PREM

The Prem Confidential Compute Infrastructure (PCCI)

Verifiable Private AI

Technical White Paper

Confidentiality. Integrity. Transparency.

Prem SA | Lugano, Switzerland | March 2026
premai.io

EXECUTIVE SUMMARY

The Prem Confidential Compute Infrastructure (PCCI) delivers verifiable private AI—enabling enterprises to run AI models within hardware-enforced Trusted Execution Environments (TEEs) and replace legal promises with safeguards rooted in silicon.

PCCI is built on three values of security:

- **Confidentiality:** Execution state and memory pages never leave PCCI enclaves unencrypted. Hardware-level memory encryption covering the full stack guarantees zero visibility to the host OS, hypervisor, datacenter operators, and employees.
- **Integrity:** Every deployed PCCI enclave generates a cryptographic attestation report. This ensures genuine hardware verification and tamper-proof code execution before any data transfer is initiated.
- **Transparency:** Verifiable privacy over vendor policy. Open-source inference codebases combined with immutable public transparency logs provide cryptographic proof of system integrity.

PCCI offers a model-agnostic inference API designed for interoperability across the AI ecosystem. Data is end-to-end encrypted and processed within hardware-sealed TEEs. PCCI cryptographically minimizes data exposure by ensuring plaintext is structurally isolated from Prem, the cloud provider, and anyone with host-level root access. By rooting security in silicon rather than operational policy, insider threats are actively mitigated (see Known Limitations for our transparency on hardware-level constraints).

This white paper details the technical architecture, security model, regulatory compliance framework, and real-world deployments that position PCCI as the foundation for verifiable, sovereign private AI in regulated verticals, including healthcare, defense, financial services, and law.

> *"Standard AI APIs require trust. PCCI removes the need for it. Instead of exposing data to host memory, we enable sovereign, protected execution."*

01 MARKET CONTEXT

The Confidential AI Imperative

The global market for confidential computing in AI infrastructure is experiencing explosive growth:

- $12–25B market size in 2025, with 34%+ CAGR through 2030 (Source: Allied Market Research, Confidential Computing Market Report 2025)
- 70%+ of enterprise AI workloads involve sensitive data (Source: Gartner, AI Infrastructure Trends 2025)
- BFSI sector represents 47% of the current market demand (Source: MarketsandMarkets, Confidential Computing Forecast 2025–2030)
- $300B+ in sovereign AI commitments globally, with confidential computing as a mandated technology (Source: European Commission EU SAFE initiative; US Executive Order on AI Safety)

Organizations can no longer afford data exposure risks during inference. Regulators, customers, public institutions, and private companies increasingly mandate confidential computing as a non-negotiable

requirement for AI deployment.

02 THE PROBLEM

Standard AI APIs Require a Leap of Faith

Pinky-Promise Security

Traditional AI infrastructure exposes data to the host memory for processing. Even with encryption in transit and at rest, plaintext remains completely vulnerable during the inference cycle. Trust is based on legal agreements (DPAs)—not cryptographic guarantees. Provider admins can technically access memory dumps, and bad actors can access data in-use.

Specific Threat Vectors

- **Insider threats:** Cloud admins, infrastructure operators, or malicious employees can access unencrypted data in memory
- **Unauthorized training:** Cloud providers may retain inference data for secondary model training, building competitive intelligence from customer workloads
- **Supply chain attacks:** Compromised infrastructure providers or third-party software can exfiltrate proprietary models and data
- **Regulatory non-compliance:** Many compliance frameworks (HIPAA, PSD3, DORA, EU AI Act) now mandate confidential computing for sensitive workloads

03 THE SOLUTION

Cryptographically Verified Isolation

Prem PCCI provides verifiable end-to-end cryptographic isolation for AI inference, ensuring that:

- **Insider threats are mitigated:** Infrastructure operators are mathematically constrained from accessing decrypted data or model weights. Enclave images contain no SSH access, no debug ports, and no administrative backdoors. All machines operate unattended. To ensure enterprise-grade reliability without compromising isolation, fleet health is managed via external telemetry, and updates are handled strictly through cryptographically signed, immutable image redeployments. Our SRE teams maintain high availability without ever requiring or possessing interactive access to the host environment.
- **Unauthorized training is prevented:** Model weights never appear in plaintext outside the enclave. Data exists only within the sealed hardware environment for the duration of the request and is immediately wiped from memory upon completion.
- **Hardware verification is cryptographically proven:** Attestation proves that inference runs on unmodified, trusted hardware. This is not a policy—it is a constraint the hardware imposes.
- **Zero visibility:** Code execution, data flows, and model internal components remain hidden from infrastructure operators. A rogue employee would have no mechanism to access TEE memory.

