AINNA

NeuralOps ESG Efficiency Methodology & Evidence Report

Technical reference for token, energy and carbon-reduction claims

PUBLIC METHODOLOGY NOTE

Published	10 August 2026
Version	1.1
Scope	AINNA NeuralOps smart routing, detached systems, published efficiency claims, and documented industry coverage
Evidence status	Internal arithmetic verification and public methodology review
Assurance status	Not independently assured; not a certified carbon audit

Key verified result

The published case-study token reduction is reproducible: (7.5M - 1.0M) / 7.5M = 86.67%, reasonably rounded to 87%. The energy and CO_2e reduction percentages are also arithmetically consistent with the study inputs, but the absolute energy and emissions values remain estimates until verified against infrastructure telemetry.

AINNA NeuralOps - Efficiency before scale

1. Executive Summary

AINNA NeuralOps is presented as an ESG-aligned AI infrastructure approach that reduces unnecessary GPU-intensive processing through smart routing, workload segmentation, deterministic/rule-based processing, detached systems and AI guardrails. The public ESG page links the headline efficiency claim to a separate controlled internal study rather than presenting it as a universal reduction rate.

This report validates the arithmetic behind the published 87% reduction claim, explains an auditable calculation method for energy and operational carbon, identifies where claims are estimates rather than measurements, and defines wording suitable for public disclosure.

Conclusion in one sentence

The 87% headline is defensible as a workload-specific internal efficiency result; it should be described as an observed/calculated token reduction in the tested scenario, while kWh and CO2e remain estimated until actual infrastructure measurements are used.

Evidence classification

Claim	Published inputs	Arithmetic check	Evidence status
Token reduction	7.5M -> 1.0M	86.67%	Verified arithmetic; workload-specific
Energy reduction	55.0 -> 7.3 kWh	86.73%	Verified arithmetic; input energy values estimated
CO2e reduction	24.0 -> 3.2 kg	86.67%	Verified arithmetic; absolute emissions depend on carbon factor
Energy saved	55.0 - 7.3 kWh	47.7 kWh	Verified arithmetic
CO2e avoided	24.0 - 3.2 kg	20.8 kg (~21 kg)	Verified arithmetic; estimate, not audit

Scope note: This report does not independently reproduce model runtime, GPU telemetry, cloud billing, power draw or data-centre metering. It verifies the published arithmetic and provides a methodology for moving from internal estimates to auditable operational emissions.

2. What NeuralOps Changes

The environmental rationale is architectural: reduce the share of work that reaches high-compute inference, and reserve large-model reasoning for tasks that genuinely require it.

Smart Routing

Classify requests and send each request to the lowest-compute processing layer that can meet the required accuracy, latency and governance controls.

Detached Systems

Move repetitive production workflows into deterministic services, database logic, scheduled jobs or local automation that can run without recurring LLM calls.

Deterministic Parsing / Rules

Use structured parsers and validation rules for supported inputs, with AI used only for ambiguous or unsupported cases.

GPU Only When Needed

Escalate complex reasoning and interpretation to high-performance models rather than making GPU-heavy inference the default path.

AI Guardrails

Reduce retries, malformed outputs, repeated context and unnecessary computational cycles through validation, format constraints and fallback logic.

Functional principle

AI builds or assists the system; the production system should use AI only where the value of inference exceeds the compute cost and risk. This is consistent with the ESG page statement that NeuralOps prioritises efficient compute utilisation rather than brute-force AI processing.

3. Published Case Study: 87% Reduction Proof

The public Token Saving Study compares two approaches for 100 SME bank-statement sets: a repeated full AI-agent workflow versus a detached-system architecture with smart routing and selective AI.

