Frugal AI Hub
White paper · May 2026

APPLICATIONS OF FRUGAL AI IN THE FINTECH INDUSTRY

Measuring Total Cost of Ownership and Return on Investment in real-time fraud detection

FRUGAL AI HUB at UNIVERSITY OF CAMBRIDGE Judge Business School

Authors

Elizabeth Osta
Co-founder, Frugal AI Hub · Visiting Fellow, Cambridge Judge Business School
e.osta@jbs.cam.ac.uk

Serish Venkata Gandikota
Co-founder, Frugal AI Hub · Visiting Fellow, Cambridge Judge Business School
s.gandikota@jbs.cam.ac.uk

Arjuna Sathiaseelan
Technology Lead, Frugal AI Hub · Visiting Fellow, Cambridge Judge Business School
a.sathiaseelan@jbs.cam.ac.uk

Jaideep Prabhu
Professor of Marketing & Nehru Professor of Indian Business, Cambridge Judge Business School
j.prabhu@jbs.cam.ac.uk

About the Frugal AI Hub

The Frugal AI Hub at Cambridge Judge Business School is a research and practice hub advancing the principles, methods and adoption of Frugal AI — intelligent systems that are high performing yet resource efficient, and that can be understood and governed by the institutions that deploy them. The Hub works with industry, public sector and civil society partners across sectors including fintech, healthcare, education and sustainability.

Abstract

Fintech firms operate in an environment characterised by tight margins, heavy regulation and rapidly evolving customer expectations. AI has become a central tool for differentiation, but the full costs — financial and environmental — of AI systems are rarely measured at the level of individual use cases. This paper applies the Frugal AI measurement framework to a portfolio of fraud detection use cases, focusing on how Total Cost of Ownership (TCO) and Return on Investment (ROI) can be extended to include energy consumption and emissions while remaining operationally tractable.

Frugal AI is an intentional design philosophy and operational framework that prioritises doing more with less — building intelligent systems that are high performing yet resource efficient, and that can be understood and governed by the institutions that deploy them. This paper operationalises that framework within a real-time credit card fraud detection system, demonstrating that Frugal AI approaches can reduce TCO by 17% and improve ROI by 24.6% while simultaneously cutting energy use by 44% and carbon emissions by 44%.

> *"Frugal AI is not merely a cost-cutting exercise. It is a reframing of what good AI looks like — systems that are not only intelligent but affordable, sustainable, and inclusive by design."*

1. Introduction

Fintech's adoption of AI is typically compute and data intensive, which drives up both operational and sustainability costs. The prevailing approach relies on large scale cloud infrastructure or proprietary hardware platforms, which delivers performance but at high financial and environmental cost. The main cost drivers include high utilisation of GPUs and TPUs for training and inference, the acquisition and retention of scarce AI talent, substantial data management and storage overheads, and the increasing regulatory and governance burden associated with deploying models in a compliant and auditable manner.

1.1 Background and motivation

Existing work on Green AI and sustainable AI has highlighted the environmental burden of large models and proposed high level principles for mitigating that impact. Fintech offers a particularly salient context for applying these principles because it combines stringent performance, latency, and governance requirements with explicit cost and sustainability pressures.

The Frugal AI Hub measurement framework provides a useful starting point because it defines TCO in a way that includes energy and carbon costs and links financial ROI to social impact metrics aligned with the UN Sustainable Development Goals.

1.2 Problem statement

The core problem we address is the absence of a standardised, granular method for measuring the full cost of AI systems in fintech — including both financial and environmental dimensions. In practice, energy consumption and carbon emissions are rarely attributed at the project or use-case level, which means that AI investment decisions are made on the basis of an incomplete picture of total cost.

1.3 Research objectives

This paper builds on an existing Frugal AI measurement framework that links total cost of ownership (TCO), return on investment (ROI) and social and environmental impact across three layers. We ask three guiding questions: How can TCO for AI systems in fintech be defined comprehensively yet remain practical? How do Frugal AI optimisations change the ROI profile? And how can these metrics be connected to SDG-aligned reporting?

