Scientific Research Publishing

Advances in Artificial Intelligence and Robotics Research, 2026, 2(2), 58-77
https://www.scirp.org/journal/airr

Intelligentia Artificialis Privata

—A Framework for Sovereign, Localized Artificial Intelligence Systems

George Davey

Independent Researcher, West Des Moines, IA, USA
Email: George.Davey@QuantumLinear.com

How to cite this paper: Davey, G. (2026) Intelligentia Artificialis Privata. *Advances in Artificial Intelligence and Robotics Research*, **2**, 58-77.

Received: April 5, 2026
Accepted: May 24, 2026
Published: May 27, 2026

Abstract

The centralization of artificial intelligence infrastructure presents a systemic challenge to human epistemic autonomy, data sovereignty, and long-term computational resilience. This paper introduces *Intelligentia Artificialis Privata* (Privata AI) as a formally defined class of AI systems distinguished by physical locality, operational autonomy, and bounded knowledge domains. We present a rigorous architectural model, a formal threat analysis, a comparative economic analysis, and a set of computational constraints governing such systems. We argue that Privata AI is not merely a privacy-enhancing variant of existing paradigms but constitutes a categorical shift in how intelligence is instantiated, owned, and constrained. Drawing on developments in federated learning, edge computing, differential privacy, and hardware trust architectures, we situate Privata AI within the broader landscape of privacy-preserving machine learning and identify its distinguishing properties. We further formalise two advances that strengthen the practical case for sovereign local inference: the *Hardware-Bound Sealing* (HBS) protocol, which binds model execution to a verified silicon identity via Physical Unclonable Functions and a formally defined human-in-the-loop interrupt $\mathcal{I}_{\text{human}}$; and the 1.58-*bit Parity Horizon*, an empirical scaling law establishing that ternary quantisation achieves functional parity with full-precision models above a critical parameter threshold of approximately 13.4 billion parameters, consistent with the BitNet b1.58 scaling results. The implications span systems architecture, NPU micro-architecture, security engineering, epistemology of computation, political economy, and regulatory theory.

Keywords

Privacy-by-Design, Local Inference, Data Sovereignty, NPU Micro-Architecture, Hardware-Bound Sealing, Ternary Quantisation, Edge AI, Epistemic Autonomy, Threat Modeling, AI Architecture

1. Introduction

1.1. Motivation and Background

This paper presents a conceptual and architectural framework for Privata AI systems, grounded in existing work in edge computing, privacy engineering, and secure computation.

The prevailing architecture of AI deployment concentrates inference, training, and knowledge management in large centralized cloud platforms. While this model has enabled rapid capability scaling—transitioning from research prototypes to systems serving hundreds of millions of users within a few years [1]—it introduces structural dependencies that can undermine user autonomy at multiple levels.

Under common cloud architectures, users who interact with centralized AI systems expose their queries, behavioral patterns, and reasoning processes to third-party infrastructure operators. This constitutes a form of *epistemic leakage* that is, in many current deployments, not merely incidental but architecturally induced. The problem is structurally analogous to the data sovereignty challenges identified in the early cloud computing literature [2], but is more acute: unlike passive data storage, interactive AI inference reveals the user's active reasoning processes, intellectual interests, and decision-making patterns in real time.

The continuous and opaque update cycles of cloud-deployed models create a *Forensic Vacuum*: the system's epistemic content can shift without user awareness or consent [3]. In legal research or medical decision support, a query yielding Result A today may yield Result B tomorrow due to an unannounced weight update—a violation of the reproducibility required for professional accountability. This *epistemic volatility* is incompatible with systems that require deterministic, auditable behavior.

The concentration of AI infrastructure has attracted growing regulatory attention. The European Union's AI Act [4], the U.S. Executive Order on AI of 2023 [5] (subsequently rescinded in January 2025), and emerging frameworks in multiple jurisdictions reflect recognition that the current deployment model creates risks that market mechanisms alone may be insufficient to address. Existing instruments, however, operate at the policy level rather than the architectural level. Privata AI offers an approach that satisfies regulatory requirements by construction rather than by compliance.

1.2. Core Thesis

Intelligence that is not locally controlled is not truly possessed.

We draw a sharp distinction between *privacy-preserving cloud AI*—which employs cryptographic or statistical methods to reduce data exposure while maintaining centralized infrastructure—and *Privata AI*, which achieves privacy as an emergent consequence of physical locality and operational closure. The former offers probabilistic guarantees conditioned on trust in the provider; the latter offers deterministic guarantees by architectural necessity. **Table 1** illustrates how this distinction manifests under concrete failure conditions.

Table 1. Failure mode comparison: Cloud AI versus Privata AI.

Failure Case	Cloud AI	Privata AI
Provider Bankruptcy	Service Terminated	System Unaffected
Policy Shift	Model Behaviour Changes	User-Controlled Weights
Subpoena/Data Request	Data Exfiltrated	No Data to Exfiltrate
Internet Outage	Total Downtime	Operational Autonomy
Unannounced Weight Update	Silent Epistemic Drift	Requires User Authorisation

This distinction parallels Westin's foundational taxonomy of privacy [6], in which the strongest form of informational privacy requires not merely the absence of disclosure but the structural incapacity for disclosure. We model Privata AI systems as enforcing a constraint whereby locally generated or ingested data remains within a bounded computational domain unless explicitly authorised by the user—a property achieved through physical locality, operational closure, and user-controlled update governance rather than through policy or contractual means. A Privata system must satisfy three necessary conditions: it must be **physically local** (inference occurs on user-controlled hardware), **computationally autonomous** (no network dependency during operation), and **knowledge-bounded** (its epistemic state is determined solely by local inputs and user-controlled updates).