04 PROTOCOL

How PCCI Works

Prem PCCI operates through four foundational steps:

1 Client-side encryption

Data is encrypted on your device using XChaCha20-Poly1305 before transmission, with session keys exchanged via XWing—a post-quantum hybrid KEM combining ML-KEM768 and X25519. This means your data is protected against both classical and quantum attacks from the very first handshake. The Prem gateway processes only ciphertext and is solely responsible for authentication and billing. Keys are never shared with Prem infrastructure.

2 Cryptographic attestation

Hardware attestation proves that inference runs on confidential-ready CPUs and NVIDIA GPUs with no unauthorized modifications. Attestation is bound to each session via a unique X-Session-Id header with a five-minute TTL and single-use constraint, preventing replay attacks.

3 Secure execution

The encrypted payload is sent to a Confidential Virtual Machine (CVM). Inside the CVM, data is decrypted, processed by the model, and re-encrypted. For stateless inference, memory is immediately wiped upon completion. For complex enterprise workflows requiring multi-turn context or Retrieval-Augmented Generation (RAG), PCCI supports securely encrypted, ephemeral data stores within the enclave boundary—ensuring proprietary vector indexes can be queried dynamically without ever exposing the underlying data to the host. Plaintext exists in only two places: your device and the hardware-sealed TEE.

4 Codebase transparency

Runtime code is auditable, deterministic, and cryptographically bound to attestation reports. The entire attestation verification stack is open-source, written in Rust, compiled to WebAssembly, and executable directly in the browser—eliminating reliance on any intermediary.

05 ARCHITECTURE

The Confidential Stack

Prem PCCI is organized as a 4-layer architecture:

Layer	Component
Layer 04	**Prem API** — End-to-end encrypted (E2EE) agentic chat, enterprise data connectors, and developer SDKs
Layer 03	**PCCI Runtime** — Enclave-native execution engines optimized for confidential model training, fine-tuning, and inference
Layer 02	**Custom Hypervisor** — Minimal-attack-surface isolation layer strictly optimized to route confidential compute primitives without host-level visibility
Layer 01	**Bare Metal Infrastructure** — Purpose-built hardware utilizing leading TEE-ready silicon, operating in Confidential Computing mode. Physically hosted in compliant datacenters

06 SECURITY MODEL

Three Values of PCCI Security

Confidentiality

Execution state and memory pages never leave the PCCI enclaves unencrypted. Hardware-level memory encryption covering the full stack guarantees zero visibility to the host OS, hypervisor, datacenter operators, and employees.

- CPU as Intel TDX and AMD SEV-SNP
- GPU as NVIDIA Hopper and Blackwell

Integrity

Every deployed PCCI enclave generates a cryptographic attestation report. This ensures genuine hardware verification and tamper-proof code execution before any data transfer is initiated.

- Genuine hardware verification
- Tamper-proof code execution

Transparency

Verifiable privacy over vendor policy. Open-source inference codebases combined with immutable public transparency logs provide cryptographic proof of system integrity.

- Public transparency log
- No hidden backdoors

07 HARDWARE

Cryptographic Primitives

Primitive	Implementation
CPU Enclaves	Full memory encryption and strictly enforced hypervisor isolation via AMD SEV-SNP and Intel TDX
GPU Enclaves	Hardware-isolated environments with accelerator support for high-performance confidential workloads (NVIDIA H100/B800 CC mode)
Attestation	Cryptographic generation and verification of enclave quotes — Hardware Root of Trust rooted in chip manufacturer (AMD, NVIDIA)
Key Management	Secure, enclave-bound cryptographic key generation and lifecycle management. XChaCha20-Poly1305 for data; Post-Quantum ML-KEM hybrid for key exchange

08 POST-QUANTUM CRYPTOGRAPHY

Quantum-Resistant by Design

In March 2026, Google published a formal migration timeline urging the industry to adopt post-quantum cryptography (PQC) by 2029, citing the accelerating threat of cryptographically relevant quantum computers. PCCI is already there.

The “Harvest Now, Decrypt Later” Threat

An adversary can record encrypted traffic today and store it indefinitely. Once a quantum computer capable of breaking classical cryptography becomes available, that adversary can decrypt everything retroactively. This is not a theoretical risk—nation-state actors are widely believed to be harvesting encrypted data at scale right now. For AI inference carrying sensitive enterprise data, the window of vulnerability is already open.

