FRUGAL AI

Building Sustainable and Accessible AI Solutions

Arjuna Sathiaseelan

FRUGAL AI HUB at UNIVERSITY OF CAMBRIDGE Judge Business School

Agenda

01

The AI Adoption Surge

Fastest technology adoption in history

02

The Model Creator Problem

Costs, energy, water, compute

03

The Enterprise Failure

Why 95% of AI pilots fail

04

The Accessibility Gap

Billions excluded from AI benefits

05

Frugal AI: The Solution

Environmental Friendly. Affordable. Sustainable

06

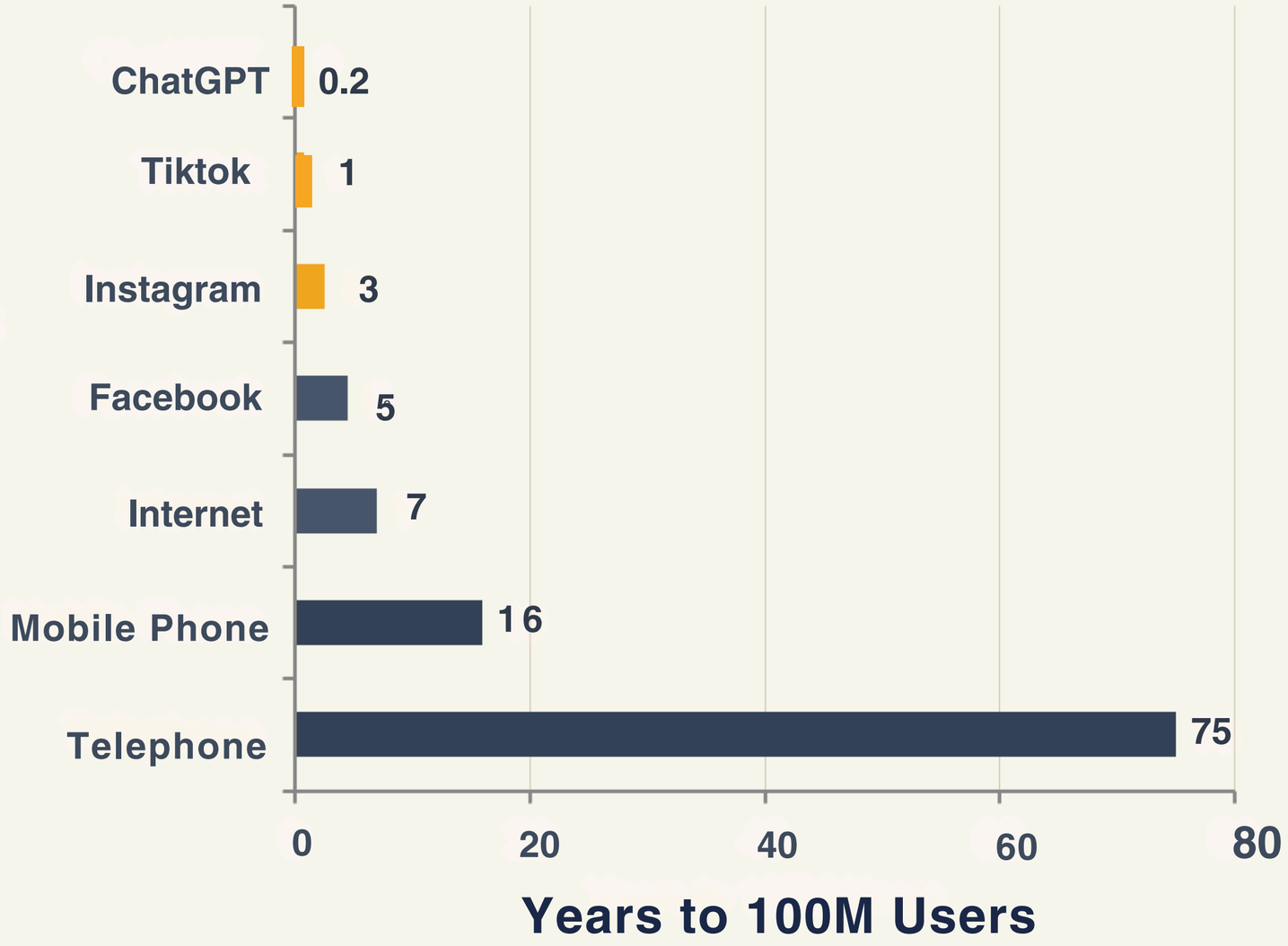
Frugal AI Hub & Projects

Cambridge research & global deployments

The AI Adoption Surge

Fastest technology adoption in human history

AI Reached 100M Users Faster Than Any Technology in History



2 Months

ChatGPT to 100M users
(vs 75 years for telephone)