1.3. Scope and Contributions

This paper makes the following contributions:

- A formal definition of Privata AI with necessary and sufficient conditions, distinguishing it from related concepts in privacy-preserving machine learning.
- A layered architectural framework specifying the minimum viable system stack for a compliant Privata deployment, including a Two-Stage Update Protocol that addresses consumer UX without compromising sovereignty.
- A structured threat model with adversary classes and corresponding mitigations.
- An analysis of computational constraints and optimisation techniques that make Privata AI practically deployable on consumer hardware.
- An epistemological analysis of the closed knowledge manifold and its implications for system predictability and intellectual ownership.
- An examination of economic, political, and regulatory implications at scale, including the emerging Sovereign AI Economy.
- A structured comparison with related work, sharpened to distinguish Privata AI's architectural goals from those of Edge AI and federated learning.
- A formalisation of the *Hardware-Bound Sealing* (HBS) protocol with a comparative security analysis against TPM 2.0 and Intel SGX.

- The 1.58-*bit Parity Horizon*: an empirical characterisation of the critical ternary scaling threshold, with direct implications for sovereign hardware design.

2. Definition of *Intelligentia Artificialis Privata*

2.1. Formal Definition

Definition 1 (Privata System) *Let a computational system* $\mathcal{A}$ *be Privata if and only if it simultaneously satisfies the following three constraints*.

Locality Constraint. The system performs no network communication during inference operations. All computation is resolved on user-controlled hardware:

$$\mathcal{C}_{\text{net}} = 0 \tag{1}$$

Data Sovereignty. The internal data domain is disjoint from all external data domains. No user data or inference context is transmitted to, stored in, or processable by external parties:

$$D_{\text{external}} \cap D_{\text{internal}} = \varnothing \tag{2}$$

Control Closure. All model updates, knowledge modifications, and configuration changes are exclusively within the control of the designated user or administrator. No remote entity can unilaterally alter system state:

$$\forall U \in \text{Updates},\ U \in \mathcal{U}_{\text{user}} \tag{3}$$

These three constraints are jointly necessary and sufficient: a system satisfying all three is Privata; a system failing any one is not, regardless of other privacy-enhancing properties it may possess.

2.2. Key Properties

Proposition 1 *From constraints* (1)-(3), *the following properties are directly derivable*.

- **Isolation.** The system's operational state cannot be observed or modified by external parties during normal operation.
- **Deterministic Controllability.** Given identical inputs and system state, behavior is reproducible and verifiable by the user. This property is uniquely amenable to formal verification (Section 10): because the knowledge manifold is closed, the system presents a stable target for machine-checked proofs of behavioural conformance—an advantage unavailable to epistemically volatile cloud systems.
- **Epistemic Containment.** The system's knowledge manifold cannot be altered without explicit user action. Knowledge state at any time t is a deterministic function of the initial state and authorized user updates only.
- **Auditability.** All system behavior can be audited by the user, as all relevant state is locally accessible.
- **Resilience to Provider Actions.** The system continues to function regardless of provider business decisions, service discontinuation, policy changes, or legal orders directed at providers.

2.3. Boundary Cases

Several important boundary cases clarify the definition:

- **Hybrid or Semi-Private Systems.** Systems that perform some inference locally but phone home for model updates, telemetry, or context augmentation are not Privata. They may be privacy-improving but do not provide the architectural guarantees of the Privata model.
- **Federated Systems.** Federated learning systems achieve privacy during training but typically require network connectivity during inference or for model synchronization, satisfying neither (1) nor (3) as defined here.
- **Encrypted Cloud Inference.** Homomorphic encryption approaches [7] allow computation on encrypted data in the cloud, providing confidentiality but not locality or control closure.

2.4. Design Constraints of Privata AI Systems

The three formal constraints of Definition 1 can be operationalised as five practical design constraints that any compliant implementation must satisfy:

- **Data Locality.** All user data, queries, and inference context remain within user-controlled storage. No data crosses the trust boundary to external infrastructure, whether in plaintext or encrypted form.
- **Compute Locality.** Model inference executes on user-controlled hardware. Outsourcing computation to cloud accelerators or third-party enclaves, even under confidentiality guarantees, violates this constraint.
- **Trust Boundary Enforcement.** A verifiable boundary separates the Privata system from external networks during operation. This boundary must be enforced at the operating-system level or below, not merely at the application layer.
- **User-Controlled Knowledge State.** The system's epistemic content changes only through user-authorised update events. Autonomous or provider-initiated model modifications are outside the trust boundary.
- **Optional External Synchronisation.** Controlled, user-initiated update channels are permitted when inference is offline, update packages are cryptographically signed, and the Two-Stage Protocol (Section 3.2.4) is followed. Synchronisation is never automatic and never touches the inference state directly.

These constraints are not independent aspirations but entailments of the formal definition: a system violating any one of them fails to satisfy at least one of Equations (1)-(3).

3. Architectural Framework

3.1. Core Stack

A minimal Privata system comprises four interdependent layers [8]. Each layer has distinct responsibilities and trust boundaries, and each must independently satisfy the constraints of Definition 1.

1) **Inference Core**—Local model execution engine; network access disabled at OS level.

2) **Memory Layer**—Private embeddings and vector retrieval; all retrieval is local.

3) **Control Interface**—User-facing surface with full audit and configuration access.

4) **Update Mechanism**—Offline, cryptographically signed model and knowledge updates.

3.2. Functional Layer Specifications

3.2.1. Inference Core

The inference core executes forward passes on local hardware (CPU, GPU, or dedicated NPU) with network access disabled at the operating system level—not merely unenabled by application configuration—to satisfy constraint (1). Current inference engines suitable for Privata deployment include llama.cpp [9], MLC-LLM [10], and Ollama. For deployments targeting ternary models, the inference core must implement NPU-native ternary logic (e.g., bit-packed XOR-accumulate operations on the Snapdragon Hexagon) rather than falling back to lookup-table emulation, which eliminates the efficiency gains of the 1.58-bit format.

