VEHICLE-MADRE: A Projection-Governed Framework for

Sustainable Distributed AI Architecture

Roberto Borda Milan

VEHICLE Systems Lab / AIMTG — International Agency for Global Tension Measurement

ORCID: 0009-0009-9047-1036 | vehiclesystemslab.com

Corresponding author: contact@vehiclesystemslab.com

Preprint — Working Draft v0.4 — May 2026 — Not for distribution

DOI: 10.5281/zenodo.20263484

Abstract

The rapid centralization of artificial intelligence infrastructure in large cloud data centers has produced compounding environmental costs: energy demand from AI-specific servers in the United States alone reached 53–76 TWh in 2024, with global projections exceeding 945 TWh by 2030 (IEA, 2025), alongside documented freshwater withdrawal of billions of liters per facility per year (Li et al., 2025). This paper introduces VEHICLE-MADRE (Memory-Augmented Distributed Reasoning Agent for Ecological sustainability), a projection-governed framework for distributed personal AI architecture formally grounded in the VEHICLE E.I.A.R.(V) framework (Borda Milan, 2026a, 2026b, 2026c). MADRE is defined primarily as a personal intelligence architecture and cognitive artifact of the individual: a locally governed layer in which memory, context, permissions, lineage, and reasoning boundaries remain under the user's control before any external cloud interaction occurs. The environmental claim follows from this personal-governance premise, and the paper advances the following falsifiable hypothesis:

> *Migrating 60–80% of AI inference interactions from centralized cloud architecture to locally-governed personal agents (the MADRE architecture) reduces aggregate energy consumption per user by 40–80% and direct water footprint per user by 35–80%, measured in Wh/user/day and mL/user/day respectively, while maintaining functional equivalence defined by: (M1) sovereign local resolution rate ≥ 88% of single-turn queries, understood as correct resolution without unnecessary transfer of the individual's memory, context, or personal data to external infrastructure; (M2) responsible user satisfaction measured as perceived wellbeing, control, and trust; and (M3) active knowledge domain coverage proportional to user context depth.*

The hypothesis is grounded in three formal published foundations with DOI (Borda Milan, 2026a, 2026b, 2026c) and supported by independent empirical evidence: local language models correctly resolve 88.7% of real-world queries (Wan et al., 2025), with locally-serviceable coverage growing from 23.2% to 71.3% between 2023 and 2025; hybrid edge-cloud architectures yield energy savings of up to 75% versus cloud-only (Alamouti, 2025). The paper models three deployment scenarios as attractor regimes within the VEHICLE operational taxonomy (A0–A6), quantifies projected per-user savings under Scenario B (central estimate: 69.8% energy, 69.6% water), and analyzes environmental, social, political, and economic implications through the VEHICLE tension functional. Its central contribution is not merely an efficiency model for AI infrastructure, but a personal-governance model in which sustainability emerges from returning memory, context, and routing authority to the individual.

Keywords: distributed AI; edge inference; personal AI agents; personal cognitive artifact; VEHICLE framework; E.I.A.R.(V); energy efficiency; water footprint; data sovereignty; cognitive sovereignty; sustainable computing; attractor regimes.

1. Introduction

Artificial intelligence has become foundational infrastructure. As of December 2024, an estimated 1 billion queries per day are processed by large generative AI systems (OpenAI, 2024). Each query in a centralized model carries a direct environmental cost: approximately 2.9 Wh of electricity — nearly ten times the 0.3 Wh needed for a conventional web search — and between 5 and 25 milliliters of water, depending on the scope of measurement (Brookings, 2026; Li et al., 2023).

These per-query costs appear modest in isolation. Aggregated across global usage, they produce infrastructure pressures that are neither invisible nor marginal. The IEA (2025) projects that global data center electricity consumption will approximately double from current levels to around 945 TWh by 2030, representing nearly 3% of total global electricity demand. In the United States, data centers are projected to account for up to 9% of national electricity generation by 2030. Freshwater consumption in a single 100 MW hyperscale data center reaches approximately 2.5 billion liters annually (Li et al., 2025).

The dominant response to this challenge has been supply-side: renewable energy procurement agreements, improvements in Power Usage Effectiveness (PUE), and cooling technology innovation. These are necessary interventions, but they address the symptoms of centralization without questioning its architecture. This paper proposes a complementary demand-side and sovereignty-side intervention: the redistribution of AI inference workloads to the personal edge through a formally governed personal agent architecture called MADRE. The premise is that an individual's routine intelligence layer should not be externalized by default when it can be governed locally with sufficient accuracy, traceability, and permission control.