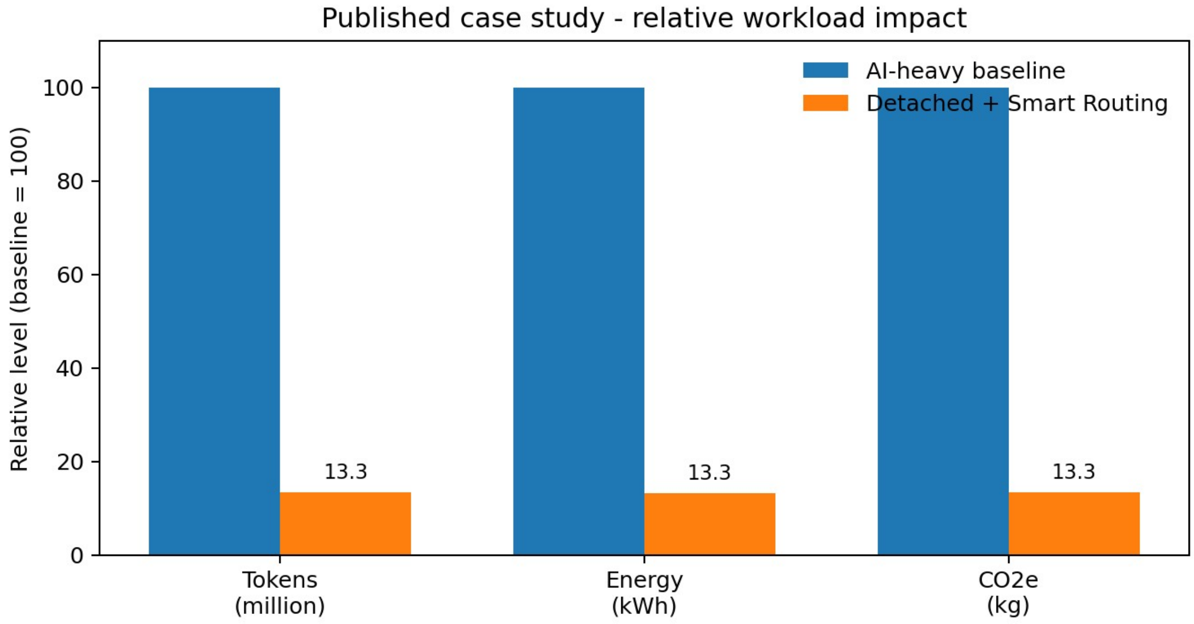

Figure 1. Relative published workload values. Baseline is normalised to 100 for visual comparison.

3.1 Token calculation

Published values: baseline = 7.5 million tokens; optimized = 1.0 million tokens.

```
Token reduction (%) = (Baseline - Optimized) / Baseline x 100
= (7.5M - 1.0M) / 7.5M x 100 = 86.67% ~= 87%
```

Result: The 87% token-reduction headline is mathematically correct for the published case-study inputs.

3.2 Energy calculation

Published values: baseline = 55.0 kWh; optimized = 7.3 kWh.

```
Energy reduction = (55.0 - 7.3) / 55.0 x 100 = 86.73%
```

The published 87% energy-reduction percentage is therefore arithmetically consistent. However, this check validates the percentage only; it does not independently validate that 55.0 kWh and 7.3 kWh were directly metered values.

3.3 Operational carbon calculation

Published values: baseline = 24.0 kg CO_2e; optimized = 3.2 kg CO_2e.

```
CO2e reduction = (24.0 - 3.2) / 24.0 x 100 = 86.67%
```

The difference is 20.8 kg CO_2e, which supports the study wording of approximately 21 kg CO_2e avoided.

Important distinction

A reduction in token volume can reduce compute demand, but token reduction is not automatically identical to energy or carbon reduction across all models and hardware. For public disclosure, token savings should be measured directly; energy and carbon should be calculated from runtime/power data and location-specific carbon intensity whenever available.

4. Recommended Carbon Accounting Method

AINNA already exposes the main variables required for a transparent carbon emulator: workload volume, routing mix, per-route energy intensity, grid emission factor and data-centre PUE. The following method converts those inputs into an auditable scenario estimate.

4.1 Workload volume

```
Monthly requests = Users x Requests per user per day x Processing days
```

If token-based reporting is required, add: Total tokens = Monthly requests x average tokens per request, or preferably use observed provider/local inference token logs.

4.2 Routed workload

```
Requests(route) = Monthly requests x Route share
```

Route shares should total 100%. Suggested route categories are GPU-heavy AI, light AI/CPU, rule-based/parsing and detached systems.

4.3 IT and facility energy

```
IT energy = Sum[(Requests(route) / 1,000) x kWh per 1,000 requests(route)]
```

```
Facility energy = IT energy x PUE
```

PUE is the ratio of total data-centre energy to IT equipment energy. If the kWh intensity already includes total facility overhead, PUE must not be applied again.