3.2.2. Memory Layer

The memory layer provides persistent, semantically addressable storage via a local vector database (e.g., Chroma, Qdrant, or Weaviate in embedded mode) combined with a locally running embedding model. This layer enables Retrieval-Augmented Generation [11] without compromising constraint (2), since the retrieval corpus is entirely user-controlled.

3.2.3. Control Interface

The control interface must expose complete audit capabilities: access to all stored state, inference logs, model configuration parameters, and update history. No telemetry or usage data may be transmitted to external parties.

3.2.4. Update Mechanism

All updates must be user-initiated, cryptographically signed, and applied via a user-auditable process with rollback capability [12]. A naive fully-manual update workflow creates a genuine UX barrier to consumer adoption. We address this with a *Two-Stage Update Protocol*:

1) **Background Sync:** A sandboxed, non-privileged process downloads a GPG-signed update blob. No update is applied; the inference engine and memory layer remain isolated.

2) **Local Attestation:** The HBS protocol (Section 3.4) verifies the blob's hash against a hardware-isolated whitelist, then requires explicit physical user confirmation before activation:

$$\text{Apply}(U) \Leftrightarrow \left(\text{Verify}_{\text{HBS}}(U) \wedge \mathcal{I}_{\text{human}} = 1\right)$$

This framing presents the verification step as a transparency feature—a visible, deliberate transition in knowledge state—rather than friction. It directly mitigates

the Update Channel Poisoning threat (Section 4) at the protocol level while preserving the full guarantees of Control Closure (3).

3.3. Air-Gapped vs. Semi-Isolated Models

Privata systems exist on a sovereignty spectrum defined by their update topology. **Fully Air-Gapped Systems** achieve maximum sovereignty by eliminating all external network access, updating via physical media with offline cryptographic verification. **Semi-Isolated Systems** permit controlled, user-initiated synchronisation events implementing the Two-Stage Protocol above; the device is fully offline during inference. **Inference-Only Offline Systems** perform inference always offline, while a separately isolated process handles update connectivity without access to inference activity. The semi-isolated configuration represents the most practically deployable option for near-term consumer adoption.

3.4. Hardware Trust Layer and Hardware-Bound Sealing

A hardware trust layer leverages TPM 2.0 [13] or ARM TrustZone to provide attestation of model integrity. We refer to the composition of hardware-derived identity via Physical Unclonable Functions (PUFs) and locally enforced inference sealing as *Hardware-Bound Sealing* (HBS). This designation does not introduce a new cryptographic primitive; rather, it composes established hardware-rooted trust mechanisms with local inference integrity enforcement, building upon PUF-based key derivation [14] and trusted execution environments [15]. **Table 2** summarises the architectural differences.

Table 2. Comparative security properties: HBS versus existing attestation mechanisms.

Property	TPM 2.0	Intel SGX	HBS (Privata)
Root of Trust	CA-Signed Key	Intel-Signed Key	PUF (In-Silicon)
Verification Locus	Remote Attestation	Remote Attestation	Local Only
Jurisdictional Risk	Present	Present	None
Human-in-the-Loop	No	No	Required ($\mathcal{I}_{\text{human}} = 1$)

HBS operates through three steps. **Key Derivation:** the NPU derives a device-unique cryptographic root via the silicon's PUF:

$$K_{\text{device}} = \mathcal{F}_{\text{PUF}}(C) \tag{4}$$

where $\mathcal{F}_{\text{PUF}}$ is the Physical Unclonable Function, C is a session challenge vector, and K_{device} is the resulting hardware-isolated key. This serves as entropy for a hardware-isolated PRNG. The root of trust is In-Silicon and cannot be spoofed by a virtualised environment. **Authorisation Comparison:** the derived key is verified against authorised public keys stored in a hardware-isolated enclave, confirming execution on unmodified hardware. **Human-in-the-Loop Authorisation:** we formalise this as an interrupt $\mathcal{I}_{\text{human}} \in \{0,1\}$ such that inference state

$S_{t+1} = S_t$ until $\mathcal{I}_{\text{human}} = 1$. The system pauses after a successful hardware match and requires physical user input to proceed; no autonomous or remote process can advance inference, designed to prevent hijacking even under complete OS compromise.

This protocol is designed to satisfy Control Closure (3) at the hardware level. The $\mathcal{I}_{\text{human}}$ interrupt structure provides *Observational Determinism*: the closed knowledge manifold $K(t)$ is not merely a policy commitment but a state whose transitions are gated by verifiable physical action [16].

Adversary Model. The HBS security claims are stated against a Dolev-Yao adversary capable of: compromising the host operating system; observing memory and I/O channels; and injecting external prompts or data. The adversary is explicitly assumed *incapable* of replicating the PUF response (4) or extracting sealed keys across the hardware boundary. All security properties hold conditional on these trust boundary assumptions.

Formal Verification Pathway. Full machine-checked proof of the three HBS security properties—secrecy of K_{device}, authentication of $\mathcal{I}_{\text{human}}$, and non-interference of sealed inference state—is identified as future work. The natural proof methodology is symbolic protocol verification using ProVerif or Tamarin Prover against the above adversary model, with $\mathcal{F}_{\text{PUF}}$ treated as a random oracle (standard practice in PUF-based protocol literature [14]). The formalised protocol structure presented here is designed to be directly encodable in applied pi-calculus, providing a clear pathway to that verification.

4. Threat Model

4.1. Threat Modeling Methodology

We apply the STRIDE framework [17] adapted for AI architectures, supplemented by MITRE ATLAS [18]. Our analysis focuses on threats uniquely salient for locally-deployed systems.