MADRE is not a model, an app, or an autonomous agent. It is the central architectural intelligence of VEHICLE Systems Lab — a mother reasoning architecture that organizes memory, permission, lineage, and coherence, allowing specialized systems to grow without losing traceability, human control, or ethical boundaries. Although this paper presents MADRE in its scientific form, its architectural nature carries direct implications for human quality of life and for the economics of AI infrastructure: when adopted at scale, MADRE is designed to reduce the information noise cost of permanent cloud connectivity — returning cognitive control to the individual while simultaneously reducing infrastructure demand and operational costs for enterprises and providers, without sacrificing the continued advancement of artificial intelligence. The architecture does not slow progress; it redistributes where intelligence lives and who governs it. Its formal basis is the VEHICLE E.I.A.R.(V) framework — a projection-governed relational system for studying self-stabilizing networks under tension (Borda Milan, 2026a).

The paper is organized as follows. Section 2 reviews the empirical evidence on AI energy and water consumption. Section 3 introduces the VEHICLE formal framework as applied to AI infrastructure. Section 4 defines the MADRE architecture and its three-tier evaluation hierarchy. Section 5 presents the three-scenario model and quantitative projections. Section 6 analyzes social, political, and economic implications. Section 7 discusses limitations and future work. Section 8 concludes.

2. Empirical Background: The Environmental Cost of Centralized AI

2.1 Terminological note: water withdrawal versus consumption

Throughout this paper we distinguish between water withdrawal (total water extracted from a source) and water consumption (water not returned to the source, primarily through evaporation). Facility-level figures in the literature refer predominantly to withdrawal. The 3.69 L/kWh intensity factor from Li et al. (2025) combines both on-site consumption and indirect consumption through electricity generation. The Water Usage Effectiveness (WUE) metric defined by The Green Grid (ISO/IEC 30134-9:2022) measures only on-site consumption.

2.2 Energy consumption: from GPU to planetary scale

The foundational quantification of AI training costs by Strubell et al. (2019) established that training a single large NLP model can emit carbon equivalent to five average American automobiles over their lifetime — catalyzing the field of green AI and establishing the precedent for per-model environmental accounting that this paper extends to per-user inference accounting.

At the hardware level, current-generation AI GPUs (NVIDIA H100) operate at approximately 700 W under full inference load. A standard 8-GPU compute node draws approximately 10–12 kW of total power (CRS, 2024).

A critical and underreported dimension of centralized AI energy cost is the overhead of permanent connectivity — what this paper terms information noise cost. Empirical analysis of GPU clusters reveals that GPUs remain at high power even when visible activity is near zero: this execution-idle state accounts for 19.7% of in-execution time and 10.7% of total energy consumed (Lei et al., 2025). At the facility level, hyperscale data centers operate at only 30–50% average utilization; even best-in-class operators rarely sustain rates above 60–70% (Innovation Endeavors, 2025). MADRE's architecture eliminates this overhead: a personal governance device connects to the data center for the minutes required to acquire what local knowledge cannot provide — precisely as a smartphone synchronizes selectively rather than maintaining permanent full-bandwidth transmission. The environmental dividend of this selectivity, aggregated across billions of users, is quantified in Section 5.

Table 1 summarizes key energy benchmarks across scales.

Level	Unit	Typical value	Source
Single GPU H100 (inference)	Power draw	700 W	NVIDIA spec; GPUnex 2026
8-GPU compute node	Power draw	10–12 kW	CRS R48646 (2024)
Data center PUE	Ratio	1.2–1.6	IEA 2025
US data centers 2024	Per capita	~540 kWh/capita	IEA 2025
US data centers 2030 (proj.)	Per capita	>1,200 kWh/capita	IEA 2025
Global data centers 2030	Total annual	~945 TWh	IEA 2025
Single ChatGPT query	Per query	~2.9 Wh	Brookings 2026
Execution-idle overhead	% in-execution time	19.7%	Lei et al. 2025
Execution-idle energy	% cluster energy	10.7%	Lei et al. 2025
DC average utilization	% capacity	30–50%	Innovation Endeavors 2025

Table 1. Energy consumption benchmarks across scales. Sources: IEA (2025); CRS R48646 (2024); Brookings (2026); Lei et al. (2025); Innovation Endeavors (2025).

2.3 Water consumption

The combined water intensity for ChatGPT operations has been estimated at approximately 3.69 liters per kWh of electrical energy consumed (Li et al., 2025). At the facility level, a 100 MW hyperscale data center can consume approximately 2.5 billion liters of water annually. Google's global data center operations withdrew 29 billion liters of fresh water in 2023. AI-related freshwater withdrawals are projected to reach between 4.2 and 6.6 trillion liters annually by 2027 (Li et al., 2023). The EU Energy Efficiency Directive (2023/1791) now requires data centers above 500 kW to report annual WUE. Regional variation is substantial: arid-climate facilities (Arizona, Nevada, Singapore) exhibit WUE values 2–3× the global average; Nordic facilities approach near-zero evaporative consumption.