XWing: Two Independent Locks on One Door

PCCI uses XWing—a hybrid Key Encapsulation Mechanism (KEM) that pairs two algorithms simultaneously:

- **ML-KEM768** (NIST FIPS 203): A quantum-resistant lattice-based algorithm specifically designed to withstand attacks from quantum computers.
- **X25519**: A well-established classical elliptic-curve Diffie-Hellman algorithm with decades of cryptanalytic confidence.

An attacker would need to break both algorithms simultaneously to compromise a session. A future quantum computer may defeat X25519, but ML-KEM768 is specifically designed to resist quantum attacks. Conversely, if a weakness were ever discovered in ML-KEM768, X25519 still provides classical security. This dual-lock architecture ensures PCCI traffic is protected against both current and future threats.

Why XWing?

We selected XWing because it has the strongest industry momentum among hybrid KEMs:

- Co-authored by researchers from Cloudflare, SandboxAQ, and Radboud University
- Recommended by Google Cloud for post-quantum migration
- Implemented in Cloudflare's CIRCL cryptography library
- Specified as an IETF Internet-Draft (draft-connolly-cfrg-xwing-kem), built entirely on NIST-standardized primitives
- Designed to be simple, opinionated, and resistant to misconfiguration

Origin: From TEE Data Persistence to Communication Layer

The adoption of XWing emerged from an initial need to manage persistent data securely within TEE environments. We recognized that the same post-quantum protection could be applied to the communication layer, providing a security solution that does not rely solely on TLS configurations that may vary across systems. The result: the communication layer of every PCCI session is quantum-resistant from the first handshake. While current TEE hardware attestation signatures still rely on classical cryptography anchored by chip manufacturers, XWing guarantees that all inference data in transit is strictly protected against 'Harvest Now, Decrypt Later' strategies today.

---

09 COMPARISON

Why Prem AI PCCI

Dimension	Traditional Cloud AI	Prem PCCI
Root of Trust	Legal contracts (DPAs) and vendor promises	Silicon-backed hardware root of trust
Data-in-Use State	Vulnerable to memory scraping & hypervisor compromise	Fully encrypted in volatile memory; isolated via TEEs
Execution Verification	Black-box proprietary runtimes	Deterministic builds with PCR (Platform Configuration Register) matching
Assurance Mechanism	Annual point-in-time manual audits (SOC 2)	Pre-execution remote cryptographic attestation
GPU & Interconnect Security	Plaintext PCIe and standard GPU execution	Hardware-encrypted PCIe & NVLink via NVIDIA Confidential Computing

10 PERFORMANCE

Encryption Without Compromise

A common concern with confidential computing is latency. PCCI is designed to impose negligible overhead:

- **Encryption speed:** XChaCha20-Poly1305 processes over 100,000 tokens per second, while LLM inference typically generates 100–200 tokens per second. Encryption overhead is insignificant relative to model inference time.
- **Compute & bandwidth overhead:** While CPU-bound CVM operations impose a minimal ~5% compute overhead, large-scale LLM inference in Confidential Computing (CC) mode introduces specific PCIe encryption and GPU memory bandwidth constraints. We have heavily optimized our stack to ensure Time-to-First-Token (TTFT) remains minimally impacted. Total generation throughput is engineered to remain highly competitive with unencrypted baselines, and we expect this footprint to shrink further with next-generation architectures like NVIDIA Blackwell.
- **Session handshake:** A brief, one-time delay occurs during the initial key exchange and attestation process. Following this step, streaming performance is virtually indistinguishable from an unencrypted API.
- **API compatibility:** PCCI is OpenAI-compatible. Same API format, same capabilities. The difference is structural—existing code requires zero modification to migrate.

11 INDUSTRY VERTICALS

Where Confidential AI is Mission-Critical

Healthcare & Life Sciences

Growth rate: 35% CAGR. Regulatory drivers: HIPAA, EU MDR.

- **Forum Health case study:** Processing 16M+ patient records with confidential inference to detect disease patterns while maintaining strict HIPAA compliance
- **Drug discovery:** Pharma companies leveraging confidential AI to train on proprietary molecular datasets without exposing compounds to cloud providers

Defense & Government

Major initiatives: EU SAFE €150B commitment, ITAR-regulated workloads.