4.4 Operational carbon

```
Operational CO2e = Facility energy (kWh) x Grid emission factor (kg CO2e/kWh)
```

```
Carbon reduction (%) = (Baseline CO2e - Optimized CO2e) / Baseline CO2e x 100
```

This is compatible with the core logic of Software Carbon Intensity (SCI), which separates operational energy, carbon intensity, embodied emissions and a functional unit. AINNA can later extend the method to embodied hardware emissions if a fuller SCI-style boundary is required.

5. Consistency Checks and Corrections

A public methodology report should also disclose where current website values require clarification. This increases credibility because visitors can distinguish verified arithmetic from assumptions.

Item	Finding	Recommended public treatment
87% token reduction	7.5M -> 1.0M gives 86.67%.	Keep. State "up to 87% in the tested workload".
Energy reduction	55.0 -> 7.3 kWh gives 86.73%.	Keep percentage, but label energy values as estimates unless telemetry is available.
CO2e values	24/55 implies ~0.436 kg CO2e/kWh; 3.2/7.3 implies ~0.438.	Publish the grid factor used for the study and use the same boundary on baseline and optimized scenarios.
Carbon emulator default	The public calculator displays a default grid factor of 0.74 kg CO2e/kWh and PUE 1.4.	Explain source/date for defaults and allow users to replace them with local/provider values.
API cost figures	The public token study includes API-cost figures that do not reconcile with the token-price assumptions shown in its methodology table.	Do not cite these figures in the ESG report until corrected and re-validated.
Live ESG dashboard scores	Scores such as "ESG Readiness" and "Responsible AI Score" are displayed without a public calculation method.	Publish formula, data source and update cadence, or relabel as internal indicators.
"Compliance" wording	The ESG page lists EECA, NETR, Bursa, IFRS, TCFD, GRI and SDGs.	Use "alignment/reference frameworks" unless a formal compliance assessment or assurance has been completed.

Why the carbon factor matters

Using the study values, the implied operational carbon factor is about 0.437 kg CO2e/kWh. The current emulator displays a default of 0.74 kg CO2e/kWh. These can both be valid only if they refer to different sources, periods or accounting boundaries; the public report should identify which factor is used for each result.

6. Secondary Validation from the Published Assumptions Table

The study also publishes a base-case assumption of 100 SMEs, 12 statements per SME per year, approximately 8,500 tokens per statement for the AI-heavy approach and approximately 1,100 tokens per statement for the optimized approach.

```
Annual baseline tokens = 100 x 12 x 8,500 = 10.20M

Annual optimized tokens = 100 x 12 x 1,100 = 1.32M

Annual token reduction = (10.20M - 1.32M) / 10.20M x 100 = 87.06%
```

This separate assumption set also produces approximately 87% token reduction. It therefore supports the direction and magnitude of the headline result, even though its absolute annual token totals differ from the 7.5M versus 1.0M headline case.

7. Documented Industry Coverage

The AINNA Cyberpunk navigation currently lists 32 distinct industry verticals under its Industry section. The separate “Industries” item is a landing/index page and is not counted as a 33rd vertical. This section records the published coverage scope and the typical NeuralOps applicability for each vertical.

32	**Industry verticals**	1	**Shared NeuralOps architecture**

Coverage principle: the same orchestration layer - Smart Routing, Detached Systems, deterministic processing, validation and selective LLM escalation - can be adapted across different operational domains without making large-model inference the default path.