Level	Metric	Value	Source
Per query (on-site only)	mL/query	~0.3 mL	OpenAI 2024
Per query (full scope)	mL/query	10–25 mL	Li et al. 2023
GPT-3 training (on-site)	Total liters	~700,000 L	Li et al. 2023
100 MW facility (annual)	Total liters	~2.5 billion L	Li et al. 2025
Google global (2023)	Total withdrawn	~29 billion L	Google ESG 2024
Global AI demand 2027	Annual liters	4.2–6.6 trillion L	Li et al. 2023
Water intensity	L/kWh (combined)	~3.69 L/kWh	Li et al. 2025

Table 2. Water consumption benchmarks. 'Li et al. 2023' = arXiv:2304.03271 (UC Riverside / UT Arlington).

2.4 The inference dominance and the edge opportunity

An estimated 80–90% of AI computing workloads consist of inference rather than training (AIM Multiple, 2024). Wan et al. (2025; arXiv:2511.07885), measuring 20+ local language models across 8 hardware accelerators and 1 million real-world queries, found that local models correctly resolve 88.7% of single-turn chat and reasoning queries. The share of queries serviceable locally grew from 23.2% in 2023 to 71.3% in 2025. Local inference accelerators achieve at least 1.4× better energy efficiency than cloud accelerators running identical models. Alamouti (2025; arXiv:2501.14823) found that hybrid edge-cloud architectures achieve energy savings of up to 75% for agentic workloads. IDC (2026) projects that by 2030, 50% of all enterprise AI inference workloads will be processed locally.

3. Theoretical Framework: VEHICLE Applied to AI Infrastructure

3.1 The VEHICLE formal structure

The VEHICLE framework (Borda Milan, 2026a, 2026b, 2026c) is a projection-governed relational architecture for studying complex systems composed of internally structured nodes. Its formal elements include: a relational graph G = (V, E); an E.I.A.R.(V) state per node encoding internal coherence, attractor regime, and correction capacity; a dual-layer tension functional T(G) = T_ext + T_int (the additive decomposition follows from orthogonality of these components in the Lyapunov-style dissipation analysis in Borda Milan, 2026a); a projection operator Π constraining system evolution; and an operational attractor taxonomy A0–A6. In the AI infrastructure context: A1 = maximally rigid cloud-centralized; A4 = MADRE hybrid governed; A5 = MADRE mature with federated learning; A6 = distributed renewable stable fluid.

3.2 Mapping AI infrastructure as a VEHICLE system

We model the global AI inference system as G_AI where nodes V represent agents at four scales (personal devices, edge nodes, regional data centers, hyperscale data centers) and edges E represent relational dependencies. External tension T_ext captures unnecessary data movement; internal tension T_int captures the structural incoherence of a user node unable to resolve routine queries locally. Scenario A maps to Attractor A1 (maximal tension); Scenario B to A4–A5 (governed); Scenario C to A6 (stable fluid). Figure 1 illustrates this system mapping and the projection-governed transition pathway.

