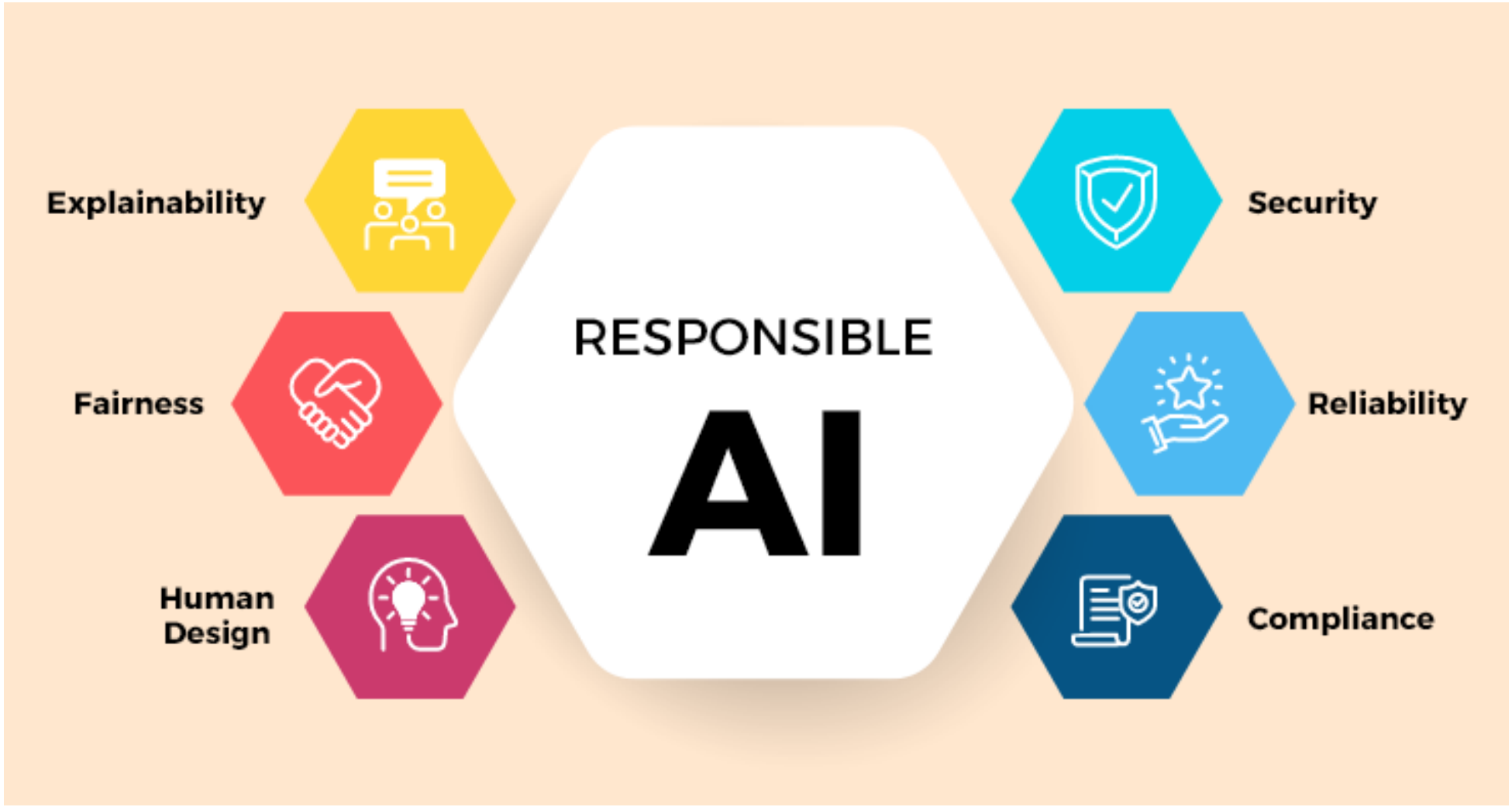


Belmont Report:

- Respect
- Justice
- Beneficience

Moral guidelines for designing, developing, and deploying AI systems - so they align with human values and respect human rights


"On life" rights
Digital flatforms regulation
Law
What it is forbidden to do
Ethics
What should be done
IP protection of AI model
Tooling for interpretable AI
Ethics by design
AI tools for regulation making
FIghting digital bias
Technology
What can be done



Overarching framework for designing, developing, and deploying AI systems - safely, ethically, and with human oversight – maximise social benefit

Sustainable AI – Path Forward

- Sustainability must be embedded into AI's design, measurement and governance from the outset
  - Not as an afterthought
- Collective action: energy providers, hardware manufacturers, cloud and data centre operators, developers, investors, regulators, customers and e-waste partners all shape the outcomes

From theory to action: 5 places to start

Design energy-efficient models & architectures

Make “sustainable by design” the default.