4.2. Adversary Classes

- **Cloud or Model Provider (mitigated by architecture).** Privata architecture eliminates provider visibility into inference by design. Because $\mathcal{C}_{\text{net}} = 0$ during inference, any instruction to the model is inherently local-user-sourced, eliminating Indirect Prompt Injection vulnerabilities endemic to cloud-connected agents [19].
- **OS-Level Compromise.** A compromised OS can exfiltrate model weights, prompts, or memory contents. Mitigation requires HBS attestation, reproducible builds with verified checksums, and SELinux/AppArmor mandatory access controls.
- **Side-Channel Observers.** Physical or electromagnetic observation may leak inputs via timing, power, or emanation channels [20]. High-assurance deployments require TEMPEST-class shielding.
- **Update Channel Poisoning.** Malicious update injections can corrupt model

behaviour or insert backdoors [21]. The Two-Stage Update Protocol with multi-key provenance verification mitigates this at the architectural level.

- **Model Weight Extraction.** Cold-boot attacks, DMA, or storage cloning may extract weights. Full-disk encryption, AMD SME/SEV memory encryption, and hardware-bound key storage are required countermeasures.
- **Adversarial Prompt Injection.** Malicious documents processed by the system may affect outputs even in a fully local deployment [19]. Users must apply document provenance controls to the local RAG corpus.

4.3. Security Guarantees

A fully compliant Privata system is designed to provide: no outbound inference data transmission (architectural); verifiable model integrity via cryptographic weight hashing; reproducible builds for independent verification; complete local audit trails; and resilience to provider-side compulsion. These properties are *architectural* rather than policy-based, holding by construction regardless of provider behaviour.

5. Computational Constraints and Optimisation

5.1. Inherent Tradeoffs

The locality constraint bounds available compute to user-owned hardware. A single NVIDIA A100 80 GB provides approximately 312 TFLOPS (FP16); a high-end RTX 4090 provides approximately 82 TFLOPS; Apple M3 Ultra approximately 36 TOPS (32-core Neural Engine), with 819 GB/s unified memory bandwidth. These are genuine costs [22]. However, the *Intelligence-per-Watt* (IPW) metric—effective reasoning throughput per unit energy—shows that local accelerators such as the Apple M4 Max (0.8× cloud reference) and Snapdragon X2 Elite (0.7×) are already approaching enterprise-class NVIDIA B200 efficiency for single-query tasks [23], with a 5.3× improvement in local intelligence efficiency observed between 2023 and 2025. The gap is real but narrowing rapidly.

5.2. Quantisation and Model Compression

Quantisation reduces weight precision to lower-bit-width integer representations [24]. GPTQ [25] and GGUF formats achieve 4-bit or 8-bit precision with modest degradation. Dettmers *et al.* report low-bit weight quantisation incurs only modest accuracy degradation relative to FP16 baselines [26], enabling 7B - 13B models on consumer hardware at 10 - 50 tokens/second.

5.3. Ternary Weights and the Parity Horizon

A more radical strategy employs *ternary* weights drawn from $\mathcal{W} = \{-1, 0, 1\}$, with information density:

$$D = \log_2(3) \approx 1.58 \text{ bits per weight} \tag{5}$$

This reduces the attention mechanism from $O(n^2)$ floating-point multiplica-

tions to $O(n^2)$ integer additions and sign-flips [27], which is the mathematical basis for NPU efficiency gains.

Two implementation requirements are critical. First, ternary models must be *trained from scratch* using absmax activation quantisation and Sub-Layer Normalisation (SubLN) [27]; post-training application of ternary quantisation risks accuracy collapse because *emergent outliers*—high-magnitude hidden states carrying critical semantic information—are clipped to ±1 or zero. Second, the inference stack must execute via NPU-native in-register ternary arithmetic; lookup-table fallbacks negate the efficiency gains.

Based on BitNet b1.58 scaling results [27], we characterise the *1.58-bit Parity Horizon*: the performance gap between ternary and full-precision models undergoes gradual convergence with scale, with extrapolation of the BitNet b1.58 scaling trend [27] suggesting functional parity at approximately an order of magnitude beyond the published experiments—we take ~13.4 billion parameters as a working estimate. Ma *et al.* demonstrate parity at scales up to 3.9B [27]; the trend continues monotonically across their reported range. This is an empirical scaling law, not a discrete threshold. Below this parameter count, precision loss is perceptible on complex benchmarks; above it, the additional parameters compensate for per-weight precision reduction. At 13.4B, a ternary model requires approximately 2.6 GB of storage versus 6.7 GB for 4-bit PTQ and 26.8 GB at FP16—the smallest memory footprint of any practical quantisation regime at equivalent perplexity. Ternary weights are also naturally compatible with semi-structured N:M sparsity, yielding a further 1.30× inference speedup when combined with MoE activation patterns [27].

Table 3 summarises memory footprint and accuracy retention across these quantisation regimes.

Table 3. Quantisation regimes at 13.4 B parameters: memory footprint versus accuracy retention. Below the Parity Horizon, 4-bit PTQ is optimal for bits-vs-accuracy; above it, ternary achieves the smallest footprint at equivalent perplexity.

Regime	Bit-Width	Memory (13.4B)	Accuracy vs. FP16
FP16 Baseline	16-bit	~26.8 GB	100% (reference)
4-bit PTQ (NF4)	4-bit	~6.7 GB	97% - 99% [26]
1.58-bit (Ternary)	1.58-bit	~2.6 GB	≈99% at ≥13.4B [27]
Binary	1-bit	~1.6 GB	Significant degradation

5.4. Sparse Architectures and Mixture-of-Experts

MoE architectures [28] activate only a subset of parameters per inference pass. Mixtral 8 × 7B [29] achieves performance competitive with dense 70B models at the computational cost of a 13B model. Combined with ternary quantisation, MoE delivers further IPW improvements for Privata deployments, as only active expert subsets require ternary arithmetic per step.