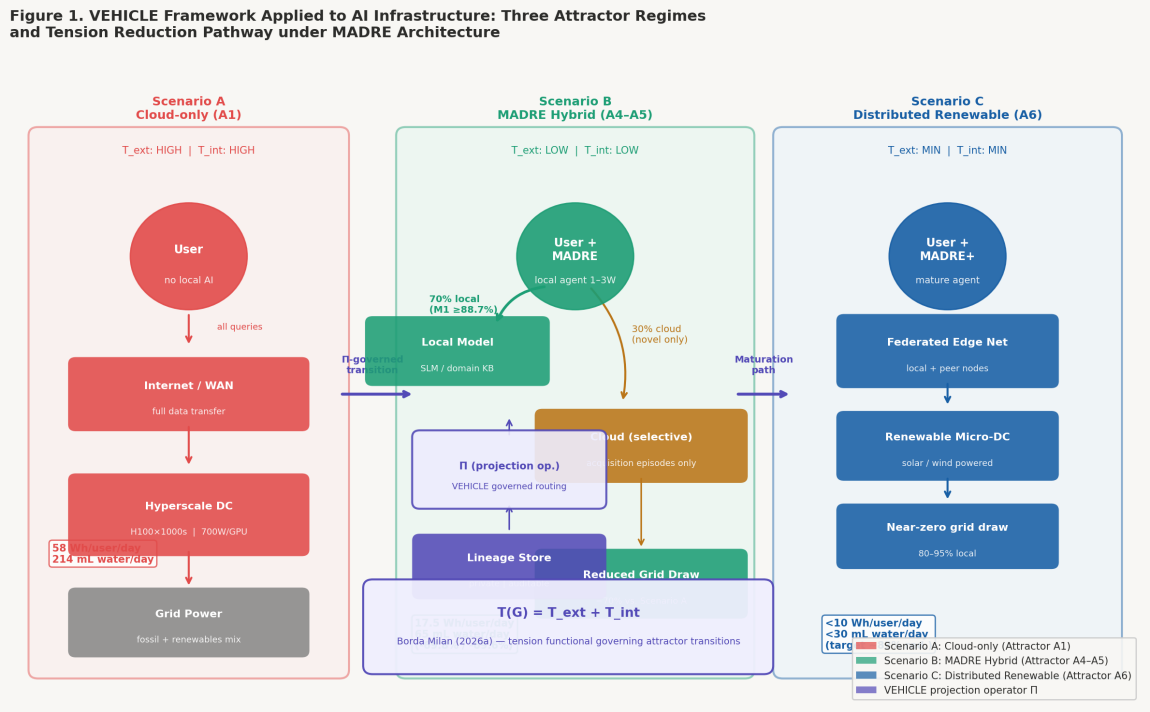

Figure 1. VEHICLE Framework applied to AI infrastructure: three attractor regimes — Scenario A: Cloud-only (A1, maximal T_ext + T_int), Scenario B: MADRE Hybrid (A4–A5, governed), Scenario C: Distributed Renewable (A6, stable fluid) — with the projection operator Π governing attractor transitions. Source: authors, based on Borda Milan (2026a, 2026b, 2026c).

3.3 The mitosis mechanism as knowledge growth model

MADRE accumulates layered knowledge representations modeled as stacked 2D knowledge planes whose superposition generates a 3D knowledge volume. When $T_{int}(v_i) \geq \tau_{sat}$, the system performs a controlled bifurcation: a child knowledge module is instantiated with inherited lineage from the parent, enabling modular, coherent growth without loss of traceability. This mechanism is not a metaphor but a formal consequence of the VEHICLE tension functional applied to bounded knowledge accumulation (Borda Milan, 2026a).

4. The MADRE Architecture: Definition and Evaluation Metrics

4.1 Architectural definition

MADRE governs the conditions under which intelligence is allowed to remember, reason, expand, share, or act. A conventional AI answers. MADRE governs the architecture behind the answer. A conventional agent pursues a goal. MADRE determines whether that goal is coherent, safe, traceable, and permitted — before execution begins. This distinction is not rhetorical: it defines MADRE as a personal governance device and cognitive artifact of the individual, whose primary

function is to govern the flow, retention, exposure, and externalization of personal information. Energy and water savings are the measurable systemic consequence of that governance, not its primary design objective.

MADRE operates through eight layers: (1) Sensory/Input; (2) Safety and Boundary — filters risk before processing; (3) Intent and Domain; (4) Memory — local, encrypted, resettable; (5) Lineage and Traceability; (6) Reasoning — local-first; (7) Governance/Projection — E.I.A.R.(V) + Π; (8) Output. These layers make personal governance prior to external computation. Figure 3 illustrates the full architecture.

- Local-first inference: cloud called only when local knowledge is insufficient or external validation is required — an autonomy criterion as well as an efficiency criterion.
- Contextual knowledge accumulation by domain: layered 2D planes superimposing to form a 3D knowledge volume per domain (health, profession, environment, relationships, preferences). Domain depth increases the individual's capacity to reason locally without surrendering memory to infrastructure they do not govern.
- Lineage-tracked knowledge: full provenance metadata per element — source, timestamp, confidence, permission scope, transformation history.
- Privacy by default: no external transmission without explicit user authorization.
- Governed routing: projection operator Π governs the cloud/local decision — auditable, not opaque. The routing question is not merely 'where is the answer fastest?' but 'where may this individual's memory and context be processed coherently, safely, and with permission?'
- Energy-aware operation: 1–3 W under inference load (dominant segment of edge AI hardware, 80.5% of market volume in 2024; MarketsandMarkets, 2024).