- **EU SAFE:** Sovereign AI computing initiative requiring confidential infrastructure for government AI workloads
- **Army deployment:** NATO-aligned defense AI systems running on Prem PCCI for intelligence and logistics optimization

Financial Services & Banking

Market share: 47% of current confidential AI demand. Regulators: DORA (Jan 2025), PSD3, CRR III.

- **Fraud detection and trading models:** Banks deploying confidential AI to detect fraud without exposing transaction data
- **Risk models:** Capital calculations running on Prem PCCI while satisfying regulatory audits and PSD3 requirements

Legal & Professional Services

Feb 2026 privilege ruling: Courts recognize AI-assisted legal work as privileged when using confidential computing. 79% of lawyers use AI; only 10% have AI usage policies.

- **Attorney-client privilege preservation:** AI-powered document review and contract analysis on Prem PCCI maintains legal privilege
- **Shadow AI risk mitigation:** Law firms deploying managed confidential AI infrastructure to replace risky public AI tools

AI Product Developers

AI product builders who need to offer their customers a verifiable privacy guarantee rooted in architecture, not terms of service. As AI agents consolidate sensitive data across work, health, finance, and travel, establishing the correct infrastructure foundation is critical before this data consolidation occurs.

---

12 REGULATORY LANDSCAPE

Compliance-Ready Architecture

Regulation	Key Requirement
EU AI Act	Effective August 2026: Mandates confidential computing for high-risk AI systems processing sensitive data
DORA	Effective January 2025: Requires EU financial institutions to implement cryptographic controls for critical AI systems
US State AI Laws	NY, CA, TX: Emerging requirements for algorithmic transparency and data protection; confidential computing satisfies attestation requirements
HIPAA / EU MDR	Healthcare: Mandatory encryption during processing; PCCI provides cryptographic proof of compliance
NYDFS Cybersecurity	Requires MFA, encryption, and third-party audits; attestation-based architecture fulfills audit requirements
GDPR	Right to erasure, data minimization: Confidential computing with client-controlled encryption keys ensures compliance

nFADP (Switzerland)	Swiss Federal Act on Data Protection: PCCI's Swiss HQ and E2EE architecture satisfy data processing requirements for Swiss organizations

13 THE STACK

Prem Private AI Stack

Prem Compute

Infrastructure layer: Confidential GPU compute (H100/B800 enclaves), hardware attestation, key management.

Prem API

OpenAI-compatible inference API: Encrypted request/response using XChaCha20-Poly1305, client-controlled keys, deterministic attestation reporting, and audit logging. The SDK manages encryption seamlessly—your code uses standard API calls.

Prem App

End-to-end application framework: No-code/low-code tools for data scientists and compliance teams to deploy models with zero-knowledge infrastructure access.

14 KNOWN LIMITATIONS

Transparency on Trade-offs

We believe transparency about limitations strengthens trust. There are three areas we want to be explicit about:

- **Side-channel attacks on TEE hardware:** While side-channel attacks could theoretically be possible, PCCI selects deployment locations that meet strict security criteria and may apply counter mechanisms to detect invalid states. These vulnerabilities exist across the industry. Manufacturers issue patches, and attestation reports include firmware versions so clients can verify patch status. Our mitigation includes prompt firmware updates, a post-quantum encryption layer that protects data in transit regardless of TEE state, and continuous monitoring of CVE databases.
- **Compromised chip manufacturers:** If AMD, Intel, or NVIDIA were to lose control of their root signing keys, attestation guarantees would weaken. This is a shared trust anchor for the entire confidential computing ecosystem—not something any single vendor can mitigate independently.
- **Metadata exposure:** The proxy observes request timing, payload sizes, and API key identifiers. Content is always encrypted, but traffic analysis on metadata remains theoretically possible.

TEEs remain the most practical architecture for encrypted AI inference today. Alternative methods such as fully homomorphic encryption are currently impractical due to performance constraints (approximately 100x slower for inference), and running models locally often requires specialized hardware that most organizations lack.

15 COMPLIANCE & DEPLOYMENT

Enterprise Readiness

Certifications & Compliance

- SOC2 Type 2
- ISO/IEC 27001 alignment (certification in progress)
- GDPR Data Processing Agreement
- HIPAA Business Associate Agreement
- EU AI Act readiness (effective August 2026)

Deployment Modes

- **Prem-Managed Cloud:** Full-service infrastructure on Prem's hardware
- **On-Premise:** Customer-deployed Prem PCCI stack on their own servers
- **Hybrid:** Mix of cloud and on-premises deployment

16 FAQ

Frequently Asked Questions

Q: What is PCCI?