#	Industry vertical	Typical NeuralOps applicability
1	**Palm Oil Intelligence**	Plantation monitoring, yield analysis, maintenance and operational decision support.
2	**Education**	Administration, learning support, institutional knowledge and workflow automation.
3	**Financial Technology**	Financial processing, reconciliation, analytics, risk controls and automation.
4	**Manufacturing**	Production monitoring, quality control, maintenance and workflow optimisation.
5	**Immersive Entertainment**	Interactive content operations, experience orchestration and digital asset workflows.
6	**Energy Trading**	Market-data analysis, monitoring, forecasting support and controlled decision workflows.
7	**Professional Services**	Knowledge automation, document processing, research and client-service workflows.
8	**Legal AI**	Document intelligence, controlled research, review workflows and auditability.
9	**Education AI**	AI-assisted learning, content workflows and institutional intelligence.
10	**Mining & Resources**	Resource monitoring, operational intelligence, maintenance and analytics.
11	**Fleet & Delivery**	Fleet utilisation, routing, delivery monitoring and exception management.
12	**Property Technology**	Property data, asset operations, transaction workflows and automation.
13	**Oil & Gas**	Asset monitoring, engineering data, maintenance and operational intelligence.
14	**Autonomous Transport**	Mobility intelligence, sensing workflows, monitoring and controlled automation.
15	**Hospitality**	Reservations, customer operations, resource planning and service automation.
16	**Marine AI Systems**	Marine operations, equipment-condition intelligence and maintenance workflows.
17	**Insurance Technology**	Claims processing, document intelligence, risk workflows and automation.
18	**Healthcare**	Operational data intelligence, scheduling, documentation and governed workflow support.
19	**Retail & Grocery**	Inventory, sales, pricing, demand and customer-operation intelligence.
20	**IT Solutions**	Infrastructure operations, service automation, monitoring and system intelligence.
21	**Property Management**	Tenant, maintenance, asset and facilities workflow management.

#	Industry vertical	Typical NeuralOps applicability
22	**Renewable Energy**	Renewable-asset monitoring, performance analysis and efficiency intelligence.
23	**Entertainment**	Content operations, audience intelligence and workflow automation.
24	**Workshop & Automotive**	Workshop operations, maintenance records, diagnostics support and scheduling.
25	**Pharmacy**	Inventory, operations, documentation and controlled information workflows.
26	**Smart Supply Chain**	Demand, inventory, procurement and supply-chain optimisation.
27	**Logistics & Supply Chain**	Warehouse, transport, fulfilment, tracking and exception intelligence.
28	**Agriculture**	Farm monitoring, productivity, resource use and operational optimisation.
29	**Construction**	Project progress, materials, documentation and site-operation monitoring.
30	**E-Commerce**	Orders, inventory, marketplaces, reconciliation and commerce automation.
31	**Food & Beverage**	Inventory, production, sales, quality and operational optimisation.
32	**Medical Technology**	Medical-device engineering, technical documentation, data and controlled workflows.

Boundary note: inclusion in this table means the vertical is documented in AINNA's public Cyberpunk menu and is considered architecturally applicable. It does not, by itself, demonstrate a live production deployment, customer engagement, or verified ESG saving in every listed industry.

ESG relevance: a shared routing-and-detached-system architecture can reduce duplicated AI infrastructure and avoid unnecessary high-compute inference across multiple domain workflows. Any quantified energy or carbon benefit should still be measured per deployment using the methodology in Section 4.

8. ESG Relevance and Reporting Boundary

Environmental

- Primary measurable lever: lower computational workload reaching energy-intensive inference.
- Recommended metrics: tokens/workflow, GPU-seconds/workflow, IT kWh/workflow, facility kWh/workflow and kg CO_2e/workflow.
- Do not treat architectural efficiency as an offset. Report actual reductions in compute, energy and emissions.

Social

- Efficiency can reduce infrastructure and inference cost, supporting more affordable AI adoption by SMEs.
- Social claims should be supported with adoption, accessibility, pricing or SME outcome data rather than inferred from environmental efficiency alone.

Governance

- Maintain traceable routing decisions, model selection logs, retry counts, parser/fallback outcomes and versioned assumptions.
- Publish the difference between measured telemetry, calculated values and illustrative scenario estimates.
- Keep audit trails for calculation inputs, emission factors and PUE sources.

Recommended disclosure labels

MEASURED = captured from system/provider telemetry. CALCULATED = derived from measured inputs using a published formula. ESTIMATED = derived from assumptions or research intensity factors. INDICATIVE = score or proxy that is not a direct physical measurement.

9. Recommended Public Wording

The following wording is suitable for the ESG page, investor material or a download link because it preserves the strength of the result without overstating assurance:

> *"AINNA NeuralOps achieved up to 87% token reduction in a controlled internal comparison of 100 SME bank-statement sets, calculated from approximately 7.5 million baseline tokens versus 1.0 million tokens using detached systems and smart routing. Energy and CO2e figures are scenario estimates unless supported by infrastructure telemetry, and vary by model, hardware, PUE and grid carbon intensity. This report is a methodology and evidence note, not an independent carbon audit."*

Suggested download label on the ESG page: "Download ESG Efficiency Methodology & Evidence Report (PDF)"

10. Path from Estimate to Audit-Ready Evidence

1. Capture workload telemetry

Log tokens, request count, route selected, model, duration, retries and fallback outcome for each workflow.