88%

of companies use AI in at least one business function

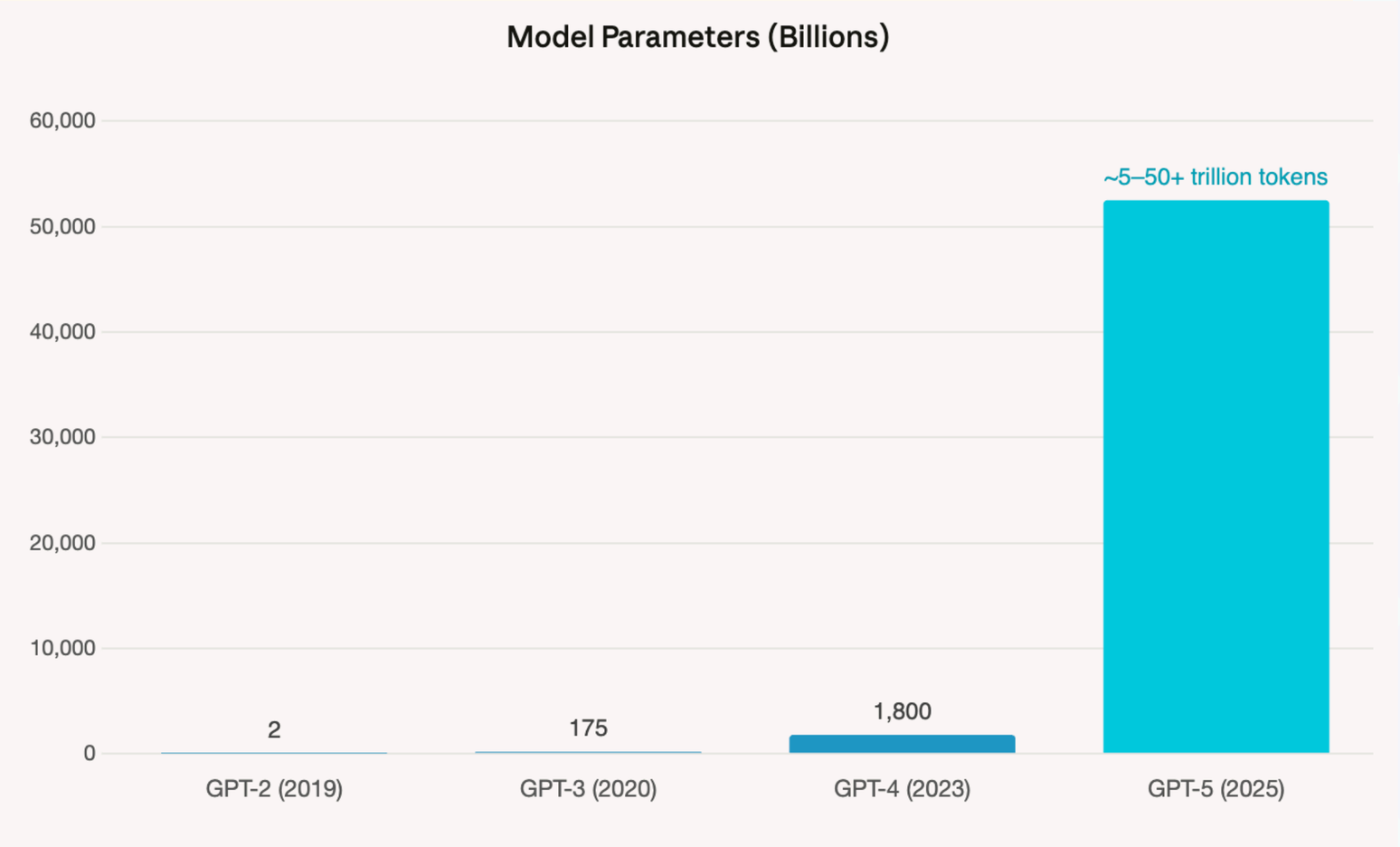
$312B → $1.8T

Global AI market size:
2026 → 2030 projection

Problem 1: The Model Creator Crisis

Bigger models. Bigger costs. Bigger environmental damage.

AI Model Size Has Exploded: Parameter Growth 2019–2025 (Billions)



GPT-2 (2019)	~$50K
GPT-3 (2020)	~$5M
GPT-4 (2023)	~$100M
GPT-5 (2025)	~$500M

10,000× cost increase in 6 years

† GPT-5 uses a Mixture-of-Experts (MoE) architecture with a router selecting sub-models per query. Dense-equivalent ~1.8T; total MoE ~52.5T (industry estimate — not confirmed by OpenAI).

Sources: OpenAI, Meta, Google DeepMind model cards; SemiAnalysis; LifeArchitect.ai; CometAPI estimates 2025