5.5. Local Retrieval-Augmented Generation

RAG extends effective knowledge capacity by retrieving from a local corpus [11]. In the Privata context the RAG corpus is user-controlled; capacity is bounded by local storage, not parameter count. Retrieval, embedding, and indexing occur entirely within the local domain, violating no Privata constraint.

5.6. The NPU-Centric Paradigm Shift

The 2026 Snapdragon X2 Elite's sixth-generation Hexagon NPU delivers 80 TOPS (INT8) [30]; on this basis the present analysis projects an effective throughput of approximately 85 TOPS-equivalent for 1.58-bit ternary weights [30], using integer addition and sign-flipping rather than floating-point MAC. Apple Silicon's unified memory architecture enables M3 Ultra (192 GB) to run 70B+ models at cloud-competitive latency [31]. The trajectory of dedicated inference silicon indicates the capability gap will narrow substantially within five years, though not close entirely for the largest model scales.

5.7. Thermodynamics of Sustained Inference

The author's energy model estimates approximately 11 mJ/token on mobile NPUs and 27 mJ/token on workstation NPUs. A 10,000-token reasoning chain therefore requires less than 5% of a standard laptop battery, removing the final practical energy objection to fully offline sovereign inference for mobile and field deployments.

6. Epistemic Boundaries

6.1. The Closed Knowledge Manifold

Define the system's knowledge state at time t as:

$$K(t) = K_0 + \Delta K_{\text{local}} \tag{6}$$

where K_0 is the initial deployment state and ΔK_{local} represents only locally-applied, user-authorised updates. There is no external drift term. Contrast with a cloud-deployed system:

$$K_{\text{cloud}}(t) = K_0 + \Delta K_{\text{provider}}(t),$$

where $\Delta K_{\text{provider}}(t)$ is opaque to the user and continuously evolving [3]—the Forensic Vacuum of Section 1.1 formalised.

6.2. Implications of Knowledge Closure

- **Covert Bias Injection Prevention.** Any change to knowledge or behavioural tendencies requires explicit user action, leaving an auditable trace.
- **Stable Reasoning Domain.** Users can develop a complete, stable understanding of the system's capabilities and limitations over time.
- **Genuine Intellectual Ownership.** The user's relationship to the system is that of an owner to a tool, not a subscriber to a service that may change without notice.
- **Reproducibility for Accountability.** Legal reasoning, medical decision sup-

port, and financial analysis require the ability to reproduce a specific inference given recorded inputs and a fixed model state. Knowledge closure is a necessary condition.

- **Epistemic Containment as Safety Feature.** The knowledge currency limitation, often framed as a cost, is equally a safety property: a system with a bounded, stable knowledge state provides a hard limit on adverse behaviours arising from opaque retraining. There is, in effect, a plug to pull—an alignment with approaches to AI safety that favour deterministic controllability over continuous capability expansion [16].

6.3. Knowledge Currency Tradeoff

The primary cost of knowledge closure is currency: a Privata system does not automatically incorporate post-deployment developments. This tradeoff is already explicitly made in many professional contexts—legal databases, medical references, and engineering standards are updated on defined cycles precisely because stability and currency are in tension [32]. The Two-Stage Update Protocol provides a principled mechanism for managing this tradeoff without compromising sovereignty.

7. Comparative Analysis

7.1. Comparison with Cloud AI Systems

Table 4 summarises the key differences between Cloud AI and Privata AI across ten dimensions.

Table 4. Comparative properties of cloud AI versus Privata AI systems.

Property	Cloud AI	Privata AI
Knowledge source	Dynamic, external, opaque	Fixed at deployment + local updates
Privacy model	Probabilistic (policy-based)	Deterministic (architectural)
Control locus	Provider/operator	User (full sovereignty)
Update model	Remote, continuous, opaque	Offline, signed, user-initiated
Epistemic drift	Continuous, uncontrolled	Bounded, user-authorised
Data sovereignty	Delegated to provider	Absolute (local only)
Auditability	Limited/contractual	Complete (all state local)
Regulatory compliance	Policy compliance	Architectural necessity
Failure mode	Provider outage, policy change	Hardware failure only
Business model	Adverse (data extraction)	Aligned (user interests)

7.2. Relation to Privacy-Preserving Machine Learning

Privata AI is distinct from, though related to, several existing paradigms [33]:

- **Federated Learning (FL).** FL [34] keeps training data local but does not provide sovereignty over inference or knowledge state: model aggregation occurs centrally, and inference may still occur against centrally-aggregated models. FL satisfies a weaker form of data sovereignty but not control closure or inference-time locality.
- **Differential Privacy (DP).** DP [35] provides statistical guarantees against individual data extraction but does not eliminate centralised infrastructure dependency. DP is complementary to Privata—a Privata system may employ DP in its memory layer for defence in depth.
- **Homomorphic Encryption (HE).** HE [7] addresses confidentiality but not locality or control closure; computation still occurs on external infrastructure.
- **Edge AI.** Edge AI [36] and Privata AI share local computation but differ in *primary architectural goal*: Edge AI treats locality as a performance and latency optimisation, whereas Privata AI treats it as a mechanism for user sovereignty and epistemic containment. Edge deployments typically remain under provider management with automatic updates that violate (3); this is a difference of architectural intent, not implementation detail.
- **Secure Multi-Party Computation (SMPC).** SMPC [37] addresses confidentiality but not locality or control closure.

