

ARCHITECTURE FRAMEWORK

Frugal AI Architecture Playbook

Build and deploy high-performance AI on standard hardware — no expensive GPUs required

Frugal AI is India's paradigm for the next generation of AI deployment — smaller contextualised models, CPU-native inference, edge computing, and carbon-aware design. This playbook is based on GCF's practical experience with Kompact AI (Ziroh Labs / IIT Madras), VigyanLabs Micro Data Centres, and the Sustainable AI CoE at Sir MVSA. Ask EcoBodhai at ecobodhai.in to explore any concept interactively.

50–80%Energy Reduction vs
GPU**10×**

More Orgs Reached

20%

Infrastructure Cost

■ 65/hr

IndiaAI GPU Portal Cost

Why Frugal AI?

The dominant narrative in AI — that bigger models, more GPUs, and larger compute clusters always produce better outcomes — is false for the majority of real-world enterprise use cases. Most business AI applications — document analysis, RAG-based search, compliance checking, sustainability reporting, conversational agents — can be served by smaller, contextualised models running on modern CPUs at a fraction of the energy and cost. India's IndiaAI Mission's emphasis on 'frugal, sovereign, and scalable AI' directly validates this approach. This playbook gives you the architecture patterns to build it.

The Four Pillars of Frugal AI Architecture

P i l l a r 1

CPU-Native Inference

Modern CPUs — particularly Intel Xeon Scalable (e.g., 8581C: 60 cores, 504W TDP) and Apple M-series — can run full LLM inference without GPUs for most enterprise workloads. Kompact AI by Ziroh Labs / IIT Madras demonstrates 50–80% energy reduction vs equivalent GPU-based inference for RAG, document AI, and conversational agents.

When to use CPU inference: RAG pipelines, document analysis, conversational AI, compliance checking, sustainability reporting — any workload where latency > 200ms is acceptable

When GPU may still be needed: Real-time sub-100ms inference at very high throughput (>1,000 concurrent users), computer vision at scale, or model training/fine-tuning runs

Reference architecture: 2x Intel Xeon 8581C (60 cores each) + 512GB DDR5 ECC RAM + 8TB NVMe + 10GbE — the exact configuration of the GCF Sustainable AI CoE at Sir MVSA (■42.85L total)

P i l l a r 2

Smaller, Contextualised Models with RAG

A 7B parameter model fine-tuned on domain-specific data with a well-designed RAG pipeline will outperform a 70B general model on your specific use case — at 10–15% of the energy cost. This is the single most impactful architectural choice for sustainable AI.

Model selection framework: Start with the smallest model that achieves your accuracy threshold on a representative evaluation set. Increase model size only when smaller models demonstrably fail.

RAG architecture essentials: Document chunking (512 tokens, 64-token overlap), dense embedding (all-MiniLM-L6-v2 or IndicBERT for Indian languages), vector store (ChromaDB, FAISS, or Weaviate), retrieval (top-k=5), re-ranking (optional but recommended for precision)

Indian language models: IndicBERT v2, MuRIL, Gemma 2B (multilingual) for embedding; Bhashini ASR for voice input; BharatGen Param2 (IndiaAI Mission) for generation

P i l l a r 3

Micro Data Centre Design

VigyanLabs IPM+ Micro Data Centres achieve PUE of 1.2 — compared to the Indian average of 1.8. This means 33% less energy wasted on cooling and infrastructure. For organisations not ready for a full data centre, a single Micro DC node provides on-premise AI capability with full data sovereignty and DPDP Act compliance.

VigyanLabs IPM+ key specs: PUE 1.2 · Scalable from 5kW to 50kW · Liquid cooling optional · Deployable in standard office space · Edge-ready form factor

When to choose on-premise vs cloud: On-premise: sensitive data (health, legal, financial), DPDP compliance requirements, predictable high-volume inference. Cloud: burst capacity, variable workloads, global distribution

Carbon advantage: On-premise with renewable energy procurement at PUE 1.2 beats major cloud providers on carbon intensity per kWh for collocated workloads in most Indian geographies

P i l l a r 4

Carbon-Aware Computing

The carbon intensity of India's electricity grid varies significantly by time of day and by state — ranging from ~0.4 kgCO₂/kWh in hydro-rich states to ~0.9 in coal-heavy grids. Scheduling non-time-sensitive AI workloads to run during low-carbon windows can reduce training and batch inference carbon by 30–50% with zero model changes.

Carbon Aware SDK: Open source library from the Green Software Foundation — integrates with any scheduler to shift workloads based on real-time grid carbon intensity data

India-specific grid data: CEA (Central Electricity Authority) publishes state-wise emission factors. Karnataka (solar + hydro heavy) is consistently one of the lowest-carbon grids in India.

BirchLogic SustainIT integration: Real-time AI carbon tracking per inference, per training run, and per deployment — integrated with MLflow and standard CI/CD pipelines. Used in all GCF CoE workloads.

Frugal AI Stack — Reference Architecture

The following table describes the recommended technology stack for a production Frugal AI deployment — as used in the GCF Sustainable AI CoE at Sir MVSA, Bangalore.

Layer	Technology	GCF Recommendation	Why Frugal
Inference Runtime	Kompact AI (Ziroh Labs/IIT Madras)	Native for CPU inference	50–80% energy reduction vs GPU
Foundation Models	Gemma 2B / Phi-3 Mini / Bhashini	Small; scale up only if needed	1–7B params; runs on CPU
Embedding	all-MiniLM-L6-v2 / IndicBERT	Indian language support built in	Runs on any modern CPU
Vector Store	ChromaDB / FAISS	Open source; on-premise	No cloud dependency; DPDP safe
MLOps	MLflow + BirchLogic	Track accuracy AND carbon per run	Prevents runaway experiment energy
Data Centre	VigyanLabs IPM+ Micro DC	PUE 1.2; edge-deployable	33% less cooling energy vs average
Carbon Tracking	BirchLogic SustainIT	Real-time kgCO ₂ e per inference	Makes sustainability measurable
Carbon-Aware Scheduler	Carbon Aware SDK (Green Software Factory)	Software Factory-based scheduling	30–50% training carbon reduction

Try it now: The EcoBodhai AI app at ecobodhai.in runs on the Frugal AI stack described in this playbook — CPU-native inference, contextualised models, RAG. It is completely free and requires no sign-up. Experience Frugal AI in action.