PREM

Prem Enclave

Confidential Compute Infrastructure for Sovereign AI Inference

The Enclave API — End-to-End Encrypted, OpenAI-Compatible AI Inference

Verifiable Private AI

Technical White Paper

Confidentiality. Integrity. Transparency.

July 2026

premai.io

Table of Contents

Revision History

Version	Date	Author	Description
1.0	March 2026	Prem AI	Initial release (PCCI white paper).
2.0	July 2026	Prem AI	Aligned to Enclave API. Updated product naming, architecture (two-server model), capabilities (multi-GPU attestation, multimodal, SDKs), performance metrics, and deployment model.

Executive Summary

Prem Enclave 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. The Enclave API provides an OpenAI-compatible interface where nobody, including Prem, can read your data.

Prem Enclave is built on three values of security:

- **Confidentiality:** Execution state and memory pages never leave Prem Enclave unencrypted. Hardware-level memory encryption covering the full stack guarantees zero visibility to the host OS, hypervisor, datacenter operators, and employees.
- **Integrity:** Every deployed Prem 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.

The Enclave API offers a model-agnostic inference platform designed for interoperability across the AI ecosystem. Data is end-to-end encrypted and processed within hardware-sealed TEEs. Prem Enclave 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 Prem Enclave as the foundation for verifiable, sovereign private AI in regulated verticals, including healthcare, defense, financial services, and law.

“Standard AI APIs require trust. Prem Enclave 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 Enclave 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.
- **CLOUD Act protection:** Even if a cloud provider were compelled by a foreign government subpoena, they can only hand over encrypted memory dumps and ciphertext. Hot data never leaves the Prem Enclave.

04 PROTOCOL

How Prem Enclave Works

Prem Enclave operates through four foundational steps, using a two-server model for defense in depth. Every service that handles plaintext material is confined inside the Prem Enclave.

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. File keys are wrapped with the customer-held master key (KEK) using AES-KWP. This means your data is protected against both classical and quantum attacks from the very first handshake. The Prem API Gateway processes only ciphertext and is solely responsible for authentication, rate limiting, usage tracking, billing, and routing. It holds no encryption keys and cannot read the data. 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. When a backend uses multiple GPUs for tensor parallelism on large models, each GPU is attested independently, producing its own per-GPU token. The model router creates a session binding between the attested GPU(s) and the inference request, so the customer has cryptographic proof that the exact GPUs they verified are the ones that processed their data.

3 Secure execution

The encrypted payload is routed by the Prem API Gateway to the Prem Enclave—a Confidential Virtual Machine (CVM) running inside a secure enclave. Inside the CVM, data is decrypted, processed by the model, and re-encrypted. The model router directs requests to the right self-hosted model. For stateless inference, memory is immediately wiped upon completion. For complex enterprise workflows requiring multi-turn context or Retrieval-Augmented Generation (RAG), Prem Enclave 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. The libraries ship as open-source Rust crates and as the @premai/reticle NPM package, published for independent audit. Prem is the first to offer NVIDIA GPU attestation directly from a web browser.

05 ARCHITECTURE

The Confidential Stack

Prem Enclave uses a two-server model for defense in depth. The Prem API Gateway acts as a blind proxy that receives encrypted payloads, handles authentication and billing, and routes requests—but holds no encryption keys. The Prem Enclave is the sealed processing environment where decryption, inference, and re-encryption occur, all inside a Confidential Virtual Machine (CVM).