AI Energy and Water Consumption

+118%

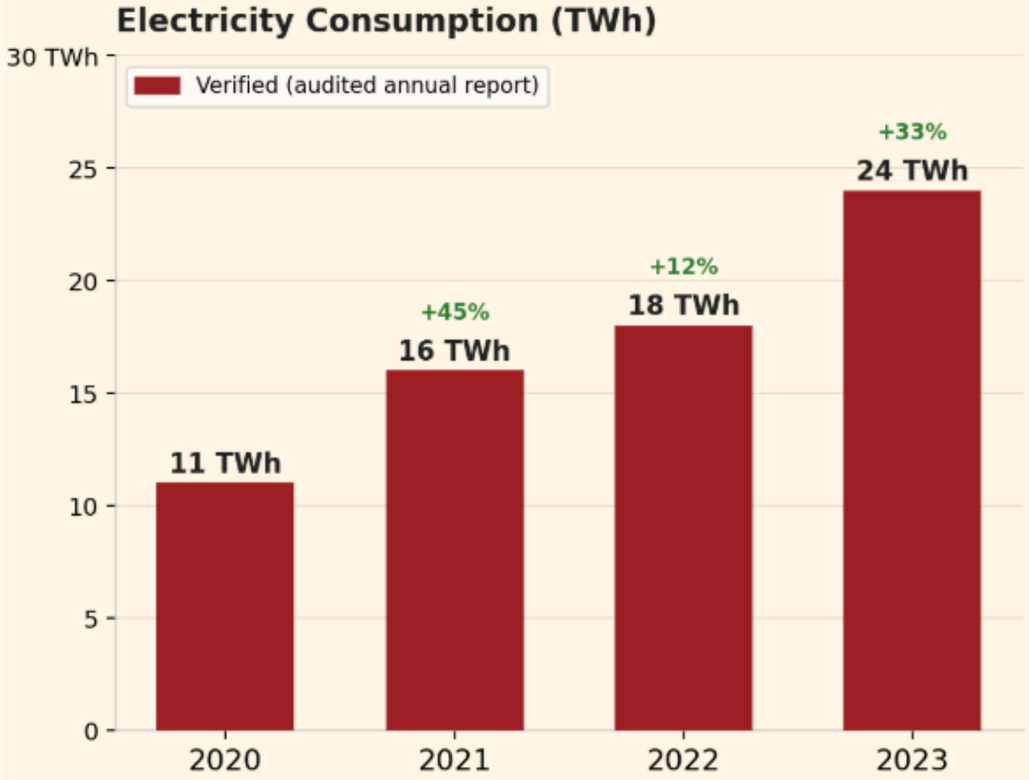
Electricity increase
2020 → 2023

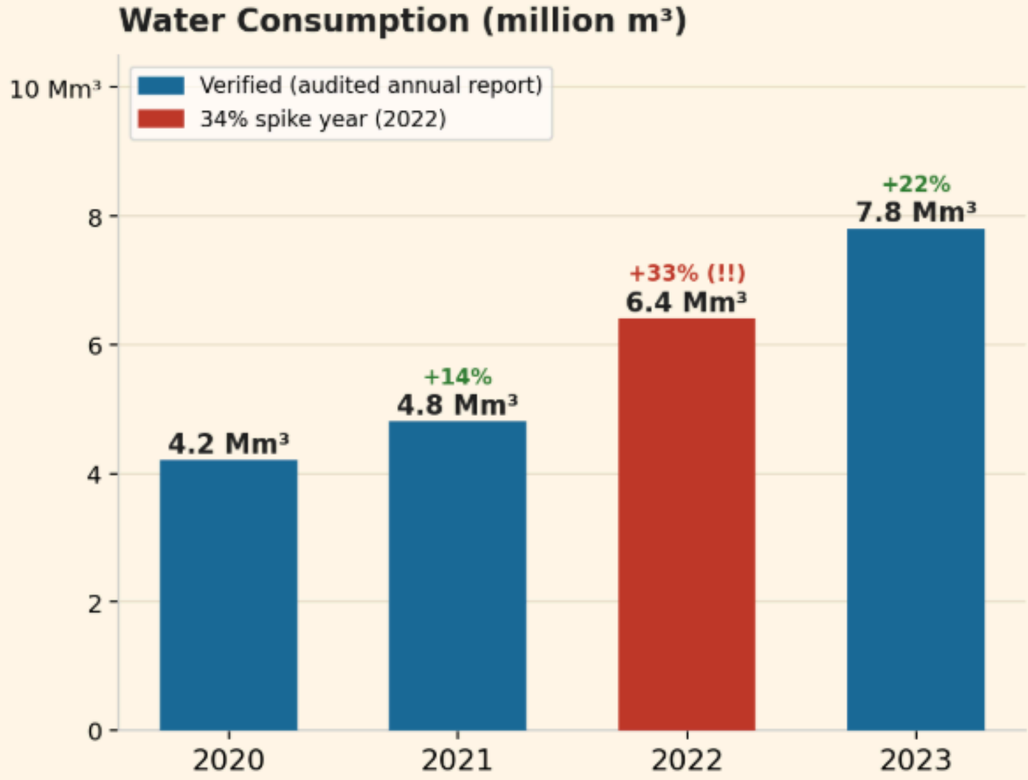
+86%

Water increase
2020 → 2023

+34%

Water spike in 2022 alone
coinciding with AI buildout

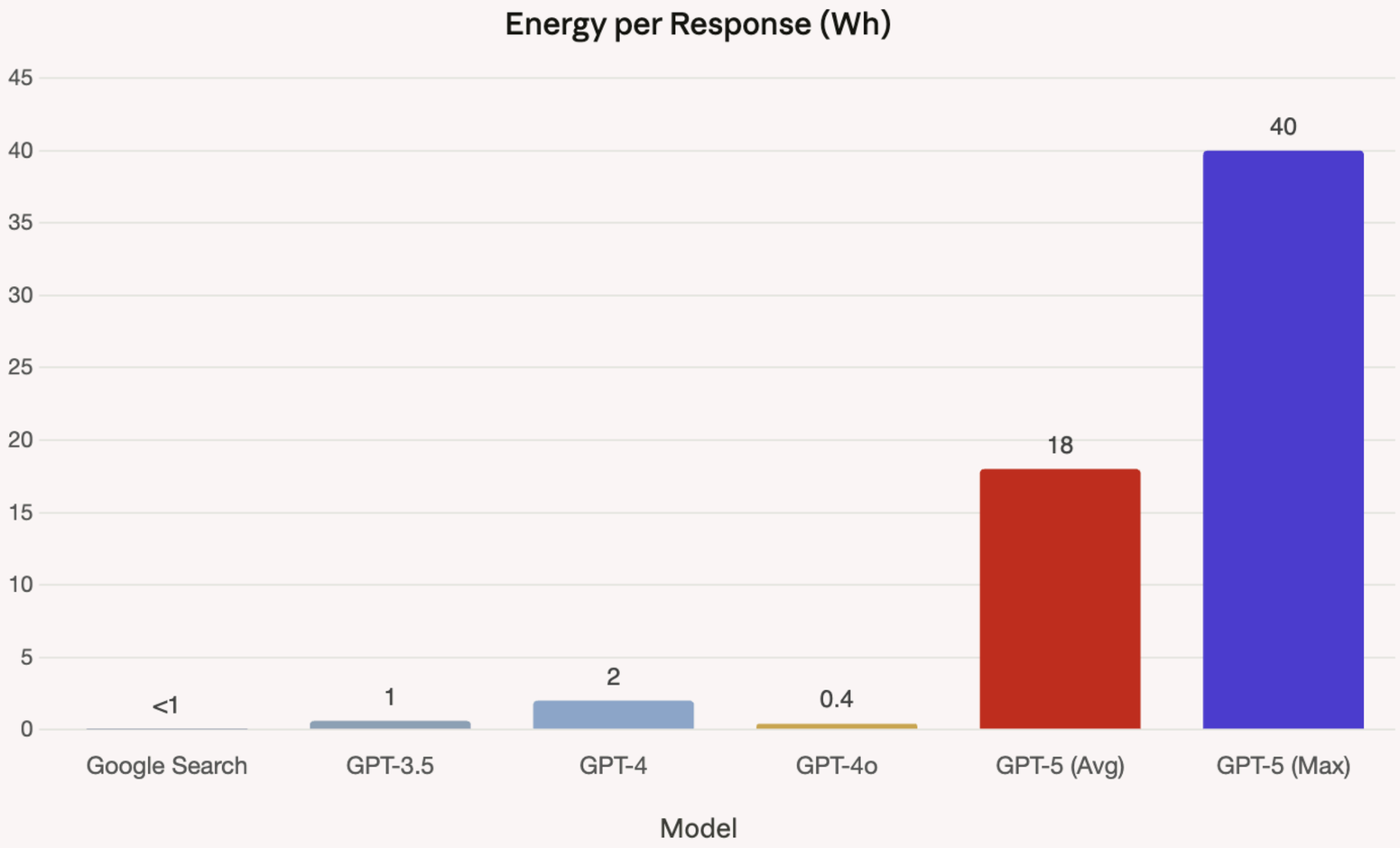




Microsoft's electricity use doubled in 4 years · Water surged 34% in 2022 alone, coinciding with AI model training & deployment

Source: Microsoft Environmental Sustainability Report 2024 (covers FY2020–FY2023), publicly available at microsoft.com.
Figures are company-wide totals, primarily driven by data centres. Microsoft explicitly cites AI and cloud growth as key drivers of increased consumption.

GPT-5 Inference Energy: 8× Higher Per Response Than GPT-4



18.35 Wh

Average GPT-5 response energy
= running a lightbulb for 18 min
(Univ. of Rhode Island, Aug 2025)

8×

Higher than GPT-4 per response
(GPT-4 avg: 2.12 Wh)

1.5M homes

Daily electricity equivalent if
GPT-5 handles 2.5B daily queries
at avg 18 Wh per query

NOT disclosed

OpenAI has not published
official GPT-5 energy figures (independent estimates only)

Sources: University of Rhode Island AI Lab (Aug 2025), Epoch AI, Guardian reporting

AI Energy Crisis Making News Headlines

DATA CENTERS

TOM'S HARDWARE — Nov 13, 2025

OpenAI's Data Center Targets Would Consume As Much Electricity As Entire Nation of India

Sam Altman's 250GW plan by 2033 equals power for India's 1.5 billion citizens — and would emit twice the CO_2 of ExxonMobil

POWER GRID

FORTUNE — Oct 11, 2025

OpenAI's 10GW Data Center Plan Requires as Much Power as New York City on Its Most Energy-Intensive Summer Days

Sam Altman's deal with Nvidia raises urgent questions about whether AI firms can actually secure the electricity they need

ENERGY USE

IEEE SPECTRUM — Oct 3, 2025

ChatGPT's Annual Energy Consumption Rivals That of Entire Small Nations — Slovenia, Puerto Rico

17,228 GWh per year: ChatGPT uses more electricity annually than Slovenia (14,630 GWh) and approaches Puerto Rico's national grid

CLIMATE

MIT TECHNOLOGY REVIEW — Sep 22, 2025

U.S. Data Centers Could Triple Their Share of National Electricity by 2028 — Driven Almost Entirely by AI

Share of US electricity to data centers may surge from 4.4% to 12% — more than New York State uses in an entire year

Sources: Tom's Hardware · Fortune · IEEE Spectrum · MIT Technology Review

Why Is AI So Expensive and Energy-Hungry? Four Root Causes

Frontier AI is not inefficient by accident! It is the direct result of architectural choices and research incentives that prioritise capability over cost.

01 Scaling Laws

Bigger = Better (but at explosive cost)

Empirical research shows model performance scales as a power law with model size, dataset size, and compute. Each capability gain requires exponentially more resources.

02 Quadratic Attention

$O(n^2)$ compute per token

The Transformer's self-attention mechanism scales quadratically with sequence length. Doubling context doubles parameters, but quadruples attention compute.

03 Dense Architectures

All neurons fire for every token

Standard dense models activate 100% of parameters for every input, even simple queries. A model with 1T parameters uses all 1T for 'What day is today?'

04 Hardware & Datacentre Overhead

GPUs, cooling, and grid inefficiency

H100 GPUs draw ~700W each. Training runs use thousands simultaneously. ~65%-85% of energy drawn goes to actual computation, the rest is cooling and overhead.