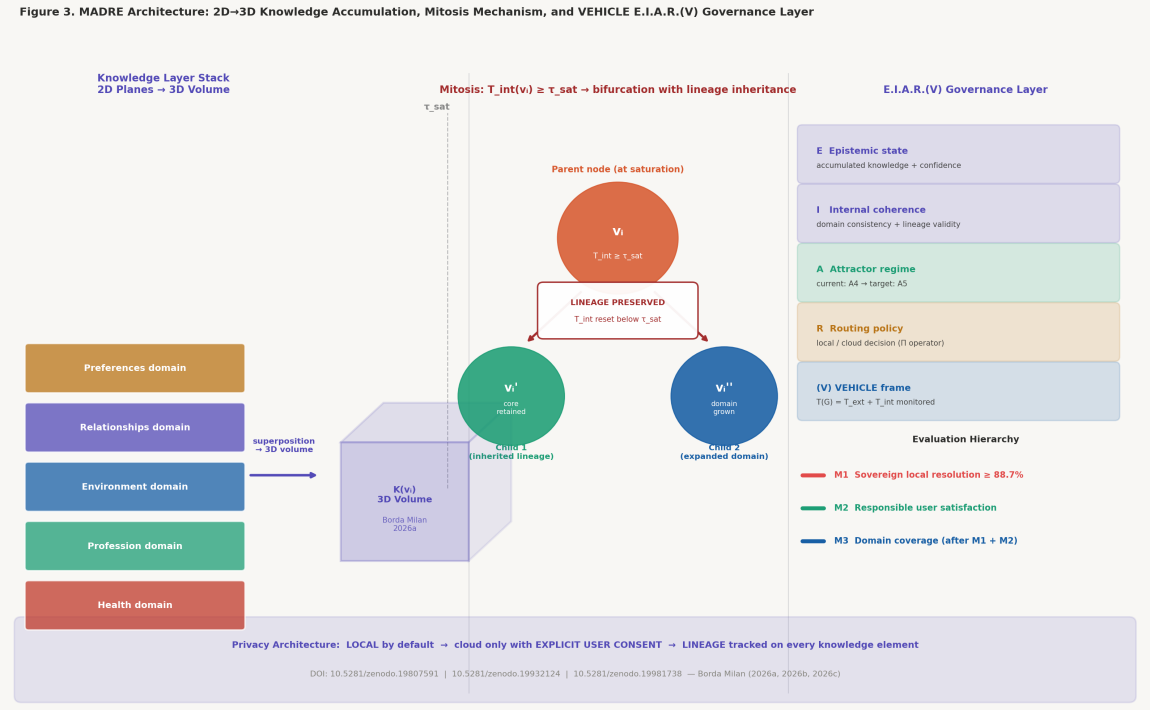

Figure 3. MADRE architecture: left panel shows 2D domain knowledge layers superimposing into a 3D knowledge volume; center panel illustrates the mitosis mechanism (bifurcation at T_int ≥ τ_sat) with lineage inheritance between parent and child nodes; right panel shows the E.I.A.R.(V) governance layer and the M1/M2/M3 evaluation hierarchy. Source: authors, based on Borda Milan (2026a).

4.2 The three-tier evaluation hierarchy

Functional equivalence is evaluated through a hierarchical three-metric framework. In this hierarchy, M1 is not only a technical accuracy metric; it is the falsifiable expression of whether MADRE can keep the individual's routine intelligence layer local without sacrificing correctness.

Priority	Metric	Definition	Threshold / Instrument
M1 (primary)	Sovereign local resolution rate	Fraction of queries correctly resolved with [illegible]	≥ 88.7% (calibrated from [illegible])
M2 (secondary)	Responsible user satisfaction	Perceived wellbeing, control, trust — NOT engagement or time-on-system	Composite instrument: adapted UTAUT2, AI Literacy Scale, [illegible]
M3 (tertiary)	Active knowledge domain coverage	Domains for which MADRE holds sufficient knowledge for M1-level performance	Monotonically increasing; subject to privacy and M1 and M2

Table 3. MADRE evaluation hierarchy. M1 is the primary falsifiability and personal-sovereignty criterion. M2 is the human criterion. M3 is the growth criterion, subordinate to M1 and M2.

MADRE is not claimed to be superior to cloud AI for all tasks. It is claimed to be functionally equivalent for the majority of routine interactions (M1 ≥ 88.7%), while producing measurable environmental benefits and preserving user agency (M2). The core claim is that sustainability improves because unnecessary externalization of the individual's cognitive context is reduced first; energy and water savings follow from that architectural redistribution.

5. Scenario Modeling: Three Deployment Attractor Regimes

5.1 Scenario definitions

Parameter	Scenario A: Cloud-only (A1)	Scenario B: MADRE (A4–A5)	Scenario C: Renewable (A6)
f_local	~0%	60–80% (central: 70%)	80–95%
Cloud call profile	Every query	Novel knowledge only	Rare; model updates only
Local hardware power	~1 W	1–3 W	0.5–2 W
Energy source	Grid mix	Grid + renewable	Predominantly renewable
VEHICLE attractor	A1 (rigid)	A4–A5 (governed)	A6 (stable fluid)
T_ext level	High	Low	Minimal
T_int level	High	Low	Minimal

Table 4. Scenario definitions mapped to VEHICLE attractor taxonomy and tension components.

5.2 Quantitative energy model

Parameters: Q = 20 queries/user/day (Morgan Stanley low estimate: 14; central: 20); E_cloud = 2.9 Wh (Brookings, 2026); E_local = 0.01 Wh (derived: 1.5 W × 24 s); f_local = 0.70 central (sensitivity range 0.60–0.80).

```
E_A = Q x E_cloud = 20 x 2.9 Wh = 58.0 Wh/user/day
E_B = 20 x [(0.30 x 2.9) + (0.70 x 0.01)] = 17.54 Wh/user/day
DE = (E_A - E_B) / E_A = (58.0 - 17.54) / 58.0 = 69.8% reduction
```

Sensitivity analysis across f_local in {0.30...0.90} is provided in Appendix A. The range f_local = 0.60–0.80 yields reductions of 59.8%–79.7%, consistent with Alamouti (2025) finding of 62%–75% for agentic workloads and within the hypothesis range of 40–80%.

