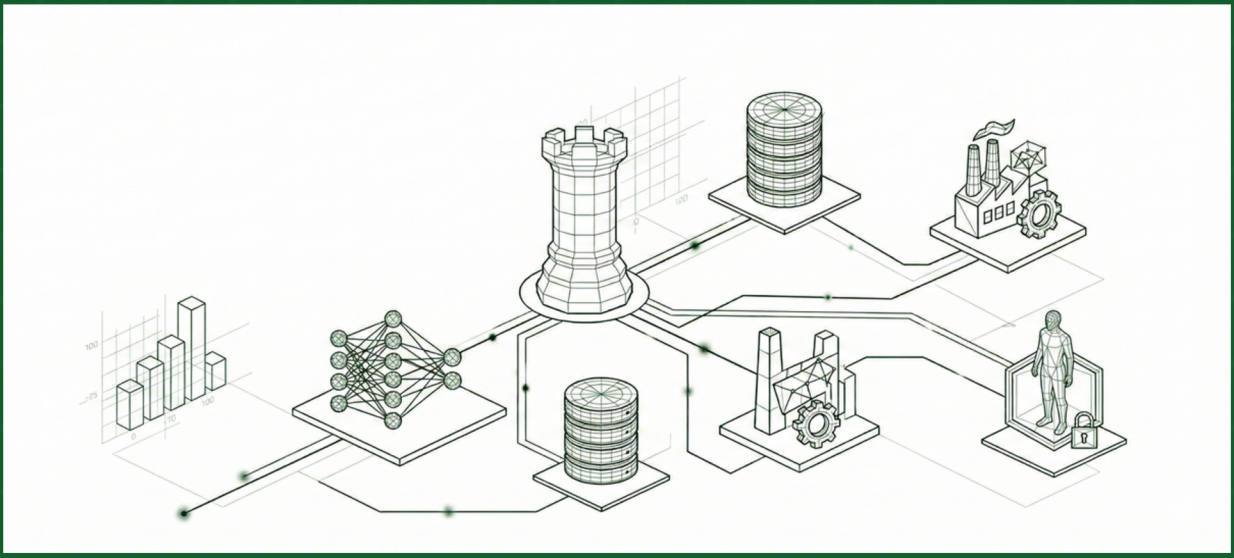
Prediction Guard

Whitepaper

Navigating the AI Security Noise

AI Gateways and the System-Level Advantage of Self-Hosting an AI Control Plane



Sharan Shirodkar
AI Engineer, Solutions Architect
February 2026

Executive Summary

Production Challenge

As enterprises transition from AI experimentation to production, they encounter a variety of external point solutions, like AI/LLM gateways or firewalls, that seem to address certain isolated concerns related to security and/or privacy. However, these external AI "wrappers" or "layers" leave companies with significant gaps as they work towards standards-aligned security, supply-chain integrity, and regulatory compliance.

The Systemic Solution

Considering the AI/LLM gateway by way of example, this whitepaper examines why an external gateway alone is insufficient to deal with the modern AI threat landscape. In particular, these gateways fall short in the current ecosystem, which is dominated by the rise of highly connected, autonomous agents.

2026 Strategic Reality: The Cost of Inaction

276+ Regulatory Burden
Distinct requirements per AI system in healthcare and finance

4.1K+ The Audit Tax
3,200-4,100+ person-hours for initial AI compliance audits

64% Shadow AI
Of workers bypass security and use personal logins and unauthorized tools

60% Sovereign Fragmentation
Of firms must split AI stacks, tripling integration and governance costs

Secure AI adoption goes beyond the convenience of a gateway. It requires a self-hosted (i.e., internal) control plane that acts as a single point of management for your entire AI ecosystem. This philosophy drives our development at Prediction Guard. We prioritize internal sovereignty and comprehensive governance, ensuring your AI assets are not just accessible, but actively monitored and fully under your control.

The Rise of the AI Gateway

Solving Part of the Problem

There has been a proliferation of various kinds of AI models (LLMs, LVMs, embeddings, etc.) from a variety of model builders (OpenAI, Anthropic, Google, Meta, DeepSeek, etc.). As the model landscape became more diverse, it became much more difficult for developers to switch between models/ providers, evaluate a variety of models for a task, and develop applications in a model agnostic manner.

To deal with this real inconvenience, AI gateways emerged as a middleware layer to manage the complexity. AI gateways provide a specialized, AI-specific API that unifies access to many different kinds of models from many different model builders.

Key Capabilities & Limitations of AI Gateways

Unified Access

A single API for all internal and external models, preventing vendor and model lock-in.

Credential Management

Centralizes API keys or other mechanisms of authentication, reducing the complexity of managing various credentials.

Consumption Control

Enforces rate limits and token budgets to prevent cost overruns.

Basic Guardrails

Provides basic, external guardrails such as PII redaction and toxicity filtering.