2. Capture compute telemetry

Collect GPU/CPU utilisation, runtime, energy where supported, or cloud-provider workload energy metrics.

3. Fix accounting boundary

Define whether energy is IT-only or facility-level; apply PUE only when required.

4. Version carbon factors

Store source, region, period and unit for each grid emission factor; avoid silent defaults.

5. Normalize by functional unit

Report gCO2e/token, gCO2e/document or gCO2e/workflow so architecture changes can be compared fairly.

6. Reconcile monthly

Compare estimated energy with actual infrastructure/cloud data and investigate material variance.

7. Independent assurance when needed

For formal ESG or investor assurance, provide logs, calculation workbook/code, factor sources and change history to an independent reviewer.

11. Alignment References

The following frameworks support the direction of transparent energy and climate reporting. Listing them does not by itself establish AINNA compliance or certification.

Malaysia Energy Efficiency and Conservation Act (EECA) 2024

Malaysia Energy Commission states that the Act regulates efficient energy consumption and conservation and came into force on 1 January 2025.

IFRS S1 and IFRS S2

ISSB standards provide general sustainability-related and climate-related disclosure requirements for decision-useful reporting.

Software Carbon Intensity (SCI)

The Green Software Foundation defines SCI as a methodology for software carbon intensity using energy, carbon intensity, embodied emissions and a functional unit.

SCI for AI

Extends SCI to AI systems and supports AI-specific functional units such as per token, per workflow or per page.

Power Usage Effectiveness (PUE)

The Green Grid defines PUE as total data-centre energy divided by ICT equipment energy.

IEA Energy and AI

IEA documents rapidly rising data-centre electricity demand and the growing role of AI, supporting the importance of compute efficiency.

12. Sources and References

1. AINNA ESG - Sustainable AI Infrastructure & Carbon-Aware NeuralOps. https://ainna.bond/esg/

2. AINNA Token Saving in SME Financial Statement Automation. https://ainna.bond/token-saving-study/

3. AINNA NeuralOps Carbon Footprint Emulator & Calculator. https://ainna.bond/carbon-emulator/

4. Green Software Foundation - Software Carbon Intensity (SCI) Specification. https://sci.greensoftware.foundation/

5. Green Software Foundation - SCI for AI. https://greensoftware.foundation/standards/sci-ai/

6. International Energy Agency - Energy and AI. https://www.iea.org/reports/energy-and-ai

7. Malaysia Energy Commission - Energy Efficiency and Conservation Act (EECA) 2024. https://www.st.gov.my/stakeholders/energy-efficiency/energy-efficiency-and-conservation-act-eeca-2024

8. IFRS Foundation - IFRS S1 General Requirements. https://www.ifrs.org/issued-standards/ifrs-sustainability-standards-navigator/ifrs-s1-general-requirements/

9. IFRS Foundation - IFRS S2 Climate-related Disclosures. https://www.ifrs.org/issued-standards/ifrs-sustainability-standards-navigator/ifrs-s2-climate-related-disclosures/

10. The Green Grid - Power Usage Effectiveness (PUE). https://www.thegreengrid.org/node/372

11. AINNA main site - Cyberpunk navigation / Industry section. https://ainna.bond/ainna/

All web sources accessed 10 August 2026. Public-site content may change after publication.

13. Methodology Disclaimer

This report is an internal methodology and evidence note prepared from AINNA public disclosures and external reference standards. It is not an independent assurance engagement, certification, lifecycle assessment or carbon audit. Arithmetic verification does not independently validate the underlying telemetry or assumptions. Industry coverage records published menu scope and architectural applicability; it does not imply a production deployment or verified ESG outcome in every listed vertical. Formal external claims should be supported by versioned logs, measured compute/energy data, documented emission factors and, where material, independent assurance.

Document owner

AINNA NeuralOps - Public ESG Methodology Note, Version 1.1. The recommended next revision should replace scenario energy intensities with observed infrastructure telemetry and publish a versioned emission-factor register.