Service	What It Does	In CVM?
Gateway	Authentication, routing, billing, rate limiting	No
Model router	Directs requests to the right self-hosted model	Yes
LLM / STT / VLM — Inference	Processes prompts/inputs through the inference stack and generates the response. GPU accelerated.	Yes

At a deeper level, Prem Enclave is organized as a 4-layer architecture:

Layer	Component
Layer 04	Enclave API & SDKs — OpenAI-compatible interface, TypeScript SDK, local proxy, AI agent integration, and developer tools
Layer 03	Prem Enclave Runtime — Enclave-native execution engines optimized for confidential inference, including multi-token prediction, CUDA graph kernel generation, speculative decoding, and hyperparameter optimization
Layer 02	Prem API Gateway — Minimal-attack-surface blind proxy handling authentication, rate limiting, billing, and routing. Processes only ciphertext; holds no encryption keys
Layer 01	Bare Metal Infrastructure — Purpose-built hardware utilizing TEE-ready silicon (AMD SEV-SNP, Intel TDX, NVIDIA Hopper/Blackwell Confidential Computing mode). Physically hosted in compliant datacenters

There is no gap in the chain between decryption and re-encryption—every service in the data path that handles plaintext is hardware-sealed and attested. The customer holds the master encryption key (KEK). Prem never receives it. If the entire platform were compromised, encrypted data remains unreadable.

06 SECURITY MODEL

Three Values of Enclave Security

Confidentiality

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

- CPU: Intel TDX and AMD SEV-SNP
- GPU: NVIDIA Hopper and Blackwell Confidential Computing

Integrity

Every deployed Prem 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
- Multi-GPU attestation with per-GPU tokens and session binding

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
- Open-source Rust attestation stack, compiled to WebAssembly, shipped as @premai/reticle NPM package

Defense in Depth — Seven Independent Layers

Even if one layer is bypassed, the others hold:

- TLS 1.3 — transport encryption
- End-to-end encryption — XChaCha20-Poly1305 (AEAD), including streaming chunks
- Quantum-resistant key exchange — XWing hybrid KEM (ML-KEM768 + X25519)
- CPU isolation — AMD SEV-SNP / Intel TDX
- GPU isolation — NVIDIA Confidential Computing (Hopper/Blackwell)
- Hardware-signed attestation — per-GPU tokens, browser-verifiable
- Key sovereignty — customer generates and holds the master key (KEK); Prem never receives it

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 Hopper and Blackwell Confidential Computing mode)
Attestation	Cryptographic generation and verification of enclave quotes — Hardware Root of Trust rooted in chip manufacturer (AMD, Intel, NVIDIA). Multi-GPU attestation with per-GPU tokens.
Key Management	Secure, enclave-bound cryptographic key generation and lifecycle management. XChaCha20-Poly1305 for data; AES-KWP for key wrapping; Post-Quantum ML-KEM hybrid (XWing) for key exchange. Customer holds the KEK.
Transport	TLS 1.3 for network-level protection

Encryption Architecture

All encryption happens client-side. The customer's plaintext data never leaves their device unencrypted.

Layer	Technology	What It Protects
Transport	TLS 1.3	Network eavesdropping
End-to-end	XChaCha20-Poly1305 (AEAD)	Data from device to enclave, including streaming chunks
Key exchange	XWing hybrid KEM (ML-KEM768 + X25519)	Post-quantum secure session establishment
Key wrapping	AES-KWP	File keys wrapped with customer master key (KEK)
CPU isolation	AMD SEV-SNP / Intel TDX	Enclave memory sealed at hardware level
GPU isolation	NVIDIA Confidential Computing (Hopper/Blackwell)	GPU memory encrypted during inference

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. Prem Enclave 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

Prem Enclave 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 Enclave API 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 Prem Enclave 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 Enclave

Dimension	Standard AI API	Prem Enclave
Who can see your data?	The provider, their staff, and potentially their cloud host	Only you
If the host is compromised?	Your data is exposed	The attacker sees only encrypted bytes
How do you know it's private?	You trust the provider's privacy policy, DPAs, and legal boundaries	You verify it with hardware-signed cryptographic proof, configuration policies, and vendor certificates
What about future threats?	Vulnerable if quantum computers break today's encryption	Protected now by quantum-resistant algorithms
What about the cloud provider?	They can inspect server memory	Hardware isolation prevents this, whoever owns the machine
API compatibility	OpenAI-compatible	OpenAI-compatible interface served through our API SDKs