The Bottom Line on AI Gateways

Gateways are tools for access, not authority. While they offer "security by proxy" (i.e., a thin, external layer that facilitates developer convenience) they sit outside the organization's infrastructure and provide only symbolic protection. They may manage the traffic, but they lack the deep architectural oversight required to secure the supply chain or ensure systematic integrity. True enterprise-grade sovereignty requires moving beyond this "wrapper" mentality toward a self-hosted control plane that governs the AI assets themselves, not just the semantics of the API calls.

The Case For a Self Hosted AI Control Plane

The most significant risk in the modern AI ecosystem is the simultaneous explosion in both complexity of AI systems and agency/ access. As different departments adopt different AI models, tools, and services (LLMs, code assistants, LVMs, MCP Servers, agent builder, etc.), the "AI sprawl" makes it impossible to maintain a coherent security posture.

If a gateway is a door, an AI Control Plane is the building's entire security and management system. It is the centralized architectural layer that sits between your users, your data, and your AI Assets. Rather than acting as a third-party intermediary, the control plane functions as a sovereign part of your internal infrastructure.

Strategic Benefits of Owning Your AI Control Plane

Financial Sovereignty

Prevent a **300% increase** in costs driven by fragmented AI stacks and sovereign zone requirements.

Compliance ROI

Drive a **94% reduction** in manual remediation via real-time, automated "governor layers."

Workflow Velocity

Cut documentation time by **40%** by embedding AI directly into runtime revenue and coding workflows.

Deployment Speed

Slash **regulatory approval timelines** from months to weeks through automated validation.

Data Sources: *Hutzschenreuter & Lämmermann (2025), Acad. Mgmt. Persp.; IJSAT (2025), Int. J. Sci. Adv. Tech. 15(1); Navigating AI Regulatory Landscape (2025), Taylor & Francis; Tallberg et al. (2023), Global Governance of AI, arXiv; The Hidden Cost of AI Agent Sprawl (2026), BuildInDigital.*

While gateways focus on the transit of API requests, a control plane focuses on the **governance of AI systems**. It provides a unified interface to manage disparate AI assets, but more importantly, it enforces your organization's specific security, privacy, and compliance policies across every interaction (including those unrelated to API interactions like adding/ removing AI assets from a system, proactively analyzing AI assets for vulnerabilities, auditing system settings and rolling back key changes, etc.).

In an era where AI agents are increasingly autonomous and interconnected, simply "securing the pipe" is no longer enough. You must **own the system** that dictates how those agents behave, what AI assets can or should be composed together, and where your data/ IP flows.

Key Capabilities of a Sovereign AI Control Plane

Sovereign Infrastructure

Operates entirely within your own VPC or on-premises environment, ensuring that sensitive data and governance logic never leave your trust boundary.

Asset Governance

Provides a centralized registry for all AI assets, including models, MCP tools, and agents, allowing for systematic tracking and auditing of the entire lifecycle.

Advanced Risk Monitoring

Goes beyond basic filtering to provide proactive analysis of AI assets (e.g., model scans), integration with SIEM or other security infrastructure, and the application of custom policies.

Standardized Compliance

Enables the automated enforcement of industry best practices, such as NIST or OWASP frameworks, across every AI interaction in the organization.

Telemetry Ownership

Consolidates and stores all logs, audit trails, and performance metrics within your own security stack, ensuring data persistence and full visibility for CISO/CIO oversight.

Supply-Chain Integrity

Tracking detailed AI BOMs for the variety of AI assets in your AI systems and probes model behaviors before they are integrated into production workflows.

Owning your control plane (i.e., hosting it within your own virtual private cloud, VPC, or on-premises environment) marks the difference between **rented convenience** and **permanent sovereignty**. When the control plane is internal, your security posture is no longer dependent on what a vendor may or may not implement. Instead, it is integrated into your existing enterprise security stack. This ownership ensures that:

1. Data never leaves your trust boundary for the sake of "security processing."
2. Audit trails and logs remain immutable and stored on your own terms.
3. Policy enforcement is instantaneous, covering everything from prompt injection defense supply-chain verification.

Prediction Guard's AI Control Plane

Prediction Guard is a security-first AI control plane designed for organizations that prioritize sovereignty over simple connectivity. While our platform includes a consolidated gateway to unify access to your AI assets, this is merely the entry point to a much broader security posture. Our goal is not to control your AI system but rather to provide the infrastructure so that **you can own it**.

This commitment to ownership is reflected in, by way of example, how we handle governance and telemetry. We have integrated standards-aligned frameworks from organizations like NIST and OWASP directly into the platform, giving you the flexibility to deploy these proven templates or implement your own unique policies. Similarly, while we facilitate seamless integration with security tools such as Datadog or Splunk, we ensure that you maintain absolute ownership of the resulting telemetry data. These capabilities are just part of a wider suite of tools designed to keep you in control.

> **For the CISO and CIO,** an AI gateway is a useful tool but it is not the destination. True enterprise readiness requires moving past fragmented security 'band-aids' toward a comprehensive and self-hosted solution that puts your organization back in total command of its AI future.

The 2026 Strategic Reality

For Agentic AI, where software now makes decisions rather than just executing them, the value of a runtime control plane is the difference between a productive asset and an uninsurable liability.

For more information, reach out to the Prediction Guard team.