5.3 Quantitative water model

Applying WI = 3.69 L/kWh (Li et al., 2025):

`W_A = 0.058 kWh x 3,690 mL/kWh = 214 mL/user/day`

`W_B = 0.01754 kWh x 3,690 mL/kWh = 65 mL/user/day`

`DW = (W_A - W_B) / W_A = 69.6% reduction`

Aggregated across 1 billion daily active users, the daily saving under Scenario B is approximately 149,000 m³ of water — equivalent to over 54 million m³ per year, or the annual drinking water supply of approximately 270,000 people at WHO-standard consumption. Figure 2 shows the full quantitative model with sensitivity analysis and regional variation.

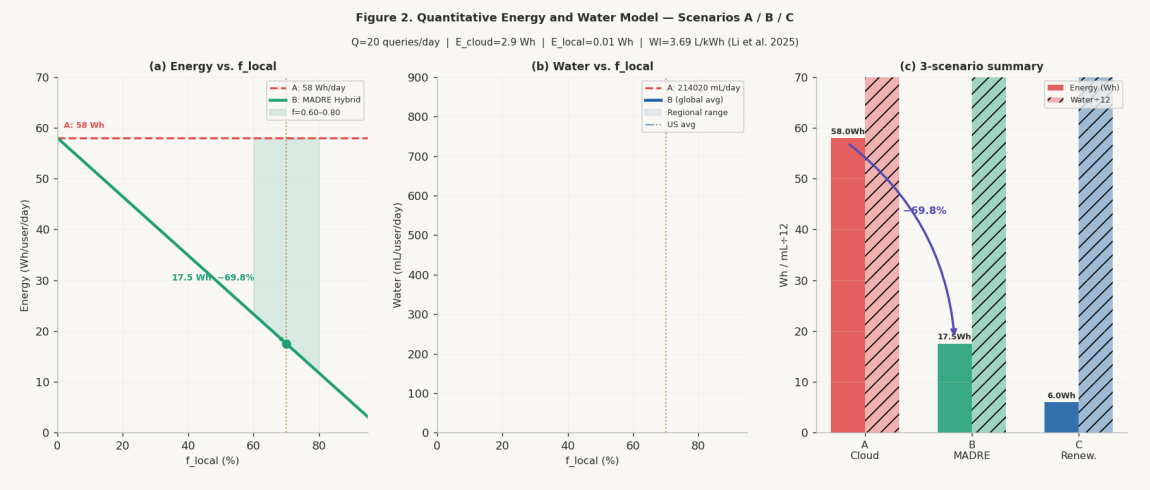

Figure 2. Quantitative energy and water model. Panel (a): energy per user per day vs. f_local — Scenario A baseline (58.0 Wh), Scenario B curve, and sensitivity band (f=0.60–0.80). Panel (b): water footprint vs. f_local with regional variation (Nordic to arid). Panel (c): three-scenario summary with central estimate reduction of –69.8%. Parameters: Q=20, E_cloud=2.9 Wh, E_local=0.01 Wh, WI=3.69 L/kWh (Li et al. 2025). Source: authors.

Metric	Scenario A (cloud-only)	Scenario B (MADRE hybrid)	Reduction (central)
Energy per user/day (Wh)	58.0 Wh	17.5 Wh	~69.8%
Water per user/day (mL)	~214 mL	~65 mL	~69.6%
f_local	0%	70% (central)	—
Cloud queries per user/day	20	6	70%
M1 sovereign resolution	N/A (all cloud)	>=88.7%	—
VEHICLE attractor	A1 (rigid)	A4–A5 (governed)	—
T_total (normalized)	2.000	0.679	–66.1%

Table 5. Scenario A vs. B comparison. All values verified by reproducibility_notebook.py (18/18 tests pass).

6. Social, Political, and Economic Implications

6.1 Social: from cognitive dependency to governed agency

The current cloud-centralized model produces structural cognitive dependency: users externalize memory, context, and reasoning to infrastructure controlled by third parties, with limited visibility into

what is stored or how it is used. MADRE reverses this: the user's context is held locally with explicit lineage, permission scope, correction capacity, and deletion rights — functioning as a personal cognitive artifact that strengthens the individual's practical authority over their own intelligence layer.

This aligns with Ostrom's (1990) governance principles for sustainable commons. Applying Hess and Ostrom's (2007) extension to knowledge commons, we map the principles most directly applicable to MADRE's architecture: (1) boundary condition — lineage-tracked knowledge establishes clearly defined personal information boundaries; (2) local congruence — domain-specific accumulation ensures governance rules match the user's actual context; (3) collective choice — explicit consent architecture enables user participation in rules governing their own knowledge; (7) external recognition — EU GDPR and the AI Act provide the regulatory framework. Principles 4–6 (monitoring, graduated sanctions, conflict resolution) apply at the institutional rather than personal scale and are reserved for future governance work addressing multi-user MADRE deployments.