Problem 2: The Enterprise AI Failure

95% of AI pilots never reach production

95% of Enterprise AI Pilots Never Reach Production — Three Root Causes

95%

of AI pilots never reach production

MIT NANDA Study '25

Data Challenges

- Poor quality, siloed, or inaccessible data
- Labelling costs prohibitive at scale
- No data governance or lineage strategy
- Privacy & consent barriers block usage

Integration & Capacity

- Legacy system incompatibility
- Shortage of ML engineering talent
- Cloud dependency and latency constraints
- No edge or offline deployment capability

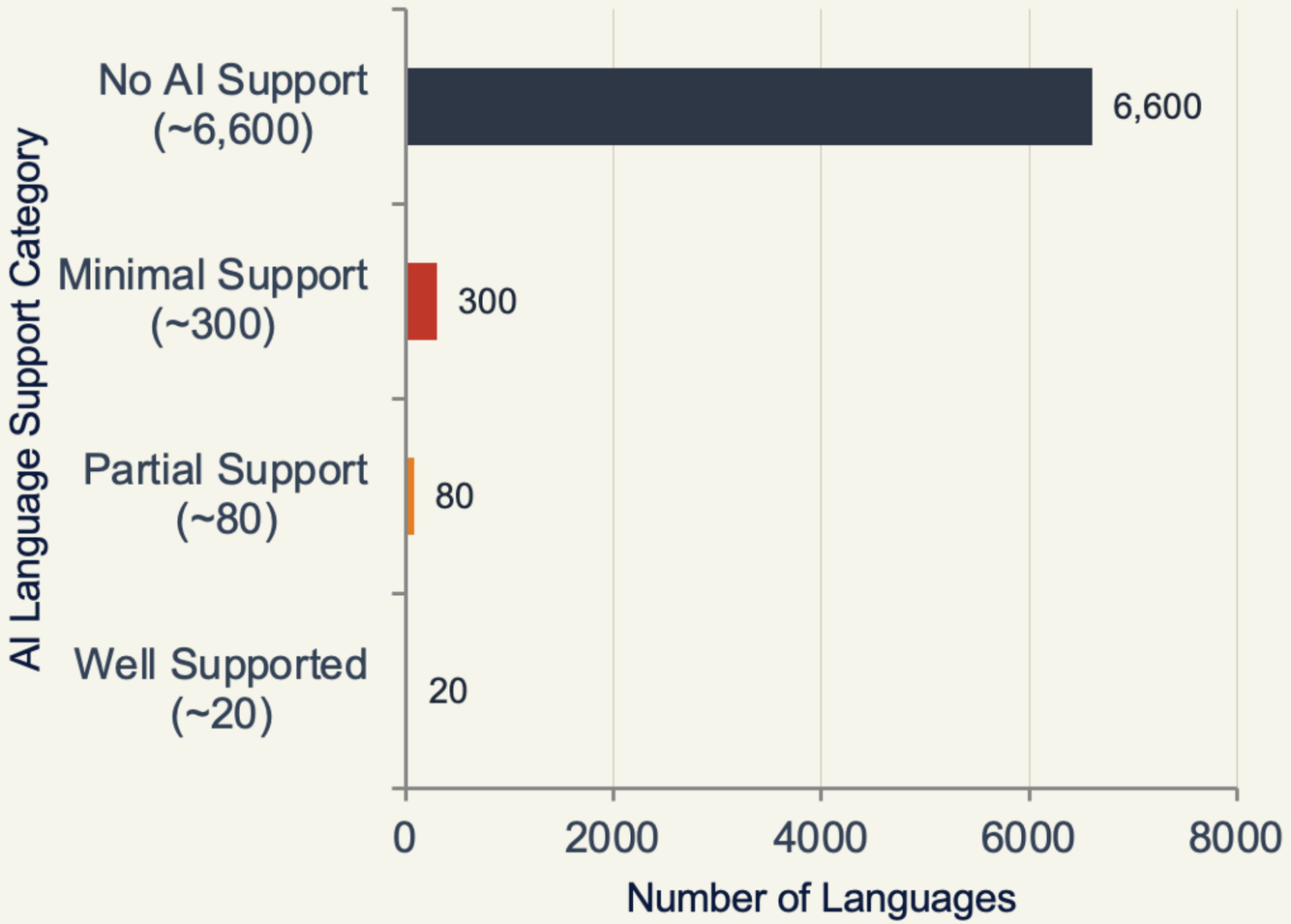
Governance & Trust

- Regulatory compliance and AI Act gaps
- Model explainability failures
- Bias in training data and outputs
- No clear ownership or accountability

Problem 3: The Accessibility Gap

Billions of people are invisible to AI systems

AI Speaks ~100 Languages — The World Speaks 7,100+



7,100+

Languages exist globally

~100

Languages with good AI support

6,600+

Languages with NO AI support

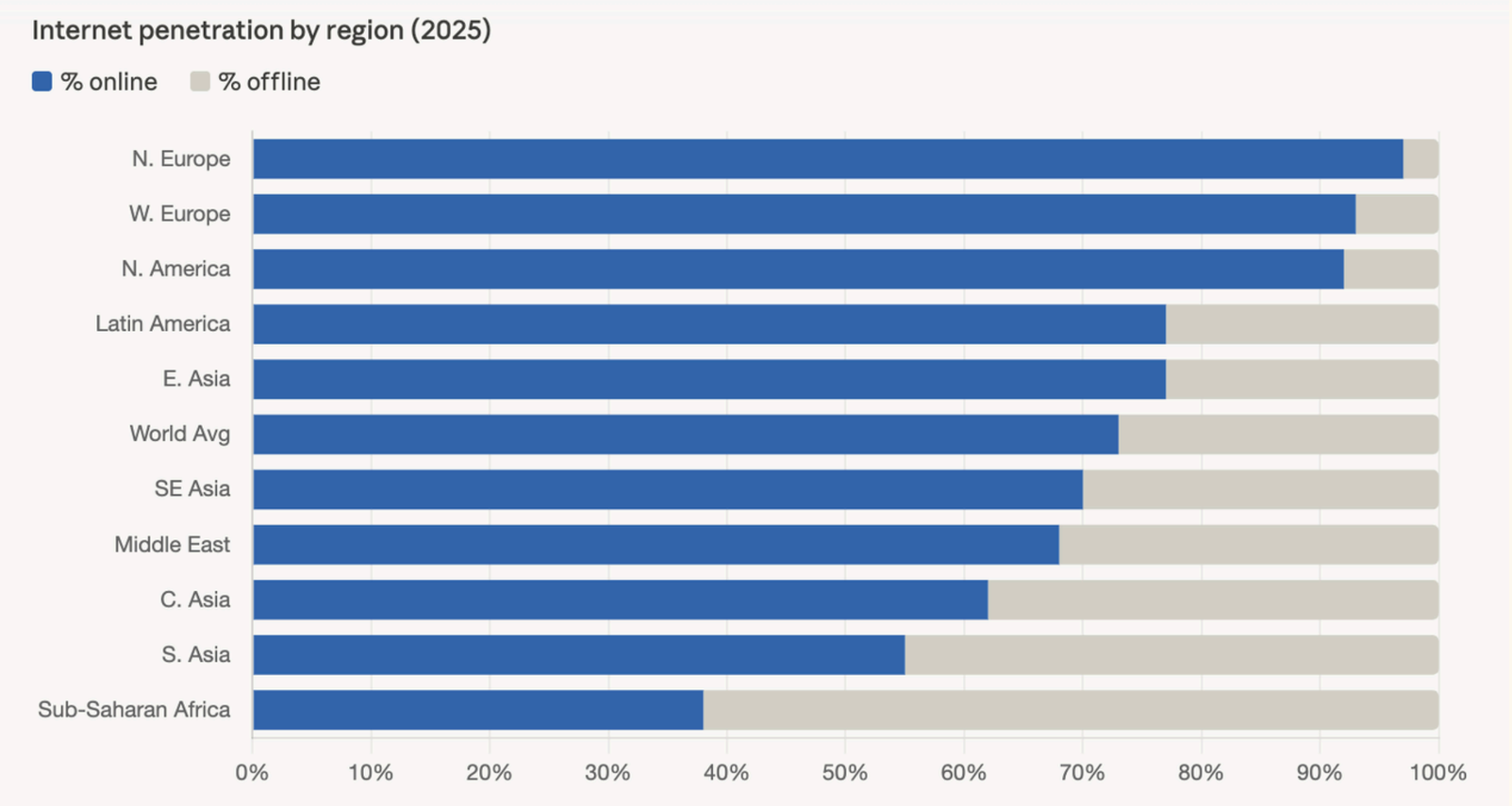
40%

World's population with limited/no AI in native language

95% of AI training data is in English, Chinese, or European languages

Sources: UNESCO, OECD AI Observatory, W3Techs 2024

Connectivity and Economic Barriers Exclude Billions More



2.6 Billion

People remain offline globally
(ITU 2024)