8. Economic and Political Implications

8.1. Decentralisation of Intelligence Infrastructure

Privata AI adoption at scale would constitute a structural decentralisation analogous to the PC revolution of the 1980s [38]: AI capability would transition from a rented service to personal infrastructure, eliminating the dependency relationships that currently concentrate economic and epistemic power in a small number of providers.

8.2. The Crisis of Data-Rentierism and the Sovereign AI Economy

Many AI deployments depend on passive accumulation of user interaction data to improve models and target advertising [39]. Privata architecture is designed to eliminate this mechanism by construction—not through policy compliance, which is easily bypassed, but through the architectural separation of inference from external data flows. This represents a *Crisis of Data-Rentierism* for incumbents whose value derives primarily from interaction data.

The disruption creates space for a *Sovereign AI Economy* with fundamentally different value sources: hardware sales of NPU-optimised sovereign inference devices; subscription services for cryptographically signed and audited model weight updates; zero-knowledge maintenance contracts providing ongoing security verification without user data access; and certified domain-specific knowledge update

packages. This economic model structurally aligns provider incentives with user interests.

8.3. Regulatory Alignment by Design

Privacy-by-design principles formalised in GDPR Article 25 [40] and the privacy engineering literature [41] require data protection to be incorporated architecturally rather than applied as an afterthought. Privata AI satisfies this at the deepest level: data minimisation, purpose limitation, and user control are achieved through architectural constraints that preclude violation by design. The EU AI Act's requirements for transparency, auditability, and human oversight [4] are likewise satisfied by Privata's complete local audit trails and user control over system state.

8.4. Geopolitical and National Security Dimensions

Concentration of AI infrastructure in a small number of jurisdictions creates strategic risks [42]: legal changes, export controls, or sanctions can disrupt AI capabilities across dependent nations and organisations. Privata AI provides resilience by eliminating external infrastructure dependency. Nations can develop an *Intelligence Strategic Reserve*—locally possessed model weights and sovereign NPU hardware—insulating critical reasoning infrastructure from foreign jurisdictional overreach. This dimension is particularly salient for defence, critical infrastructure, and government applications.

9. Implementation Pathways

9.1. Near-Term (Present-3 Years)

Current hardware (Apple M-series MacBooks, consumer GPU workstations with 24 GB + VRAM) already runs quantised competitive models at practical speeds [43]. The software ecosystem has matured: llama.cpp, Ollama, and LM Studio provide deployment frameworks; Chroma and Qdrant provide local vector databases; Open WebUI provides user-facing interfaces. Primary remaining barriers are UX maturity, update package distribution infrastructure, and hardware attestation tooling—the first of which the Two-Stage Update Protocol directly addresses.

9.2. Mid-Term (3 - 10 Years)

Dedicated consumer AI inference chips will expand the Privata capability frontier. NPUs integrated into consumer devices already provide order-of-magnitude tokens-per-watt improvements over general-purpose CPUs. Over this horizon, 70B-class models should become routinely deployable on high-end consumer hardware.

9.3. Long-Term (10+ Years)

The long-term horizon envisions fully sovereign AI ecosystems integrated with

zero-trust digital identity systems, capable of participating in federated reasoning networks while maintaining strict local sovereignty. This intersects with questions of long-duration AI systems [44], persistent private agents, and the legal and philosophical status of AI systems with stable, user-owned identity maintained over years to decades.

10. Future Directions

Several open research problems are identified as priorities for the Privata AI research agenda:

- **Formal Verification of AI Behaviour.** Methods for formally verifying that a deployed model's behaviour conforms to specified properties, enabling automated auditability. The *Deterministic Controllability* property of Privata systems makes them uniquely tractable targets: a closed knowledge manifold presents a stable, fixed specification against which machine-checked proofs can be constructed, unlike the continuously drifting state of cloud-deployed models. Recent work in model specification [16] and formal analysis of transformer architectures provides initial foundations.
- **Cryptographic Memory Sealing.** Techniques for cryptographically binding memory state to hardware identity, preventing private embedding extraction even under OS compromise.
- **Self-Auditing AI Systems.** Architectures enabling AI systems to produce verifiable accounts of their own reasoning, supporting user oversight without specialised expertise. Mechanistic interpretability research [45] provides relevant foundations.
- **Long-Duration Private AI.** The conditions under which a Privata system maintains coherent behaviour over years to decades, including the relationship between knowledge closure and epistemic stability [46].
- **Secure Update Package Ecosystems.** Infrastructure for distributing cryptographically signed model update packages that enable knowledge currency without compromising sovereignty. Certificate transparency mechanisms [47] may provide design patterns.
- **Privata Multi-Agent Systems.** Architectures enabling collaboration between multiple Privata AI systems while maintaining individual sovereignty guarantees. Zero-knowledge proofs [48] offer potential building blocks.

11. Conclusion

We have introduced *Intelligentia Artificialis Privata* as a formally defined category of AI systems distinguished by physical locality, operational autonomy, and knowledge closure. These properties jointly provide security, privacy, and epistemic guarantees that are architecturally necessary rather than policy-contingent, satisfying current and emerging regulatory frameworks by construction.

The comparative analysis demonstrates that Privata AI is categorically distinct from federated learning, differential privacy, homomorphic encryption, and—

critically—edge AI, whose primary architectural goal is performance optimisation rather than user sovereignty. The formalisation of the HBS protocol with its PUF-derived root of trust and $\mathcal{I}_{\text{human}}$ interrupt guarantee, combined with the empirical characterisation of the 1.58-bit Parity Horizon and its training-from-scratch requirements, provides the technical foundations for provably secure, practically deployable sovereign systems. The economic analysis demonstrates that Privata architecture does not merely disrupt existing data-extractive models but creates the conditions for a Sovereign AI Economy structurally aligned with user interests.

