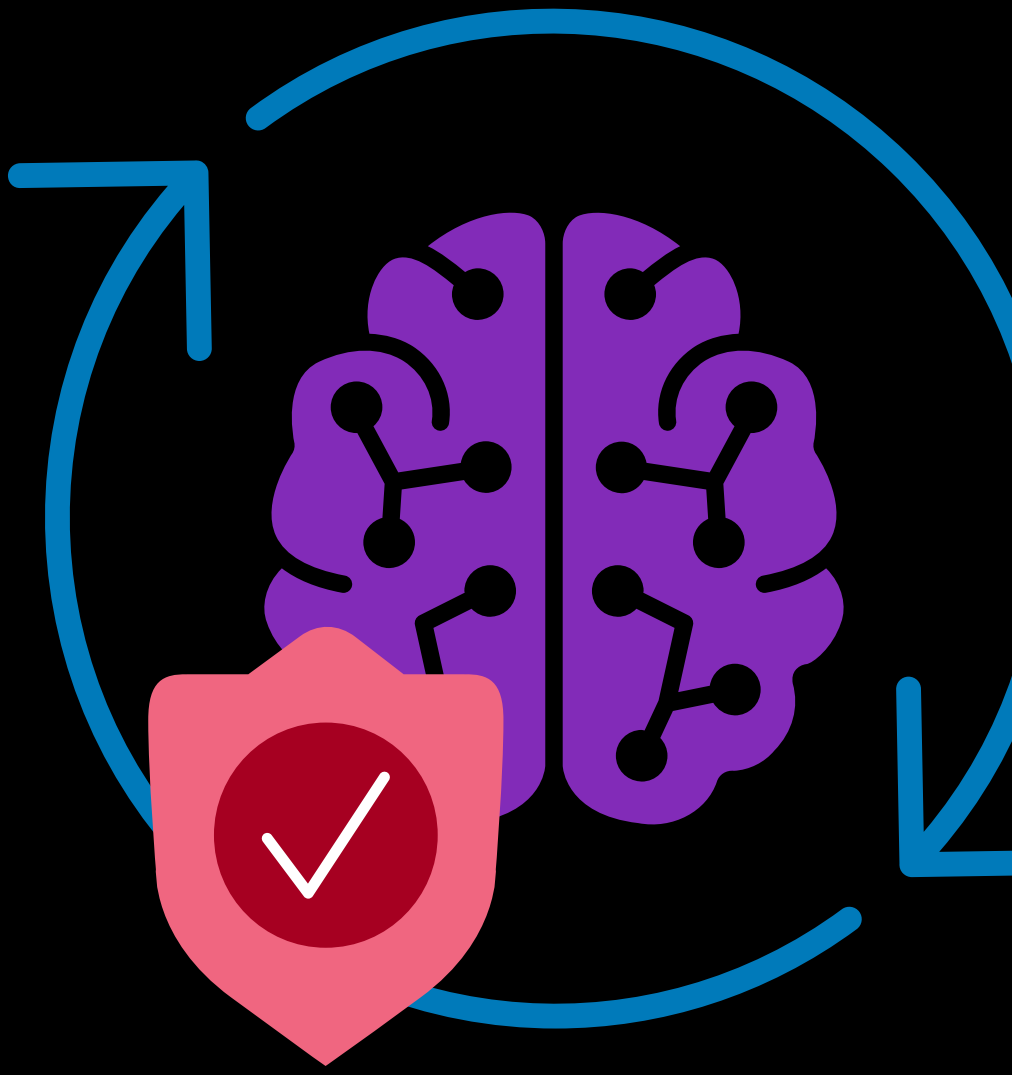
Google Cloud

SOLUTION OVERVIEW

Secure your AI applications with end-to-end API protection

Discover, protect, and optimize APIs powering your AI applications at scale with F5 and Google Cloud.



Key benefits

Reduce risk
Keep APIs secure throughout their lifecycle to protect AI models and sensitive data.

Secure every API
Discover all API endpoints in your environment and locate sensitive data for comprehensive defenses.

Optimize performance and cost
Apply rate limiting and resource controls to prevent overuse and reduce infrastructure spend.

Enforce policies consistently
Enforce uniform API security controls across cloud, on-premises, and hybrid AI deployments.

Gain unified observability
Streamline monitoring with centralized metrics and logging integrated across F5 and Google Cloud tools.

Every stage of the AI app lifecycle relies on APIs

From data ingestion and model training to inference and real-time responses, AI applications depend on APIs to function effectively. As AI becomes embedded in enterprise IT ecosystems, each step in the pipeline introduces new API connections that link models to data and end-user services.

Tasks like fine-tuning or retrieval-augmented generation (RAG) add more endpoints, often without visibility or oversight from security teams. This rapid growth in API traffic is accelerating. Two billion APIs are expected by 2030.[1] And Gartner predicts over 30% of new API demand will stem from AI and LLM use cases by 2026.[2] With this expansion comes a critical need: organizations must secure every API across the AI lifecycle to protect sensitive data, proprietary models, and the integrity of their AI investments.

An evolving threat landscape exposes AI APIs to new risks

Securing AI applications requires more than traditional defenses. The complexity of AI infrastructure, including training, inference, and data exchange, has created a growing web of API interfaces that attackers actively probe for weaknesses. These APIs not only expand the attack surface but often operate in ways that security teams are unprepared to manage. From misused inputs to invisible endpoints, AI-specific API risks are becoming increasingly sophisticated and impactful.