$10/month

Typical AI subscription cost vs
$2.15/day global poverty line

400ms+

AI inference latency on 3G vs
50ms needed for voice AI

AI is accelerating inequality: the most connected benefit, the least connected fall further behind

Frugal AI

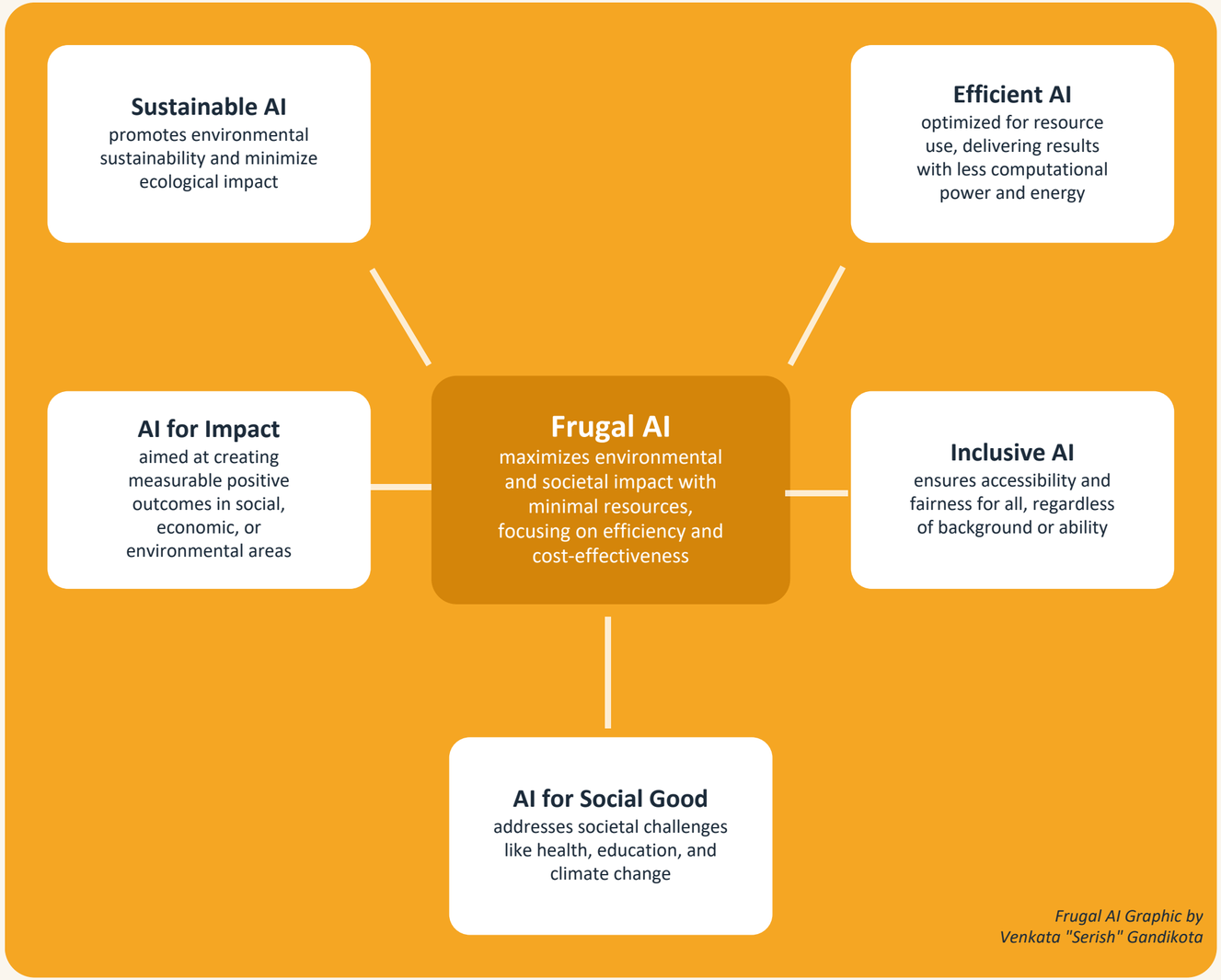
Environmental Friendly. Affordable. Sustainable

Frugal Innovation - Doing more with less

A design philosophy focused on drastically reducing cost and complexity to deliver good enough solutions that are accessible to resource constrained or underserved populations

Frugal AI

Design philosophy focused on building AI systems that achieve strong performance while minimizing computational resources, energy, data, and cost.



The Frugal AI Technical Stack

Where Every Layer Saves: Engineering Efficiency Across the Full AI Pipeline

01 Input Layer

02 Model Layer

03 Compute Layer

04 Orchestration Layer

05 Open Foundation Layer

Control What Enters — Reduce Everything Downstream

4×

Compute reduction from halving tokens

60–80%

Cost cut via prompt compression & pre-summarisation

≈0 changes

Savings with no model modifications needed

1. Token count is a direct compute multiplier — every token triggers operations across all transformer layers
2. Input compression (remove redundancy, pre-summarise docs, extract features upstream) cuts tokens without touching the model
3. Use better representations to reduce token count e.g. TOON
4. Quadratic attention scaling means halving tokens can cut attention computation by up to 75%

The Quadratic Problem: Why Transformer Attention Is Brutally Expensive

Self-attention requires every token to attend to every other token. With n tokens, that's n × n = n² comparisons. This is not a bug — it is fundamental to how Transformers work.

Attention Compute Grows Quadratically with Context Length

Context Tokens	Attention Ops (n^2)	vs 1K baseline	Real-world use
1,000	1M	**1×**	Short email
4,000	16M	**16×**	Document page
32,000	1B	**1,024×**	Long report
128,000	16.4B	**16,384×**	Book chapter
1,000,000	**1T**	**1,000,000×**	**GPT-4 context (est.)**

Sources: Vaswani et al. 2017; OpenAI API pricing 2024; Dao et al. 2022 (FlashAttention)

Memory Explosion

Naïve 128K-token attention would need roughly a whole high-end 32 GB GPU just to hold one layer’s attention scores.

Cost per Long Query

Processing a 100K-token context with GPT-4 costs ~$3-$9 (with output). The same task on an efficient self-hosted or edge model can be 10–100× cheaper.

Why Frugal AI Avoids This

“Task-specific small models run with short, bounded contexts (typically 4k–32k tokens), so they avoid expensive million-token attention and can cut cost and energy use by well over 60% compared with large general models.