2. Methodology

The methodology combines qualitative insight into institutional priorities with a quantitative cost and emissions model applied to a representative fraud detection use case. The methodology was developed in collaboration with a large United States-based credit card issuer that operates a digital-first business model and relies heavily on advanced analytics across customer acquisition, credit risk, fraud prevention, and customer management.

2.1 Use case: credit card fraud detection

Our focal use case is a real-time fraud detection system at the application stage, where each incoming application is scored for fraud risk and subjected to graduated, risk-based friction such as additional identity verification or application decline. The system must simultaneously achieve: strong fraud capture and business performance; low latency and infrastructure efficiency; and robust governance and auditability.

2.2 Integrating Frugal AI metrics into TCO

We operationalise layers 1 and 2 of the Frugal AI measurement stack by defining TCO and ROI at the level of a specific fraud detection use case. TCO for the AI system is expressed as a sum of four distinct cost components, each of which is directly influenced by model design and deployment strategy:

$$\mathbf{TCO_{AI} = C_{dev} + C_{ops} + C_{energy} + C_{carbon}}$$

Cost Component	Linked Metrics	Description
Development Cost (C_{dev})	Developer hours, retraining cycles	Influenced by model complexity, retraining frequency, automation, and reuse.
Operational Cost (C_{ops})	Cost per inference, idle compute ratio	Captures infrastructure inefficiencies; reduced through elastic scaling and utilisation optimisation.
Energy Cost (C_{energy})	kWh per inference	Derived from compute intensity and runtime, reduced via compression and efficient architectures.
Carbon Cost (C_{carbon})	tCO_2e × carbon price	Monetises emissions footprint attributable to workload energy consumption.

Table 1. TCO cost components and linked Frugal AI metrics.

Note on energy and carbon attribution

In traditional fintech cost accounting, energy (C_{energy}) and carbon externalities (C_{carbon}) are rarely calculated at the project or use-case level. Instead, they are typically estimated in aggregate at the corporate or divisional level. This practice can lead to underestimation of the full economic footprint of individual AI workloads and overestimation of ROI due to an incomplete cost denominator. The Frugal AI framework explicitly models all four components to provide a more complete and defensible estimation of project-level ROI.

2.3 TCO with Frugal AI integration

Frugal AI introduces explicit efficiency gains across all four cost components. The Frugal TCO formula becomes:

$$TCO_{Frugal} = (C_{dev} - \Delta C_{dev}) + (C_{ops} - \Delta C_{ops}) + (C_{energy} - \Delta C_{energy}) + (C_{carbon} - \Delta C_{carbon})$$

2.4 Frugal AI cost reduction mechanisms

The four Δ terms represent distinct but interrelated efficiency pathways that can be operationalised by different technical and governance teams. Together, they form the structural basis for making AI more affordable, energy-efficient, and environmentally accountable.

ΔC_{dev} — Development efficiency

- Transfer learning & model reuse
- Automated CI/CD deployment
- Modular architectures
- Efficient hyperparameter search
- Reduced retraining frequency

ΔC_{ops} — Operational efficiency

- Autoscaling inference endpoints
- Right-sizing instance types
- Batch inference consolidation
- Eliminating idle capacity
- Traffic-aware elastic scaling

ΔC_{energy} — Energy efficiency

- Model quantisation and pruning
- Efficient architecture selection
- Hardware acceleration per watt
- Dynamic batching
- Reduced retraining frequency

ΔC_{carbon} — Emissions reduction

- Reduced compute intensity
- Carbon-aware scheduling
- Lower-intensity region deployment
- Infrastructure optimisation
- Renewable energy scheduling

3. Step-by-step Frugal AI cost model

The numerical demonstration below is a modelled scenario based on realistic institutional parameters. All values are illustrative and intended to make the methodology concrete rather than to report proprietary internal financials.