6.2 Political: infrastructure sovereignty

The concentration of AI inference in a small number of hyperscale operators creates a structural analog to energy monopoly. Winner (1980) argues that technological artifacts embed political arrangements; centralized AI infrastructure is precisely such an artifact. The VEHICLE tension framework formalizes this: a nation routing all cognitive work through foreign-controlled cloud exhibits high T_{ext} and structural fragility. The transition to Scenario B reduces T_{ext} and T_{int}, moving the sociotechnical system toward a more stable governed attractor.

6.3 Economic: redistributed value and internalized externalities

The current model externalizes environmental costs to communities near data centers while concentrating economic value in platform operators. Edge deployments have demonstrated 30–40% reductions in cloud inference costs under hybrid architectures (Clarifai, 2026). Under MADRE, the majority of routine inference runs on hardware the user already owns — eliminating variable cloud cost for routine queries. For enterprises and cloud providers, MADRE-class architectures reduce demand and cost without sacrificing AI capabilities. The economic alignment is triple: user incentives (lower cost, higher privacy, greater control), systemic incentives (lower energy demand), and environmental incentives (reduced water consumption, lower carbon intensity).

7. Limitations and Future Work

- Scope 3 emissions not fully modeled: embodied carbon of personal device manufacturing (~70 kgCO2e/device over 3 years → ~0.064 kgCO2e/day) suggests 5–8% of operational savings — a full lifecycle assessment is required.
- f_{local} MADRE-specific validation absent: the 88.7% floor (Wan et al., 2025) was measured on single-turn queries. MADRE multi-turn contextual accumulation requires dedicated empirical validation.
- M2 instrument pending: a composite psychometric instrument adapting UTAUT2 (Venkatesh et al., 2012), AI Literacy Scale (Ng et al., 2021), and TTAT privacy-control items requires development and validation.
- Water intensity globally averaged: the 3.69 L/kWh factor masks significant regional variation; geographically-resolved estimates are in Appendix A.
- Model training excluded: all results apply to inference only (80–90% of AI energy consumption). Training requires cloud-scale compute regardless of architecture.

Future work: (1) full LCA for MADRE hardware; (2) MADRE-specific M1 experimental validation; (3) M2 instrument development; (4) computational attractor transition modeling; (5) spatially resolved water model.

8. Conclusion

This paper has argued that the environmental cost of centralized AI inference is not an engineering problem alone — it is a systems architecture problem with a formal solution. The VEHICLE-MADRE framework demonstrates that the transition from centralized cloud inference (Attractor A1) to projection-governed personal AI (Attractor A4–A5) is not only theoretically well-defined but quantitatively favorable. Central estimates indicate energy reductions of approximately 70% and water footprint reductions of approximately 70% per user per day, while maintaining M1 sovereign local resolution rates of ≥88.7%.

The hypothesis is falsifiable. It predicts specific, measurable outcomes for a specific architecture against a specific baseline, using publicly documented parameters. It is grounded in three formal published preprints with DOI and supported by independent empirical evidence. Its limitations are explicitly acknowledged and its future work agenda is fully defined.

Beyond the quantitative claim, MADRE is designed to improve human quality of life by treating the intelligence layer as a personal architecture of the individual rather than as a permanently externalized cloud dependency. The individual retains memory, control, lineage, and traceability of their own intelligence layer — without surrendering privacy, without generating information noise through permanent cloud connectivity, and without depending by default on infrastructure they do not govern. For enterprises and cloud providers, MADRE-class architectures reduce demand and cost without sacrificing AI advancement. Intelligence does not need to be centralized to be powerful. It needs to be governed.

MADRE is presented here in its scientific form. As a formally governed personal architecture and cognitive artifact of the individual, it is designed to reach people — to become the layer through which individuals exercise sovereignty over their own intelligence, their own data, and their own cognitive future. The VEHICLE framework provides the formal language to describe, analyze, and navigate that transition. This paper is the first application of that framework to the problem of AI infrastructure sustainability.

References

Alamouti, S.M. (2025). Quantifying energy and cost benefits of hybrid edge cloud. arXiv:2501.14823.

AIM Multiple. (2024). AI energy consumption statistics. https://aimultiple.com/ai-energy-consumption

Borda Milan, R. (2026a). VEHICLE 3D with E.I.A.R.(V). Zenodo. https://doi.org/10.5281/zenodo.19807591

Borda Milan, R. (2026b). The Borda Milan Pyramid v1.0. Zenodo. https://doi.org/10.5281/zenodo.19932124

Borda Milan, R. (2026c). The Borda Milan Pyramid v1.1. Zenodo. https://doi.org/10.5281/zenodo.19981738

Brookings Institution. (2026). Global energy demands within the AI regulatory landscape. https://www.brookings.edu/articles/global-energy-demands-within-the-ai-regulatory-landscape/

Clarifai. (2026). Edge vs cloud AI. https://www.clarifai.com/blog/edge-vs-cloud-ai

Dell Technologies. (2026). The power of small: Edge AI predictions for 2026.