Optimize data center operations & renewable usage

Where and when you compute matters.

Extend hardware lifecycles through circularity

Treat hardware as an asset, not a consumable.

Use data responsibly & reduce redundant processing:

Data bloat drives unnecessary computing.

Deploy governance guardrails

Make sustainability stick by ensuring accountability is built in.

NTT DATA

Approach to AI Sustainability

Driving impact through smart use, efficient models, and clean energy

Smart Demand

Using AI wisely

- Agentic Architecture
  - Agentforce
- AI Transparency
  - AI Energy Score

Efficiency

Doing more with less

- Small, Domain-Specific Models
- Efficient Data & Hardware
  - Data Cloud

Clean Supply

Low-impact resources

- Investing in clean power

- Collaborating to advance the industry

AI Ecosystem

COALITION FOR SUSTAINABLE AI

GOUVERNEMENT
Liberté
Égalité
Fraternité

ITU

UN environment programme

Mindful use of AI as an Academic

- AI truly needed? – don’t use AI out of habit
- Can I use local AI instead of cloud AI?
- Does AI use maintain academic integrity?
- Does AI use encourage bias and discrimination?
- Is the AI model trustworthy and secure?
- Does using AI effect my mental health?
- Does using AI effect my skills and creativity?

How to Practice Conscious AI Use

- Master Prompt Engineering
  - Craft clear, concise, and specific prompts. Vague prompts lead to trial-and-error, which wastes significant computing power and energy on the backend.
- Limit Generative AI for Simple Tasks
  - Ask yourself, "Would I do this differently without AI?" If a task can be solved by a quick search or basic script, avoid deploying heavy, multi-purpose generative models.
- Batch Your Queries
  - Consolidate your questions into a single session rather than opening multiple, separate chat.

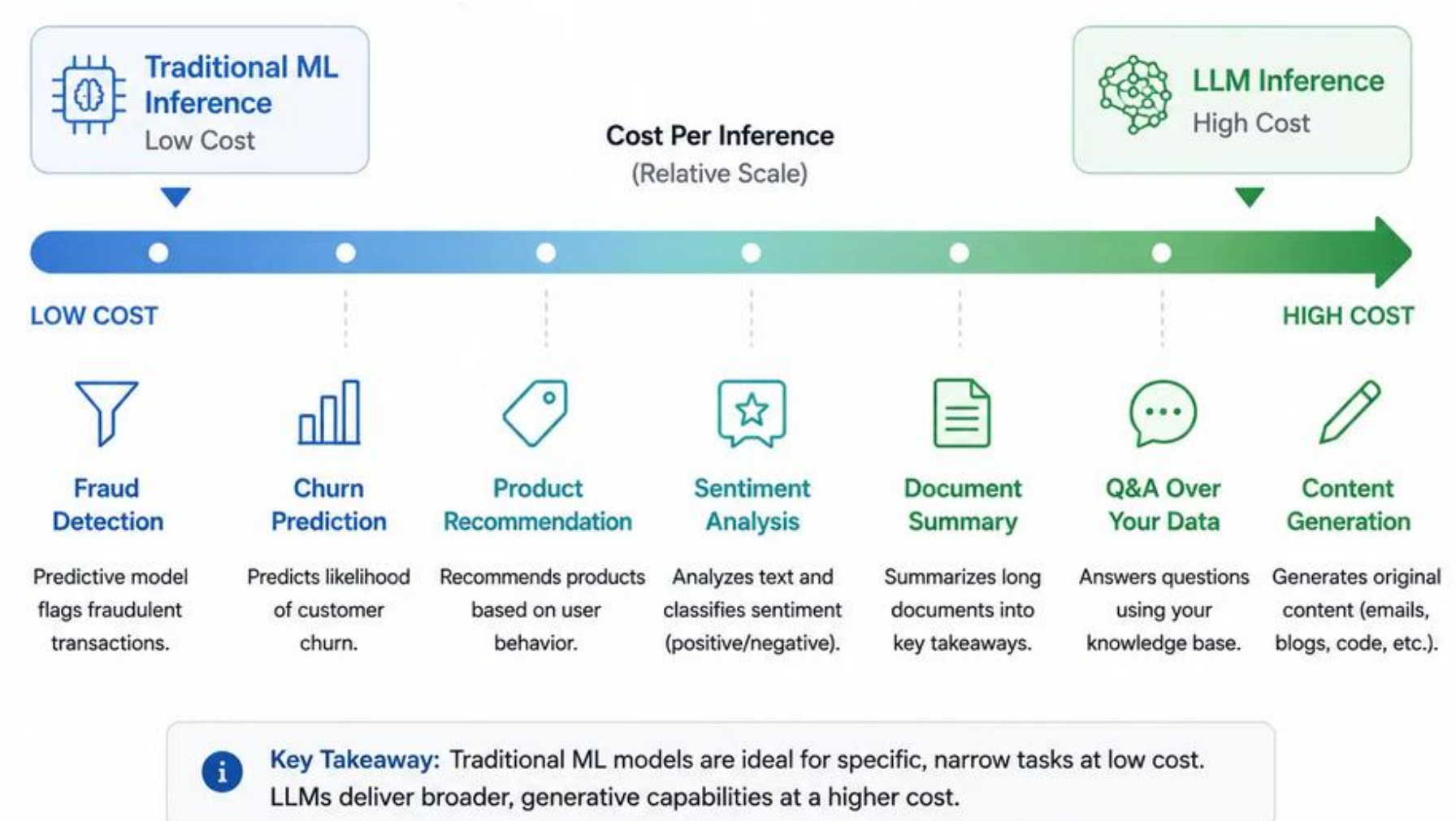
GenAI is Costly!