Technical Comparison

Dimension	Traditional Cloud AI	Prem Enclave
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
Multi-GPU Attestation	Not available	Per-GPU tokens with session binding

10 PERFORMANCE

Encryption Without Compromise

A common concern with confidential computing is latency. Prem Enclave 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 overhead: The Prem Enclave adds approximately 10–15% compute overhead on Hopper GPUs, less with newer GPU architectures such as Blackwell. We have heavily optimized our stack—including multi-token prediction, CUDA graph kernel generation, and speculative decoding (both drafting and verification happen inside the enclave)—to ensure Time-to-First-Token (TTFT) remains minimally impacted. Total generation throughput is engineered to remain highly competitive with unencrypted baselines.
- 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: The Enclave API 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.

- Italy's second-largest hospital: Patient-doctor transcription processed inside encrypted enclaves, with attestation proving no data leaked to third parties.
- 40-clinic AI assistant across North America: Medical records processed inside encrypted enclaves, 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.
- On-premise deployment: The on-prem enclave strengthens the national perimeter and lets remote operations run as though they were inside it—same security posture, same privacy guarantees, no loss of sovereign control at the edge.

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 Enclave 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: Top-tier US law firms process privileged documents through Fluso, which runs on Prem Enclave. Client data never leaves the encrypted boundary.
- 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; Prem Enclave 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: Prem Enclave's Swiss HQ and E2EE architecture satisfy data processing requirements for Swiss organizations

13 THE STACK

Prem Enclave Platform

Prem Compute

Infrastructure layer: Confidential GPU compute (NVIDIA Hopper/Blackwell enclaves), CPU attestation (AMD SEV-SNP, Intel TDX), GPU attestation with multi-GPU support, and key management.

Enclave API

OpenAI-compatible inference API: Encrypted request/response using XChaCha20-Poly1305, client-controlled keys (KEK), deterministic attestation reporting, and audit logging. The SDK manages encryption seamlessly—your code uses standard API calls. Supports chat completions (streaming, tool calling, multi-step reasoning, vision payloads), audio transcription (Whisper and Deepgram models), and audio translation.

Available Models

All models are self-hosted on Prem's infrastructure with no requests are forwarded to third-party AI providers, enterprise customers can also bring their own open-weight models. All Prices are pay-as-you-go per usage, learn more at: https://docs.prem.io/developer-resources/billing/models-and-pricing.

Capability	Details
Chat completions	OpenAI-compatible. Streaming, tool calling, multi-step reasoning, vision payloads
Audio transcription	Whisper and Deepgram models, fully encrypted
Audio translation	Audio to English text
CPU attestation	AMD SEV-SNP and Intel TDX, independently verifiable
GPU attestation	NVIDIA Hopper and Blackwell, per-GPU tokens, multi-GPU support
TypeScript SDK	Automatic encryption, attestation, session management
Local proxy	OpenAI and Anthropic compatible, any language, zero code changes
Organizations and teams	Multi-org support with role-based permissions
Scoped API keys	Fine-grained permissions with IP restrictions
Rate limiting	Four-tier system with automatic graduation

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, Prem Enclave 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 Prem API Gateway observes request timing, payload sizes, and API key identifiers. Content is always encrypted, but traffic analysis on metadata remains theoretically possible. A compromised gateway is an availability risk, not a data exposure risk.

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

- SOC 2 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 Models

Tier	Infrastructure	Models	Rate Limits
Explorer	Shared multi-tenant	Curated selection	10 req/sec
Developer	Dedicated instances	Custom model choice	Scalable
Enterprise	Air-gapped or on-premise	Custom models, bring your own	Dedicated

Organizations can choose between a managed environment or an air-gapped on-premise license to suit their specific requirements. Enterprise deployments support bring-your-own open-weight models and dedicated rate limits.

Shared Responsibility

Prem secures the hardware isolation and attestation infrastructure, data in transit / in use / at rest, model execution, and physical and platform security.