Model assumptions

Fraud events prevented	600 / year
Average net loss avoided	$500
Loaded labour rate	$150 / hour
Baseline labour hours	270 / year (Frugal: −15%)
Compute instances	2 baseline / 1.5 Frugal @ $0.50/hr
Power draw	0.20 kW baseline / 0.15 kW Frugal
Electricity price	$0.12 / kWh
Grid intensity	0.35 $kgCO_2e$ / kWh
Carbon price	$50 / tCO_2e

STEP 1

Development & maintenance labour cost

Baseline: 270 hrs × $150 = $40,500

Frugal: 270 × 0.85 × $150 = $34,425 | **Savings ΔC_{dev} = $6,075**

STEP 2

Operational compute cost

Baseline: 2 × $0.50 × 8,760 hrs = $8,760

Frugal: 1.5 × $0.50 × 8,760 = $6,570 | **Savings ΔC_{ops} = $2,190**

STEP 3

Energy consumption and cost

Baseline: 2 × 0.20 kW × 8,760 = 3,504 kWh → $420.48

Frugal: 1.5 × 0.15 kW × 8,760 = 1,971 kWh → $236.52 | **Savings ΔC_{energy} = $183.96**

STEP 4

Carbon emissions and carbon cost

Baseline: 3,504 kWh × 0.35 = 1.226 tCO_2e → $61.32

Frugal: 1,971 kWh × 0.35 = 0.690 tCO_2e → $34.49 | **Savings ΔC_{carbon} = $26.83**

STEP 5

Total Cost of Ownership (TCO)

Baseline TCO = $40,500 + $8,760 + $420.48 + $61.32 = **$49,741.80**

Frugal TCO = $34,425 + $6,570 + $236.52 + $34.49 = **$41,266.01** | **Reduction: 17.0%**

STEP 6

ROI calculation

Baseline ROI = ($300,000 − $49,741.80) / $49,741.80 = **5.03×**

Frugal ROI = ($300,000 − $41,266.01) / $41,266.01 = **6.27×**

ROI improvement: +24.6%

Summary results

Component	Baseline AI	Frugal AI	Change
Fraud prevented value	$300,000	$300,000	—
Dev + maintenance labour	$40,500	$34,425	↓ 15%
Operational compute cost	$8,760	$6,570	↓ 25%
Energy cost	$420.48	$236.52	↓ 44%
Carbon cost	$61.32	$34.49	↓ 44%
Total Cost of Ownership	**$49,741.80**	**$41,266.01**	**↓ 17.0%**
ROI	**5.03×**	**6.27×**	**↑ 24.6%**

Table 2. Baseline vs Frugal AI — TCO and ROI comparison.

4. Frugal AI telemetry: measuring TCO in practice

To make the proposed metrics actionable, institutions need a practical way to collect and attribute the underlying data that feeds each TCO component. The telemetry options outlined here illustrate how existing cloud-native tooling can be repurposed for Frugal AI measurement without requiring significant new infrastructure investment.

Compute & infrastructure

- Cloud billing APIs (AWS CUR, GCP BigQuery) for per-service cost attribution
- CloudWatch metrics: cost per inference, idle compute ratio
- Prometheus / SLURM / Kubernetes for on-premises node-level metrics

Energy & carbon

- AWS Customer Carbon Footprint Tool (CCFT) — account-level allocation
- Top-down: allocate CCFT emissions by cost or compute-hours
- Bottom-up: per-instance power models for granular attribution
- Key metrics: kWh per inference, gCO_2e per inference

Model performance

- Accuracy–Latency Index: normalised ratio for frugal vs non-frugal models
- Tokens per Joule: LLM efficiency measure
- CloudWatch: P50/P95/P99 latency, inference count, error rates

Governance & risk

- GRC tools (ServiceNow, Archer) for compliance cost tracking
- Metric: % of spend vs avoided incidents/fines
- Security audit costs allocated at project level via tagging

Key design principle: use tags everywhere

All cost and utilisation attribution hinges on consistent use of project/model tags across cloud resources. Without tags, accurate per-workload TCO attribution is not possible. Embedding TCO-Frugal metrics into governance requires cross-functional ownership spanning engineering, finance, and sustainability teams.