Comparison with Everyday Activities

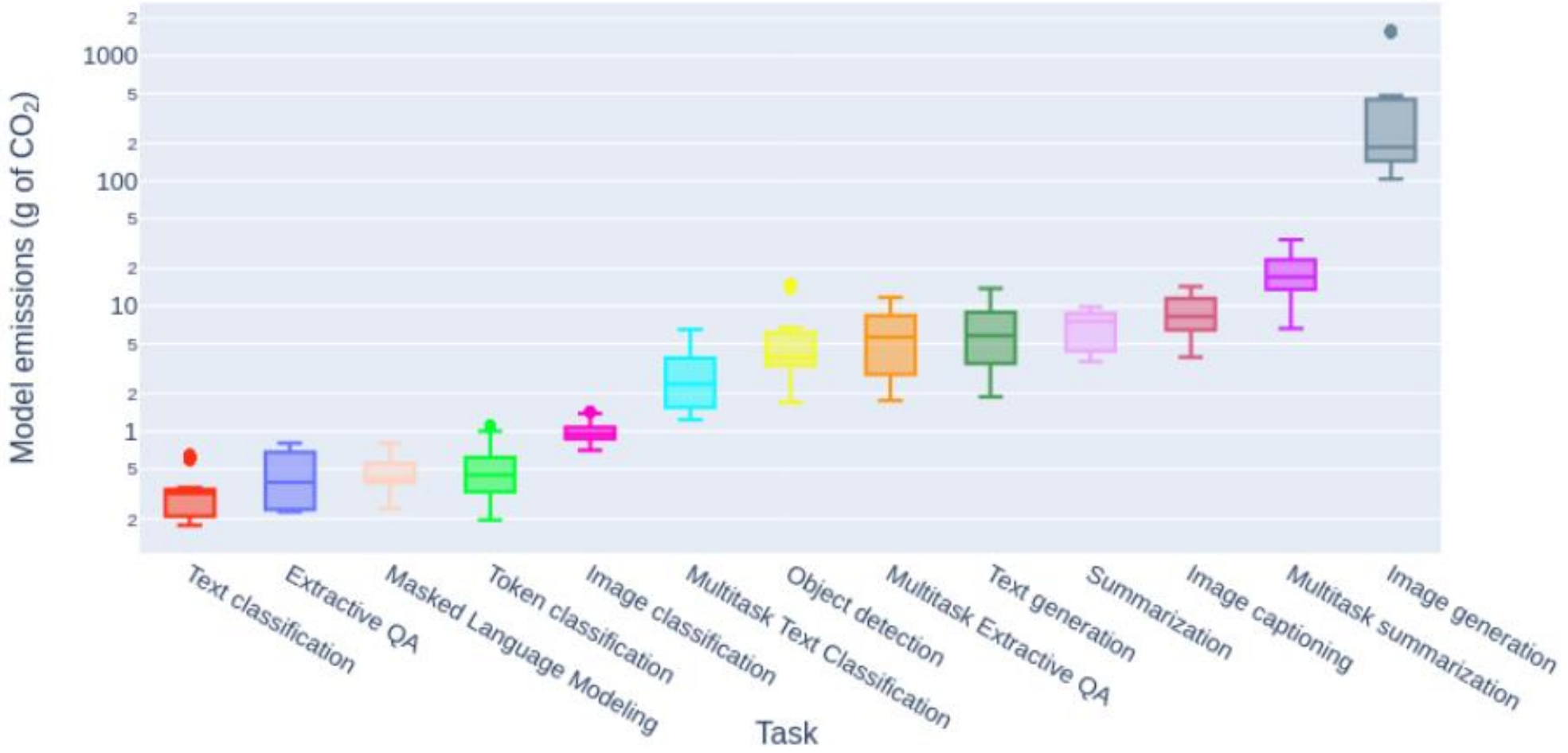
Carbon Footprint

Activity (typical use)	CO_2 Emissions (grams)	Comparison to AI Queries
Generative AI Query	~4.3	Baseline
Driving a gasoline car (1 mile)	~404	Significantly higher than AI queries
Hot shower (10 minutes)	~2,000	Substantially higher than AI queries
Clothes dryer (1 load, ~1 hour)	~1,000	Substantially higher than AI queries
Electric oven (1 hour)	~675	Substantially higher than AI queries
Boiling water in kettle (~1 liter)	~50–70	Significantly higher than AI queries
Microwave (5 minutes)	~40–50	Significantly higher than AI queries
Video Conferencing (per minute)	2.5–16.7	Higher or comparable to AI queries
Streaming Video (per minute)	0.6–1.7	Slightly lower or comparable to AI queries
Online Gaming (per minute)	0.3–1.0	Lower or comparable to AI queries
Social Media (per minute)	0.3–0.5	Lower than AI queries
Web Browsing (per minute)	0.1–0.2	Notably lower than AI queries
Email Usage (per email)	0.02–0.17	Significantly lower than AI queries

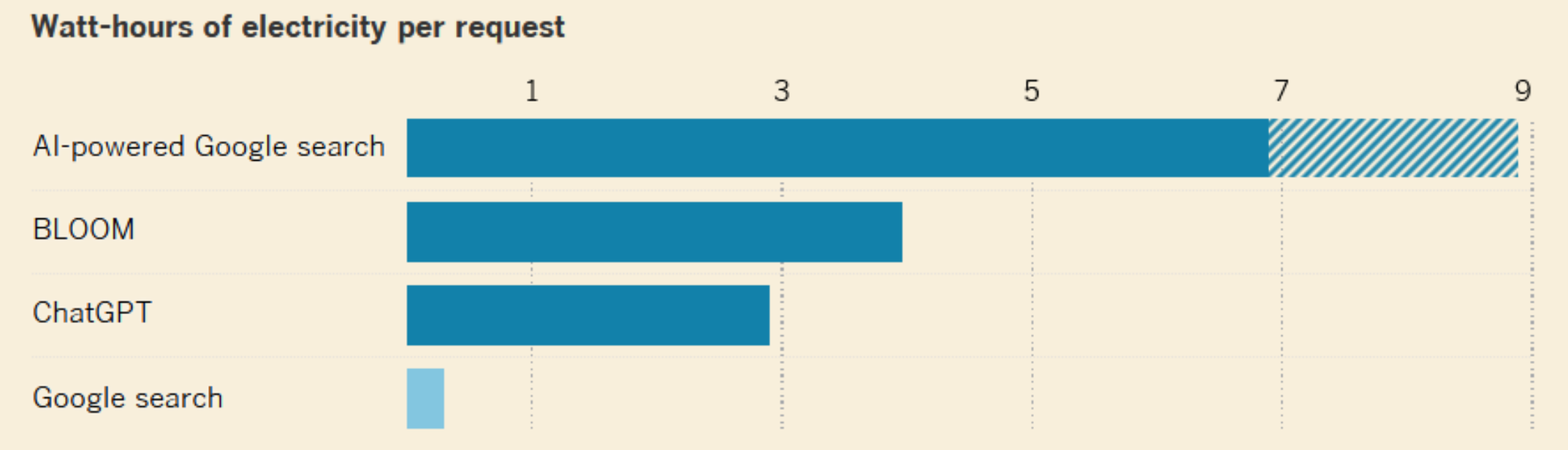
Generative AI Costs



Climate Impact of GenAI Tasks



Searching for Information



Right Sized Intelligence in Practice

- Right Model for a Task
  - Simple Tasks: For text classification, sentiment analysis, or simple data extraction, use compact models (e.g., Gemini Nano, Phi, or Llama 8B)
  - Complex Reasoning: Reserve larger, energy-heavy models only for advanced synthesis, creative writing, or complex problem-solving.
- Model Routing
  - Model Cascading: Send a prompt to a lightweight model first. If it fails to yield a confident answer, only then query a heavier model.

Right Model for a Task

- Edge vs. Cloud Computing
  - Client-Side/On-Device: Running models (Gemini Nano) locally on your phone or PC drastically reduces cloud computing and data transmission energy costs.
- Best-in-Class Efficient Models
  - On-Device & Mobile: Google's Gemini Nano for optimized edge inference
  - Open-Weight Workhorses: Deploy Meta's Llama 3 (8B parameters) or Microsoft's Phi Series for highly capable, low-footprint language reasoning.
  - Sparsely Activated Models: Utilize Mixtral 8x7B, which uses a Mixture of Experts (MoE) to activate only a fraction of its parameters at a time

Thank You!