The Open Worldwide Application Security Project (OWASP) has identified the top risks to AI apps in the OWASP Top 10 for LLM Applications, many of which require robust API security to mitigate, including:

- **Prompt injection.** Attackers manipulate API inputs to make AI models disregard safety controls or expose sensitive information.
- **Unbounded consumption.** When APIs that lack proper rate limiting allow attackers to overwhelm AI models with requests.
- **System prompt leakage.** AI systems accidentally reveal their internal instructions, configuration details, or operational parameters through API responses.

Shadow APIs multiply your risk

When APIs exist outside the knowledge and control of security teams, these unprotected shadow APIs are an open invitation for attackers. You can't protect what you can't see.

Key features

Continuous discovery
Maintain real-time visibility into all API endpoints that connect to AI services.

Authentication and access control
Ensure every API endpoint connecting to AI models or data has proper authentication controls and clearly defined access policies.

Rate limiting
Set and enforce specific thresholds for API calls to prevent both accidental and malicious resource consumption, as well as denial-of-service (DoS) attacks.

Anomaly detection
Monitor API traffic patterns continuously to identify potential model theft attempts, system prompt leakage, and other suspicious behaviors that could indicate an ongoing attack or unauthorized access to your AI systems.

Unified observability
Provide comprehensive visibility into API performance, usage patterns, and security events across your entire AI infrastructure.

Comprehensive protection across the AI application lifecycle

F5 and Google Cloud have partnered to meet the specialized demands of AI security with an integrated approach. By combining Google Cloud's AI-optimized infrastructure and native security capabilities with the F5® Application Delivery and Security Platform (ADSP), teams get end-to-end API security across every stage of AI development and deployment.

Discover every API in your AI ecosystem

Effective API security starts with visibility. Together, F5 and Google Cloud help you identify every API—authorized or not—across hybrid multicloud environments. Application traffic is scanned to uncover live endpoints, while source code inspection reveals risks early in the development process. The solution also helps locate sensitive data and enforce policies that prevent exposure through APIs.

Protect against AI-specific and traditional threats

F5 uses machine learning to detect and block threats in real time. Automated rate limiting helps contain excessive requests and defends against Layer 7 DoS attacks. Strong authentication and access control safeguard model and data access, while OpenAPI Specification (OAS) enforcement ensures API requests are valid and safe. These protections are consistently applied across all environments, including Google Cloud.

Monitor and visualize your security posture

Security insights are centralized across all environments. F5 aggregates global API metrics into an interface that integrates with Google Cloud's observability tools for real-time monitoring and alerting. Logs can also be routed to BigQuery for long-term storage and advanced analysis, with visualization available in Looker Studio.

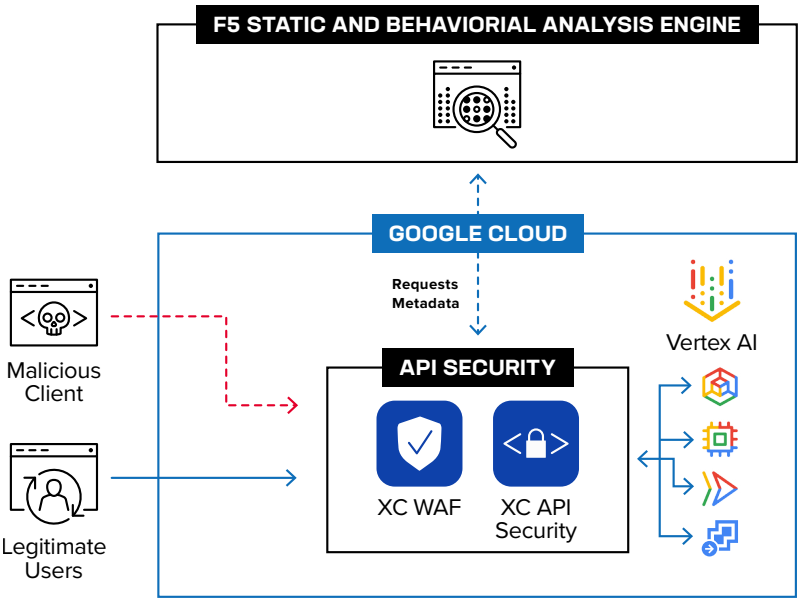


Figure 1: Advanced API security for AI apps and services on Google Cloud with API discovery and protection from F5.

Elevate your API security with F5 and Google Cloud

F5® Distributed Cloud API Security identifies APIs by analyzing live traffic to uncover shadow or undocumented APIs. Code repository scanning integrates with platforms like GitHub to identify APIs early in the development lifecycle, while client-side web crawling navigates dynamic AI frontends. Behavioral monitoring, risk scoring, and enforcement improve protection.

Google Cloud provides essential API security capabilities that integrate seamlessly with AI workloads. Native tooling in Google Cloud helps you build, manage, and secure APIs using AI and machine learning. It can detect undocumented and unmanaged APIs in your Google Cloud environment and identify critical API abuses.

Together, these solutions provide broad security coverage, leveraging Google Cloud's native tooling to manage APIs within Google Cloud infrastructure and F5 solutions to identify APIs across on-premises, multicloud, and third-party environments. F5 security events integrate with Google Security Command Center and Google Security Operations for holistic threat detection and response, while native OpenTelemetry support enables seamless log flow to Google Cloud Logging, Cloud Monitoring, and BigQuery for unified observability.

As AI applications continue to reshape the enterprise landscape, APIs have become both the connective tissue and the most critical attack surface. F5 and Google Cloud deliver a unified, end-to-end solution that discovers, protects, and monitors every API across any environment and every stage of the AI lifecycle.

Learn more about F5 and Google Cloud at f5.com/gcp.

1 F5, API Security Evaluation Guide, Dec 2023

2 Gartner Press Release, Gartner Predicts More Than 30% of the Increase in Demand for APIs will Come From AI and Tools Using Large Language Models by 2026, Mar 2024