The customer secures key management (generating, storing, and rotating the KEK), the endpoints where the SDK runs and data is decrypted, and the prompt and model-level defenses within their own AI workflows.

Key Sovereignty

The customer generates and holds the master key (KEK). Prem never receives it. Lost keys mean permanent data loss, because Prem cannot recover what it cannot decrypt. This is a feature, not a bug—it is the structural guarantee that no one, including Prem, can access your data without your keys.

16 DEVELOPER EXPERIENCE & SDKS

No Encryption Expertise Required

If a team has built anything with the OpenAI API, they already know how to use Prem Enclave. The Enclave API is fully OpenAI-compatible—just swap your client SDK and go live.

TypeScript SDK

Install from npm (@premai/api-sdk). Initialize with API key and master encryption key. The SDK handles all cryptography, attestation, and session management automatically. The code looks identical to an OpenAI integration.

Local Proxy

For Python, Go, Java, or any language with an OpenAI-compatible client. Run one command, point the existing base URL at localhost, encryption happens transparently. Zero code changes to application logic. Supports both OpenAI and Anthropic compatibility modes.

AI Agent Integration

A skill.md file ships with the docs.prem.io and the GitHub repository (github.com/prem-research/api-skills), so AI coding agents like Claude Code, Cursor, and Windsurf can use the Enclave API out of the box with one command. The @premai/api-sdk includes a specialized binary designed to plug Prem Enclave directly into existing development harnesses.

Quickstart Guide

Follow these steps to perform your first encrypted inference call:

- Explore the Documentation: Head over to docs.prem.io for a comprehensive walkthrough. This guide covers billing structure, supported LLMs, and instructions for integrating the Prem SDK into your current development workflows.
- Obtain Your Credentials: Create an account and generate your unique API key via the Prem Dashboard at dashboard.prem.io.
- Universal Compatibility: Whether you use Python, Go, or other languages, simply execute the bundled confidential-proxy. By redirecting your OpenAI client to localhost, you gain transparent encryption with absolutely no changes to your code.

17 FAQ

Frequently Asked Questions

Q: What is Prem Enclave?

A: Prem Enclave 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. The Enclave API provides an OpenAI-compatible interface where nobody, including Prem, can access your data in plaintext outside of the secure enclave. Unlike traditional API providers, Prem Enclave ensures that no one—not even Prem—can read your data.

Q: Who is Prem Enclave designed for?

A: Prem Enclave 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 the Enclave API, the client SDK encrypts your prompt using XChaCha20-Poly1305 with a session-specific key derived during attestation. The encrypted payload travels through the Prem API Gateway (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. Prem Enclave is architected so that Prem employees, infrastructure operators, and cloud providers cannot access your data. Encryption keys are controlled exclusively by the client. The customer generates and holds the master key (KEK); Prem never receives it. 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. Prem Enclave 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 Enclave API 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. Prem Enclave

also supports browser-based GPU attestation using Rust compiled to WebAssembly, a first in the industry. When a backend uses multiple GPUs, each GPU is attested independently with per-GPU tokens and session binding.

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 Prem Enclave 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. The Prem Enclave adds approximately 10–15% compute overhead on Hopper GPUs, less with newer GPU architectures such as Blackwell. 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. A brief one-time delay occurs during the initial session handshake (key exchange plus attestation); after that, streaming performance is virtually indistinguishable from an unencrypted API.

Q: Which compliance frameworks does Prem Enclave address?

A: Prem Enclave 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 Prem Enclave?

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, Intel, 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 Prem API Gateway 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.

Q: What models are available on the Enclave API?

A: The Enclave API offers state-of-the-art open-weight models including multimodal features and Deepgram for speech-to-text. All models are self-hosted on Prem's infrastructure—no requests are forwarded to third-party AI providers. Enterprise customers can also bring their own open-weight models.

Get in Touch

Email: jaipal@premai.io

Website: www.premai.io

Documentation: docs.prem.io

Dashboard: dashboard.prem.io