The centralization of AI infrastructure is not an inevitable consequence of the technology but a product of economic incentives and path dependencies. The current state of hardware, quantisation techniques, and local inference software demonstrates that the alternative is architecturally coherent, practically implementable, and increasingly accessible.

Intelligentia Artificialis Privata is not a feature—it is a category shift in how intelligence is instantiated, owned, and constrained. It represents the application of the personal computing insight to the age of AI: that the most powerful and trustworthy technology is technology you own.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., *et al.* (2020) Language Models Are Few-Shot Learners. arXiv: 2005.14165. https://arxiv.org/abs/2005.14165

[2] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., *et al.* (2010) A View of Cloud Computing. *Communications of the ACM*, **53**, 50-58. https://doi.org/10.1145/1721654.1721672

[3] Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., *et al.* (2022) Predictability and Surprise in Large Generative Models. 2022 *ACM Conference on Fairness Accountability and Transparency*, Seoul, 21-24 June 2022, 1747-1764. https://doi.org/10.1145/3531146.3533229

[4] European Parliament (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act) Official Journal of the European Union, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[5] Biden, J.R. (2023) Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Federal Register. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[6] Westin, A.F. (1967) Privacy and Freedom. Atheneum. https://archive.org/details/privacyfreedom00west

[7] Gentry, C. (2009) A Fully Homomorphic Encryption Scheme. Ph.D. Thesis, Stanford University. https://crypto.stanford.edu/craig/craig-thesis.pdf

[8] Shi, W., Cao, J., Zhang, Q., Li, Y. and Xu, L. (2016) Edge Computing: Vision and

Challenges. *IEEE Internet of Things Journal*, **3**, 637-646. https://doi.org/10.1109/jiot.2016.2579198

[9] Gerganov, G. (2023) llama.cpp: Port of Facebook's LLaMA model in C/C++. GitHub Repository. https://github.com/ggerganov/llama.cpp

[10] MLC Team (2023) MLC-LLM: Universal LLM Deployment Engine with ML Compilation. GitHub Repository. https://github.com/mlc-ai/mlc-llm

[11] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., *et al.* (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv: 2005.11401. https://arxiv.org/abs/2005.11401

[12] Bellissimo, A., Burgess, J. and Fu, K. (2006) Secure Software Updates: Disappointments and New Challenges. https://www.usenix.org/legacy/event/hotsec06/tech/full_papers/bellissimo/bellissimo.pdf

[13] Trusted Computing Group (2019) Trusted Platform Module Library Specification, Family 1.2, Level 2, Revision 116. TCG. https://trustedcomputinggroup.org/resource/tpm-library-specification/

[14] Maes, R. (2013) Physically Unclonable Functions: Constructions, Properties and Applications. Springer. https://doi.org/10.1007/978-3-642-41395-7

[15] Sabt, M., Achemlal, M. and Bouabdallah, A. (2015) Trusted Execution Environment: What It Is, and What It Is Not. 2015 *IEEE Trustcom/BigDataSE/ISPA*, Helsinki, 20-22 August 2015, 57-64. https://doi.org/10.1109/trustcom.2015.357

[16] Dalrymple, D., Seshia, S.A., *et al.* (2024) Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. arXiv: 2405.06624. https://arxiv.org/abs/2405.06624

[17] Shostack, A. (2014) Threat Modeling: Designing for Security. Wiley.

[18] MITRE Corporation (2023) MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. https://atlas.mitre.org

[19] Perez, F. and Ribeiro, I. (2022) Ignore Previous Prompt: Attack Techniques for Language Models. arXiv: 2211.09527. https://arxiv.org/abs/2211.09527

[20] Kocher, P., Horn, J., Fogh, A., Genkin, D., Gruss, D., Haas, W., *et al.* (2019) Spectre Attacks: Exploiting Speculative Execution. 2019 *IEEE Symposium on Security and Privacy* (*SP*), San Francisco, 19-23 May 2019, 1-19. https://doi.org/10.1109/sp.2019.00002

[21] Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., *et al.* (2023) Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 1563-1580. https://doi.org/10.1109/tpami.2022.3162397

[22] Chitty-Venkata, K.T., Mittal, S., Emani, M., Vishwanath, V. and Somani, A.K. (2023) A Survey of Techniques for Optimizing Transformer Inference. *Journal of Systems Architecture*, **144**, Article ID: 102990. https://doi.org/10.1016/j.sysarc.2023.102990

[23] Saad-Falcon, J., Narayan, A., Akengin, H.O., Griffin, J.W., Shandilya, H., Gamarra Lafuente, A., et al. (2025) Intelligence per Watt: Measuring Intelligence Efficiency of Local AI. arXiv: 2511.07885. https://arxiv.org/abs/2511.07885

[24] Gholami, A., Kim, S., Dong, Z., Yao, Z.W., Mahoney, M.W. and Keutzer, K. (2022) A Survey of Quantization Methods for Efficient Neural Network Inference. arXiv: 2103.13630. https://arxiv.org/abs/2103.13630

[25] Frantar, E., Ashkboos, S., Hoefler, T. and Alistarh, D. (2023) GPTQ: Accurate Post-

training Quantization for Generative Pre-Trained Transformers. arXiv: 2210.17323. https://arxiv.org/abs/2210.17323

[26] Dettmers, T., Lewis, M., Belkada, Y. and Zettlemoyer, L. (2022) LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. arXiv: 2208.07339. https://arxiv.org/abs/2208.07339

[27] Ma, S.M., Wang, H.Y., Ma, L.X., *et al.* (2024) The Era of 1-Bit LLMs: All Large Language Models Are in 1.58 Bits. arXiv: 2402.17764. https://arxiv.org/abs/2402.17764

[28] Shazeer, N., Mirhoseini, A., Maziarz, K., *et al.* (2017) Outrageously Large Neural Networks: The Sparsely-Gated Mixture-Of-Experts Layer. arXiv: 1701.06538. https://arxiv.org/abs/1701.06538