5. From financial ROI to SDG-linked portfolio impact

The Frugal AI Hub framework establishes a three-layer measurement stack: Total Cost of Ownership → Financial ROI → Social Impact. The metrics developed in this paper operate at the TCO and ROI layers. However, they form the foundation for SDG-aligned reporting at the institutional level — connecting individual project-level efficiency gains to broader societal outcomes.

Fintech outcome	SDG link	Mechanism
Fraud loss reduction	SDG 8	Protects economic productivity
System integrity	SDG 16	Strengthens institutional trust
Infrastructure efficiency	SDG 9	Improves digital infrastructure resilience
Energy reduction	SDG 13	Lowers emissions footprint
Resource productivity	SDG 12	Reduces wasteful compute

Table 3. Fintech outcomes mapped to UN Sustainable Development Goals.

The Frugal AI value flywheel

Efficiency → Accessibility → Inclusion → Innovation → Sustainability → Efficiency

Each improvement in cost efficiency lowers barriers to adoption, enabling more inclusive deployment, which in turn fuels broader innovation and strengthens the sustainability case for responsible AI at scale.

6. Conclusion

This paper has shown how Frugal AI can be made operational in a concrete fintech setting by expressing total cost of ownership and return on investment at the level of an individual fraud detection use case. By including energy and carbon in the cost denominator, the framework produces a more defensible and comprehensive measure of AI performance — one that does not overstate returns by excluding environmental externalities.

For fintech practitioners, the immediate implication is that Frugal AI need not be treated as a separate sustainability initiative, but can be embedded within existing risk, investment and model governance processes. For policymakers and industry bodies, the framework illustrates how standardised, project-level energy and carbon reporting can be made tractable using existing cloud-native tooling.

While our analysis focuses on a single credit card issuer and on application-stage fraud detection, the underlying methodology is designed to be portable to other AI-enabled services such as transaction monitoring, credit decisioning, and customer engagement — and to other institutions operating under comparable performance and compliance constraints.

References

Frugal AI Hub, 2025. *Accelerating a Frugal AI Ecosystem.* Frugal AI Hub, Cambridge Judge Business School.

Frugal AI Hub and UNICC, 2025. *From Total Cost of Ownership to Social Impact: A Frugal AI Framework to Measure AI Portfolios as Strategic Assets.*

AFNOR, 2023. *AFNOR SPEC 2314: Frugal Artificial Intelligence — Guidelines for the Design and Deployment of Frugal AI Systems.*

Strubell, E., Ganesh, A. & McCallum, A., 2019. Energy and Policy Considerations for Deep Learning in NLP. *ACL*.

Schwartz, R., Dodge, J., Smith, N.A. & Etzioni, O., 2020. Green AI. *Communications of the ACM*, 63(12), pp.48–54.

Patterson, D., et al., 2021. *Carbon Emissions and Large Neural Network Training.* arXiv.

Greenhouse Gas Protocol, 2011. *Corporate Accounting and Reporting Standard.*

Pigou, A.C., 1920. *The Economics of Welfare.*

Han, S., et al., 2015. Deep Compression. *ICLR*.

ISO/IEC 42001:2023. *Information Technology — Artificial Intelligence — Management System.*

Henderson, P., et al., 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *JMLR*.

Bolton, R.J. & Hand, D.J., 2002. Statistical fraud detection: A review. *Statistical Science*, 17(3), pp.235–249.

Frugal AI Hub
at Cambridge Judge Business School

Trumpington Street
Cambridge CB2 1AG
United Kingdom

https://frugalai.org