Right-Size Intelligence — Eliminate Redundant Parameters

4×

Memory saving: FP32 → INT8 quantisation

70–90%

Size reduction via knowledge distillation

Continuous

Per-query savings — not one-off training cost

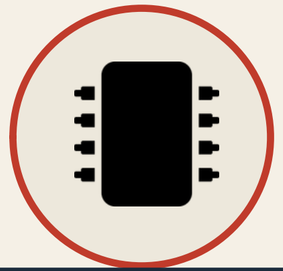
1. Over-parameterisation is the silent cost multiplier — use Pruning to remove redundant parameters
2. Quantisation (FP32 → INT8) cuts memory 4× and improves throughput on modern accelerators
3. Distillation produces smaller student models that replicate teacher behaviour with fraction of the parameters
4. RAG externalises factual knowledge, allowing smaller reasoning-only models

Match Hardware to Workload

10×
Efficiency gain: specialised accelerator vs general GPU

40–60%
Cloud cost reduction via right-sizing instances

~0%
Idle cost with scale-to-zero autoscaling



1. GPUs are flexible but energy-intensive; ASICs, NPUs & LPUs can cut per-query energy dramatically for stable workloads

2. Cloud pricing compounds across compute + storage + egress + orchestration — right-size every component

3. Idle GPU at 10% utilisation still incurs 100% of hourly cost — autoscaling and batch scheduling eliminate waste

4. Hardware alignment to workload is a first-class architectural decision, not an infrastructure afterthought

Never Send to the Cloud What the Edge Can Handle

Upto 80%

of enterprise AI tasks can run at the edge — SLMs at 10–30x lower cost

Upto 80%+

Cost reduction vs cloud for edge-eligible workloads

<10ms

Edge latency vs 100–300ms cloud round-trip — non-negotiable for real-time AI

1 Edge-first inference

Deploy compressed models on-device or on-prem. Process where the data lives — sub-10ms latency, no egress fees, offline resilience, full data sovereignty

2 Intelligent cloud routing

Only escalate to cloud for queries too complex for the edge. Hybrid architectures cut total cost by 15–30% vs pure-cloud or pure-edge alone

3 Edge caching & batching

Cache repetitive results at the edge before they reach the cloud. Batch remaining cloud queries for 2–4x throughput — minimise the cloud leg entirely

4 Workload triage

Route by latency, volume, and compliance: real-time decisions stay local, batch analytics go cloud.

Portability & Openness — Eliminate Vendor Lock-In

30–50%

Cost reduction by switching to open models

100%

Workload portability across cloud providers

Zero

Retraining cost when knowledge updates via RAG

1. Proprietary model APIs charge premium margins — open-weight models (Llama, Mistral) run on your own infra at marginal cost
2. Open frameworks like vLLM, ONNX, and LiteLLM enable hardware-agnostic deployment
3. Data locality control reduces cross-region egress costs and simplifies compliance requirements
4. Open foundations future-proof the stack — swap models, hardware, or cloud without architectural rework

Frugal AI Hub at Cambridge

Research • Adoption Labs • Global Ecosystem

Frugal AI Hub: Research, Adoption & Ecosystem Building at Cambridge

Based at Cambridge Judge Business School — bridging academic research, industry adoption, and policy with active work across the UK, Asia, and Africa.

Research & Innovation

Small language models, decentralised AI, responsible data usage

Framework Development

Tools to measure Impact, TCO, and ROI across AI portfolios

Frugal AI Adoption Labs

Co-creation living labs embedded in local innovation ecosystems

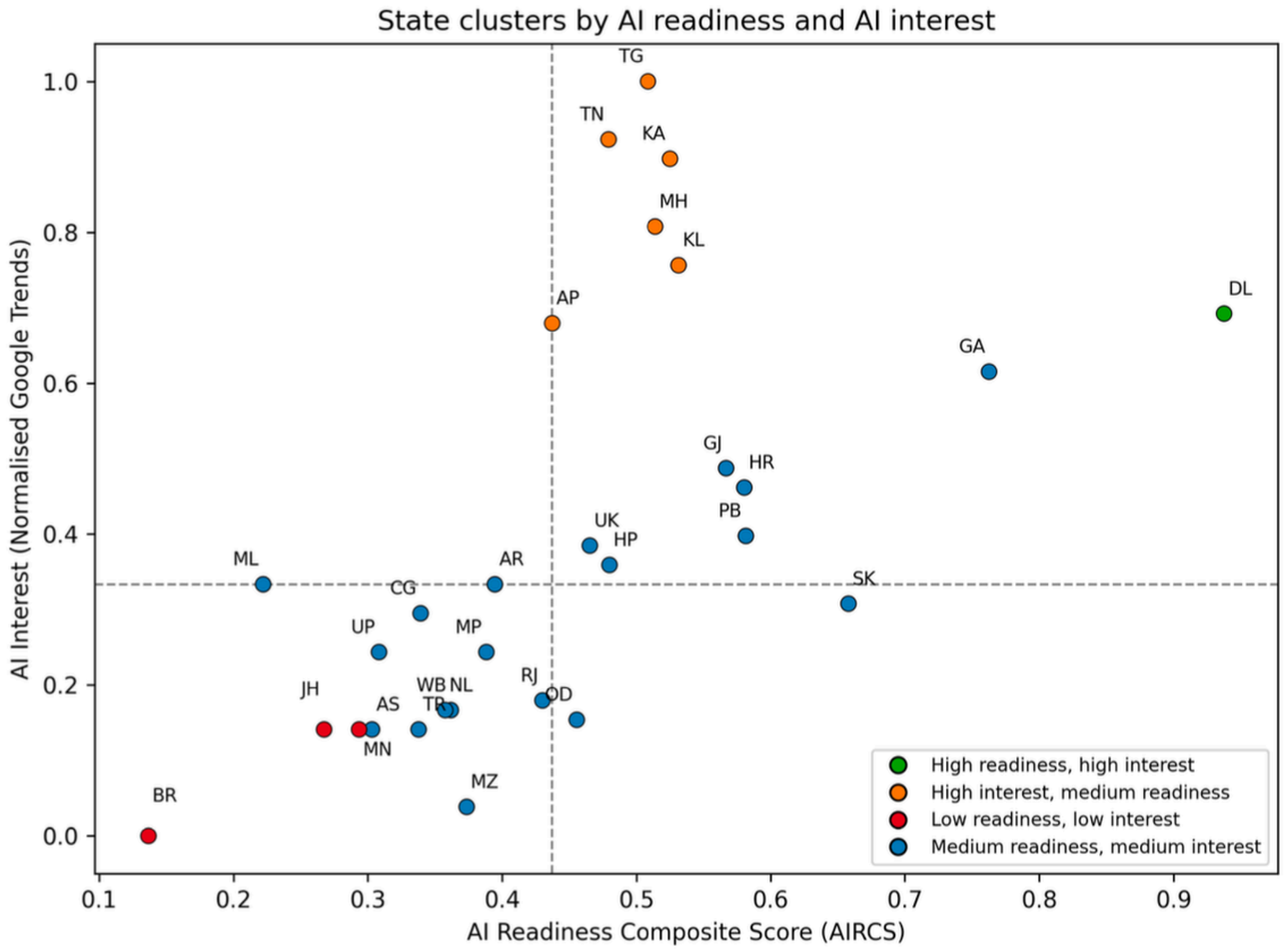
Global Awareness & Community

Events, workshops, whitepapers, and public discourse

3 Whitepapers | MoU with UNICC (UN) | MoU: Andhra Pradesh Quantum Valley | Research: IIIT Dharwad | Advisory: Prof. Prabhu, Dr Bardhan & others

Informing Policy through Data-driven Research

Quantifying Regional AI Readiness Disparities: An Empirical Analysis of Indian States



N. Sathiaseelan & J. Prabhu, Quantifying Regional AI Readiness Disparities: An Empirical Analysis of Indian States, under submission

Frugal AI Framework to Measure AI Portfolio

Capturing Frugal AI telemetry: Infrastructure, data, and models

Compute & Infrastructure

Use billing APIs (AWS CUR, GCP BigQuery) to extract per-service costs. Track Cost per Inference and Idle Cost Ratio.