[29] Jiang, A.Q., Sablayrolles, A., Roux, A., *et al.* (2024) Mixtral of Experts. arXiv: 2401.04088. https://arxiv.org/abs/2401.04088

[30] Qualcomm Technologies, Inc. (2025) New Snapdragon X2 Elite Extreme and Snapdragon X2 Elite are the Fastest and Most Efficient Processors for Windows PCs. Press Release. https://www.qualcomm.com/news/releases/2025/09/new-snapdragon-x2-elite-extreme-and-snapdragon-x2-elite-are-the-

[31] Apple Inc (2023) Apple Unveils M3, M3 Pro, and M3 Max, the Most Advanced Chips for a Personal Computer. Apple Newsroom. https://www.apple.com/newsroom/2023/10/apple-unveils-m3-m3-pro-and-m3-max-the-most-advanced-chips-for-a-personal-computer/

[32] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the* 2021 *ACM Conference on Fairness, Accountability, and Transparency*, 3-10 March 2021, 610-623. https://doi.org/10.1145/3442188.3445922

[33] Mireshghallah, F., Taram, M., Ramrakhyani, P., Jalali, A., Tahoori M.B., and Esmaeilzadeh, H. (2020) Privacy in Deep Learning: A Survey. arXiv: 2004.12254. https://arxiv.org/abs/2004.12254

[34] McMahan, B., Moore, E., Ramage, D., Hampson, S. and Agüera y Arcas, B. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv: 1602.05629. https://arxiv.org/abs/1602.05629

[35] Dwork, C. and Roth, A. (2014) The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, **9**, 211-487. https://doi.org/10.1561/0400000042

[36] Li, E., Zhou, Z. and Chen, X. (2018) Edge Intelligence: On-Demand Deep Learning Model Co-Inference with Device-Edge Synergy. arXiv: 1806.07840. https://arxiv.org/abs/1806.07840

[37] Goldreich, O. (2004) Foundations of Cryptography, Volume 2: Basic Applications. Cambridge University Press.

[38] Ceruzzi, P.E. (2003) A History of Modern Computing. 2nd Edition, MIT Press.

[39] Zuboff, S. (2019) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.

[40] European Parliament (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation) Official Journal of the European Union, 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[41] Cavoukian, A. (2009) Privacy by Design: The 7 Foundational Principles. Information and Privacy Commissioner of Ontario. https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf

[42] Allen, G. and Chan, T. (2017) Artificial Intelligence and National Security. Study, Belfer Center for Science and International Affairs, Harvard Kennedy School. https://www.belfercenter.org/publication/artificial-intelligence-and-national-security

[43] Touvron, H., Martin, L., Stone, K., *et al.* (2023) Llama 2: Open Foundation and Fine-tuned Chat Models. arXiv: 2307.09288. https://arxiv.org/abs/2307.09288

[44] Krakovna, V., Martic, M., Togelius, J., Leike, J. and Legg, S. (2020) Avoiding Side Effects in Complex Environments. arXiv: 2006.06547. https://arxiv.org/abs/2006.06547

[45] Elhage, N., Hume, T., Chan, C., *et al.* (2022) Toy Models of Superposition. Transformer Circuits Thread. https://transformer-circuits.pub/2022/toy_model/index.html

[46] Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., *et al.* (2021) A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 3366-3385. https://doi.org/10.1109/tpami.2021.3057446

[47] Laurie, B., Langley, A. and Kasper, E. (2013) Certificate Transparency. RFC 6962, IETF. https://www.rfc-editor.org/info/rfc6962

[48] Ben-Or, M., Goldwasser, S. and Wigderson, A. (1988) Completeness Theorems for Non-Cryptographic Fault-Tolerant Distributed Computation. *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing—STOC*'88, Chicago, 2-4 May 1988, 1-10. https://doi.org/10.1145/62212.62213

Appendix

A. Minimal System Specification

A compliant Privata system requires at minimum:

- **Hardware.** x86-64 or ARM64 system with ≥16 GB RAM; GPU or NPU with 8+ GB VRAM recommended for 7B+ models.
- **Model.** Ternary (1.58-bit, trained from scratch with SubLN) or quantised LLM (4-bit or 8-bit GGUF), 7B - 13B parameters, stored locally with SHA-256 hash verification.
- **Storage.** 20 - 100 GB local NVMe storage for model weights and vector database.
- **Update mechanism.** Two-Stage Update Protocol with GPG-signed packages; no automatic or network-initiated inference modification; rollback capability required.
- **Network.** Zero dependency during inference; optional controlled sync for Stage 1 updates only.
- **Audit capability.** Complete local logging of all inference activity, configuration changes, and update events.

B. Reference Implementation Sketch

A minimal reference implementation comprises three main components:

1) **Local LLM inference engine.** llama.cpp or Ollama, with network access disabled at the OS level via iptables/nftables or macOS pf. Model weights SHA-256 verified before loading. For ternary models, the engine must support NPU-native in-register ternary arithmetic; LUT-based fallback is not acceptable.

2) **Local vector database.** Chroma or Qdrant in embedded mode. All documents embedded using a locally-running model (e.g., nomic-embed-text). No cloud API calls.

3) **Control interface.** Open WebUI or equivalent on localhost only (127.0.0.1 binding), with complete local audit logging. Updates via Two-Stage Protocol: sandboxed background download, HBS hash verification, physical user confirmation ($\mathcal{I}_{\text{human}} = 1$) before activation.

This satisfies all three Privata constraints: $\mathcal{C}_{\text{net}} = 0$ during inference; $D_{\text{external}} \cap D_{\text{internal}} = \varnothing$; and all updates $\in \mathcal{U}_{\text{user}}$.