A: PCCI stands for Prem Confidential Compute Infrastructure. It is a secure AI inference platform that combines Trusted Execution Environments (TEEs), end-to-end encryption using XChaCha20-Poly1305, and cryptographic attestation to ensure that sensitive data remains protected during AI processing. Unlike traditional API providers, PCCI ensures that no one—not even Prem—can access your data in plaintext outside of the secure enclave.

Q: Who is PCCI designed for?

A: PCCI is built for enterprises operating in regulated industries—such as finance, healthcare, legal, and government—where data confidentiality is not optional. It is ideal for organizations that want to use frontier AI models without exposing sensitive data to the model provider, cloud infrastructure operator, or any third party.

Q: How does end-to-end encryption work with AI inference?

A: When you send a request through PCCI, the client SDK encrypts your prompt using XChaCha20-Poly1305 with a session-specific key derived during attestation. The encrypted payload travels through Prem's proxy (which never sees plaintext) and is only decrypted inside a Confidential Virtual Machine (CVM) running on TEE-enabled hardware. After the model processes the request, the response is re-encrypted inside the CVM before being sent back to the client. At no point does unencrypted data leave the hardware enclave.

Q: Can the Prem team access my data?

A: No. PCCI is architected so that Prem employees, infrastructure operators, and cloud providers cannot access your data. Encryption keys are controlled exclusively by the client. Even if someone gained physical access to the server, the TEE hardware prevents reading memory contents from outside the enclave. This is verified through cryptographic attestation before any data is transmitted.

Q: Is my data used to train or improve models?

A: No. PCCI does not use customer data for training, fine-tuning, or any purpose beyond fulfilling your inference request. The architecture enforces this: data is encrypted with client-held keys and only decrypted inside the TEE for processing. Prem has no mechanism to extract or retain your data.

Q: What is attestation, and why does it matter?

A: Attestation is a cryptographic verification process that proves the hardware and software running inside a TEE have not been tampered with. Before sending any data, the PCCI client SDK requests an attestation report from the enclave, verifies it against known-good measurements, and only then establishes an encrypted session. This means you don't have to trust Prem's claims—you can verify them independently. PCCI also supports browser-based GPU attestation using Rust compiled to WebAssembly, a first in the industry.

Q: Is this truly 'end-to-end encrypted' if the TEE sees plaintext?

A: Yes. The term "end-to-end encryption" in the context of PCCI means that data is encrypted from the client to the enclave and from the enclave back to the client. The only place plaintext exists is inside the TEE—a hardware-isolated environment that is inaccessible to the operating system, hypervisor, or any external process. This is analogous to how end-to-end encrypted messaging apps decrypt messages on your device: the device (or in this case, the enclave) is the trusted endpoint.

Q: Does encryption add latency?

A: The overhead is negligible. CVM-level encryption adds approximately 5% overhead to compute operations. The XChaCha20-Poly1305 cipher operates at over 100,000 tokens per second, while LLM inference typically generates 100–200 tokens per second. Encryption and decryption are therefore not the bottleneck. In practice, PCCI response times are comparable to standard (unencrypted) API calls.

Q: Which compliance frameworks does PCCI address?

A: PCCI is designed to satisfy requirements across multiple regulatory frameworks, including GDPR, HIPAA, the EU AI Act (effective August 2026), DORA, NYDFS Cybersecurity Regulation, nFADP (Swiss Federal Act on Data Protection), and US state-level AI laws. Its architecture provides cryptographic proof of data protection, which simplifies compliance audits and regulatory reporting.

Q: What are the known limitations of PCCI?

A: We are transparent about three areas: (1) Side-channel attacks on TEE hardware—these are industry-wide vulnerabilities mitigated through firmware updates and continuous CVE monitoring. (2) Compromised chip manufacturers—if AMD or NVIDIA lost control of root signing keys, attestation guarantees would weaken. This is a shared trust anchor across the confidential computing ecosystem. (3) Metadata exposure—the proxy can observe request timing and payload sizes, though content remains encrypted. Despite these trade-offs, TEEs represent the most practical approach to encrypted AI inference available today.

Get in Touch

Email: hello@premai.io | **Website:** www.premai.io

Prem SA · Crocicchio Cortogna 6 · 6900 Lugano · Switzerland

901 N Market Street, Suite 100 · Wilmington, Delaware 19801 · USA

Via Giuseppe Verdi 6 · 70017 Putignano, Bari · Italia