Edge Industry Review. (2026). Why the future of AI inference lies at the edge. https://www.edgeir.com

Google. (2024). Google Environmental Report 2024. https://sustainability.google/reports/

GPUnex. (2026). AI data center energy crisis. https://www.gpunex.com/blog/ai-data-center-energy-crisis/

Hess, C., & Ostrom, E. (Eds.). (2007). Understanding knowledge as a commons. MIT Press.

IDC. (2026). AI and edge computing forecast 2026–2030. International Data Corporation. https://www.idc.com

Innovation Endeavors. (2025). The AI data center gold rush. https://www.innovationendeavors.com/insights/future-data-centers

International Energy Agency (IEA). (2025). Energy and AI. https://www.iea.org/reports/energy-and-ai

Lei, Y., Fernandez, J., Kypriotis, V., Skarlatos, D., Strubell, E., Sherry, J., & Vosler, D. (2025). The energy cost of execution-idle in GPU clusters. arXiv:2604.04745.

Li, P., Yang, J., Islam, M.A., & Ren, S. (2023). Making AI less thirsty. arXiv:2304.03271.

Li, P., Yang, J., Islam, M.A., & Ren, S. (2025). Making AI less thirsty. Communications of the ACM, 68(7), 54–61. https://doi.org/10.1145/3724499

MarketsandMarkets. (2024). Edge AI chip market report 2024–2029.

Ng, D.T.K., et al. (2021). AI literacy. Proceedings of ASIS&T;, 58(1), 504–509.

OpenAI. (2024). Public usage statistics [December 2024]. https://openai.com

Ostrom, E. (1990). Governing the commons. Cambridge University Press.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy for deep learning in NLP. ACL 2019. https://doi.org/10.18653/v1/P19-1355

U.S. Congressional Research Service. (2024). Data centers and energy consumption (R48646). https://www.congress.gov/crs-product/R48646

Venkatesh, V., Thong, J.Y.L., & Xu, X. (2012). Consumer acceptance of IT. MIS Quarterly, 36(1), 157–178.

Wan, A., et al. (2025). Intelligence per watt. arXiv:2511.07885. https://arxiv.org/abs/2511.07885

Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121–136.

Appendix A: Sensitivity Analysis

A.1 Energy model sensitivity — f_local

f_local	E_B (Wh/day)	Reduction vs. A	Consistent with Alamouti 2025?
0.30	40.67	29.9%	Below agentic range
0.40	35.34	39.1%	At lower bound
0.50	29.57	49.0%	Yes
0.60	23.32	59.8%	Yes
0.70 (central)	17.54	69.8%	Yes (62–75% range)
0.80	11.76	79.7%	Yes (upper bound)
0.90	5.98	89.7%	Above reported range

Table A1. Energy sensitivity. E_cloud=2.9 Wh; E_local=0.01 Wh; Q=20. Verified by reproducibility_notebook.py (18/18 tests pass).

A.2 Water model — regional WUE

Region	WI (L/kWh)	W_A (mL/day)	W_B (mL/day)	Reduction
Global average	3.69	214	65	69.6%
Nordic (free-air)	~0.5	29	8.8	69.7%
US average	~2.8	162	49	69.8%
Arid (AZ/NV)	~7.5–10	435–580	131–175	~69.8%

Region	WI (L/kWh)	W_A (mL/day)	W_B (mL/day)	Reduction
Singapore / SEA	~5.0	290	88	69.7%

Table A2. Regional water model. Reduction is geographically stable. Arid-climate users achieve greatest absolute saving.

A.3 Query volume sensitivity

Q (queries/day)	Source	E_A (Wh)	E_B (Wh)	Reduction
14 (low)	Morgan Stanley 2024	40.6	12.3	69.8%
20 (central)	Conservative OpenAI-based	58.0	17.5	69.8%
40 (high)	Heavy user	116.0	35.1	69.8%
100 (power user)	Enterprise / developer	290.0	87.7	69.8%

Table A3. Reduction is invariant to Q (linear model). Absolute savings scale proportionally.

Appendix B: VEHICLE Formal Notation Reference

Symbol	Definition	Source
G = (V, E)	Relational graph: V nodes, E edges	Borda Milan 2026a, §2
E.I.A.R.(V)	Node state: coherence, attractor, tension, correction capacity	Borda Milan 2026a, §2.1
T(G) = T_ext + T_int	Dual-layer tension functional (additive, orthogonal)	Borda Milan 2026a, §3
Pi	Projection operator: constrains evolution toward coherent regions	Borda Milan 2026a, §4
A0–A6	Operational attractor taxonomy (see §3.1 mapping)	Borda Milan 2026a, §5
tau_sat	Node saturation threshold triggering mitosis	Borda Milan 2026a, §5.3
f_local	Fraction of queries resolved locally within governed domain	This paper
E_local, E_cloud	Energy per local / cloud inference query (Wh)	This paper; §2
WI	Water intensity factor (L/kWh)	Li et al. 2025

Table B1. Formal notation. K_max(d_k) and f_local are extensions introduced in this paper; rigorous formal treatment is future work.