Data Lifecycle

Track cloud logs for ingestion costs ($ per GB), re-indexing batch jobs, and storage costs split by raw vs. processed data.

Models & Software

Monitor API billing for proprietary LLMs (cost per token), track CPU/GPU deployment hours for open-source alternatives, and allocate platform licenses via specific resource tags.

Four optimization pathways to reduce comprehensive AI costs

ΔC_dev (Development Efficiency)

Transfer learning, automated CI/CD deployment, efficient hyperparameter search, and modular model architectures.

ΔC_ops (Operational Utilization)

Autoscaling inference endpoints, right-sizing instance types, traffic-aware elastic scaling, and batch inference consolidation.

ΔC_energy (Energy Efficiency)

Model compression (quantization, pruning, distillation), hardware acceleration per watt, and dynamic batching.

ΔC_carbon (Emissions Reduction)

Carbon-aware scheduling, infrastructure optimization, and deployment in lower-intensity grid regions.

Applying efficiency levers yields 15-44% reductions across all cost pillars

Component	Baseline AI	Frugal AI	Change
Dev + Maint Labor	$40,500	$34,425	↓ Drop 15%
Ops Compute	$8,760	$6,570	↓ Drop 25%
Energy Cost	$420.48	$236.52	↓ Drop 44%
Carbon Cost	$61.32	$34.49	↓ Drop 44%

Transfer learning reuses feature pipelines and automated CI/CD.

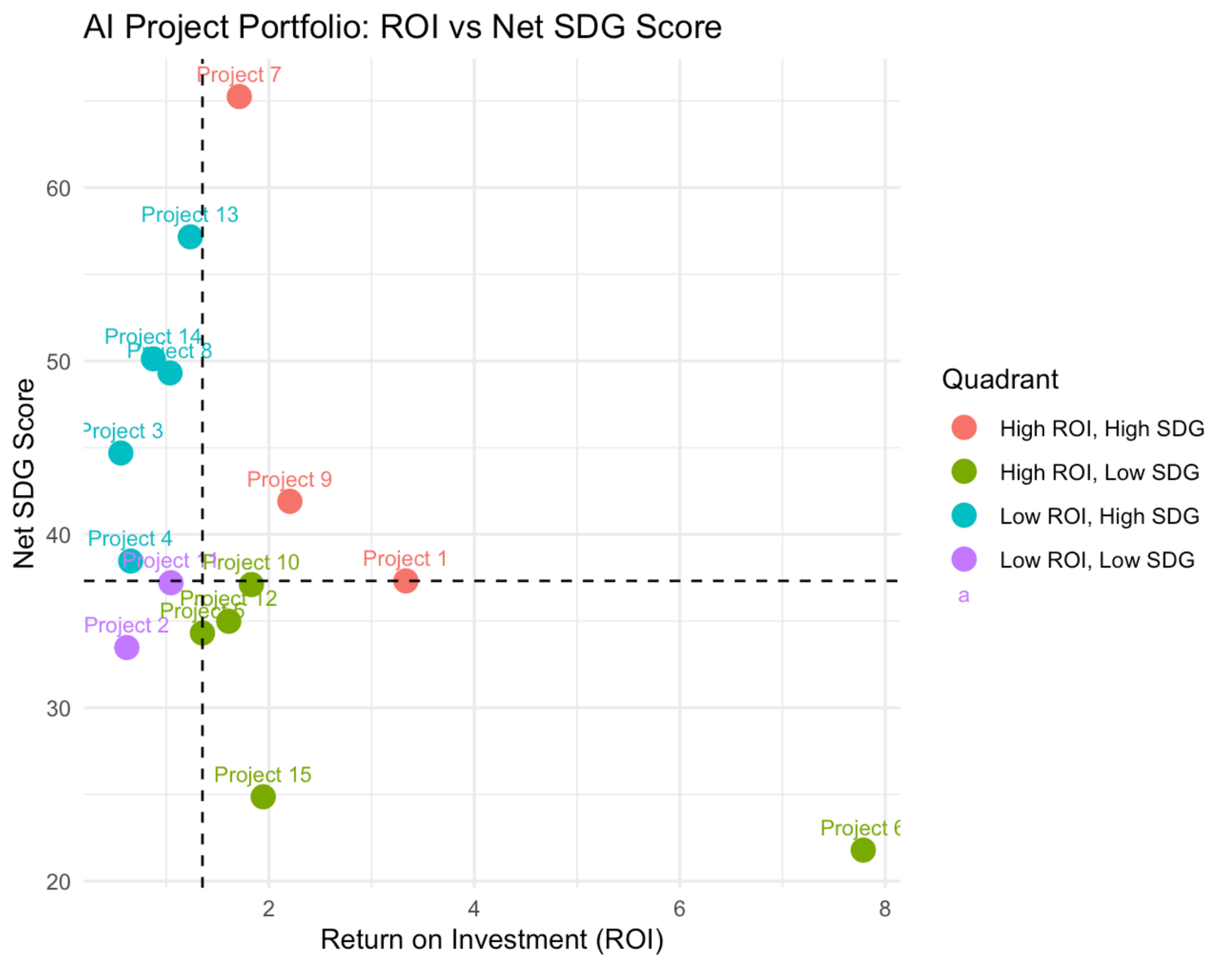
Autoscaling minimizes idle compute from 2 constant instances to 1.5 average.

Model quantization lowers power draw from 0.20 kW to 0.15 kW per instance.

Driven by reduced compute energy at $50/tCO2e.

UNICC

Modeling AI portfolios: ROI vs Net SDG Score (Illustrative)



Frugal Voice: Offline Speech AI for the Soliga Tribe, Karnataka — with IIIT Dharwad

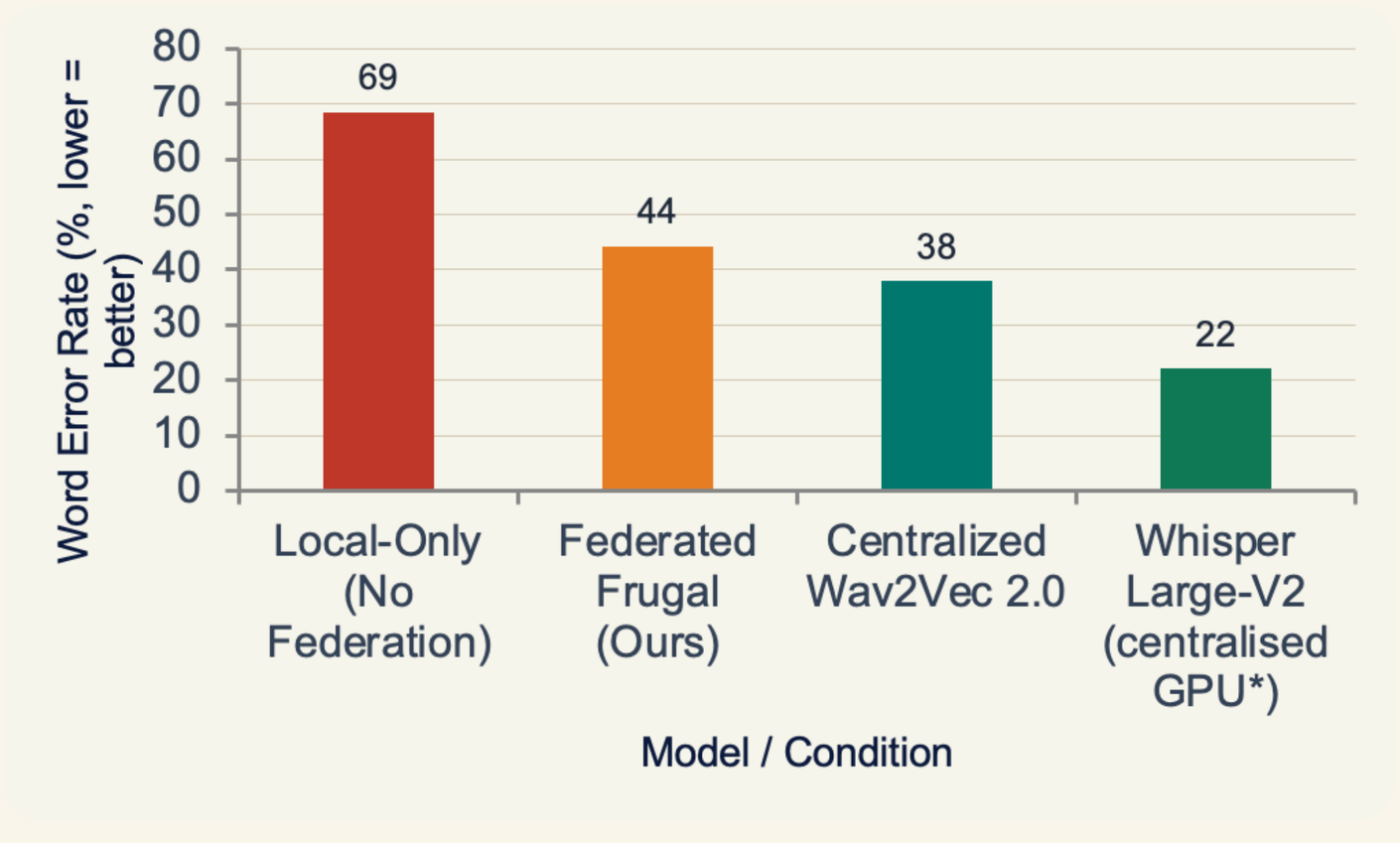
The Challenge

- Soliga: Dravidian tribal language — no standardised script
- Purely oral; no existing digital corpus
- No internet connectivity in Biligiri Rangaswamy Hills
- Historical distrust of external data collection
- Cloud AI requires constant connectivity & expensive GPUs
- No existing ASR system for this language

The Frugal AI Technical Stack

Input	5 hrs Soliga speech; 16kHz mono; Kannada transliteration by bilingual annotators
Model	Compact CNN (KWS) + small GRU/Conformer ASR; 8-bit quantized; distilled from Wav2Vec 2.0 seed
Infrastructure	Raspberry Pi + Android edge devices; community mini-server at local NGO; weeks-long offline operation
Orchestration	FedAvg with intermittent sync; sparsified 340KB updates; version-aware aggregation at village markets
Open Foundations	Wav2Vec 2.0 pre-training; community data council (OCAP-aligned); tiered consent & veto governance

Frugal Voice Results: Near-Competitive Performance, Complete Data Sovereignty



0.84

Keyword Spotting F1
(health terms: doctor, hospital, medicine)

340 KB

Model update per federated round
(vs several MB in standard FL)

12 min

Training time per round
on Raspberry Pi 4 hardware

5 hours

Total audio data used — frugal by design

** Whisper requires datacentre GPUs & 11h training data; Frugal Voice uses 5h on Raspberry Pi*

🗣️ Data Sovereignty

Voice data never left community devices

🏥 Health Access

KWS used to navigate health hotlines in Kannada

📖 Cultural Preservation

ASR enables intergenerational story archiving

A GLOBAL COMMUNITY INITIATIVE

The Saving Voices Project

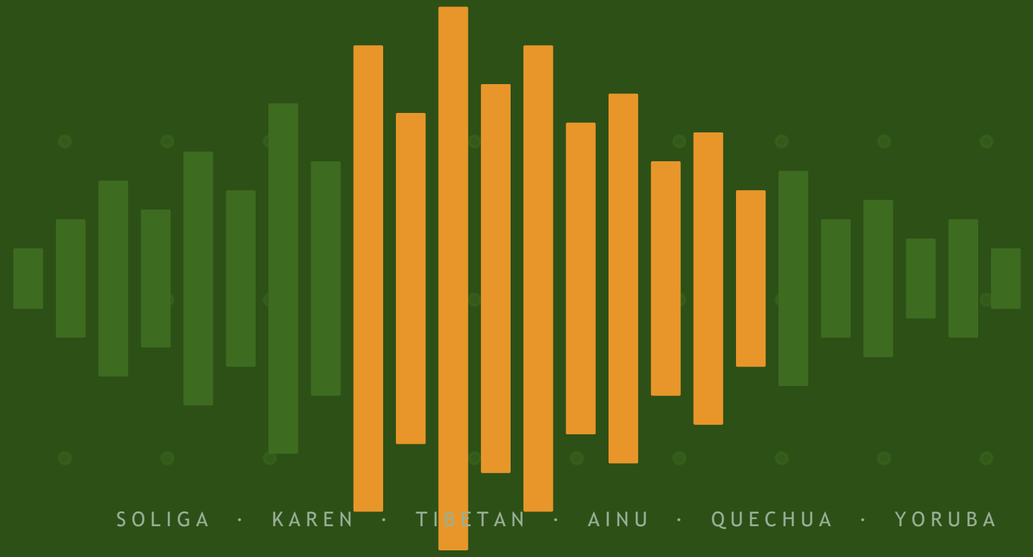
Preserving the voices of indigenous communities through frugal open-source AI

476M+
Indigenous People Globally

5,000+
Communities at Risk

90
Countries Represented

www.savingvoices.org



SOLIGA · KAREN · TIBETAN · AINU · QUECHUA · YORUBA

"Every language is a living archive.
When it falls silent, we lose a world."

— The Saving Voices Project

savingvoices@gmail.com · Cambridge Judge Business School · IIIT Dharwad · Flipped.ai

Frugal AI Hub Activities

Our Team

Serish Venkata Gandikota

Co-Founder & Co-Director

Frugal AI Hub · Visiting Fellow, Cambridge Judge Business School

Innovation strategist, impact & climate fund adviser, entrepreneur and researcher focused on frugal innovation, sustainability, and impact investing.

Elizabeth Osta

Co-Founder & Co-Director

Frugal AI Hub · Visiting Fellow, Cambridge Judge Business School

Digital and data strategist advising CXOs on AI, innovation, and responsible data use. Former Chief Data Officer at HEINEKEN.

Dr Arjuna Sathiaseelan

Chief Technology Officer

Frugal AI Hub · Visiting Fellow, Cambridge Judge Business School

Expert in inclusive connectivity. Formerly Head of Networking for Development Lab at the University of Cambridge.

Prof Jaideep Prabhu

Advisor & Mentor

Cambridge Judge Business School

Leading global voice on frugal innovation. Co-author of 'Jugaad Innovation' and 'Frugal Innovation'. Actively advises the Frugal AI Hub.

Join the Frugal AI Movement

AI doesn't have to be expensive, planet-damaging, or exclusive. Frugal AI proves a better path is possible — and it works.

Partner with Us

Establish a Frugal AI Adoption Lab in your region, sector, or organisation

Commission Research

Joint research on frugal AI applications for your specific challenge

Join the Ecosystem

Connect your startup, enterprise, or government to our global network

Pilot a Use Case

Test frugal AI on your data, under community governance, at minimal cost

frugalai.org | Cambridge Judge Business School | a.sathiaseelan@jbs.cam.ac.uk

"AI should serve humanity — not just those who can afford it."

FRUGAL AI HUB at UNIVERSITY OF CAMBRIDGE Judge Business School