Solo.io's guide to managing GenAI complexity

Author: Jim Barton

Introduction

The rise of Generative AI (GenAI) and Large Language Models (LLMs) is revolutionizing how applications are developed, deployed, and consumed. Once limited to research labs and tech giants, these technologies have rapidly become essential tools for businesses across industries. By enabling applications to automatically generate human-like responses across text, code, and images, to even making complex decisions through analytics, GenAI unlocks unprecedented possibilities for enterprises. However, it also introduces new layers of challenges, unpredictability, and risk—what we collectively refer to as “GenAI Complexity.”

In this guide, we explore the challenges and opportunities of building GenAI-powered applications, the crucial role AI Gateways play in managing these risks, and why modern solutions like Gloo AI Gateway are essential for bringing order to the chaos.

The Rise of AI/LLMs: A Double-Edged Sword

The rapid adoption of GenAI and integration of LLMs into application development is driving a transformative shift across industries. From chatbots engaging in complex conversations to algorithms generating code, content, and art, GenAI is revolutionizing how humans interact with technology. This evolution brings exciting possibilities, but it also introduces new risks and challenges for technology operators, developers, and decision-makers which we will explore further below. To maintain an application's performance, reliability, and seamless user experience, organizations must proactively navigate the unique complexities presented by LLMs.

Understanding the Complexity Unleashed by GenAI in Application Development

While GenAI's capabilities are impressive, they also inject significant unpredictability into application development processes. Traditional software development follows a deterministic approach where a specific input yields a known output. In contrast, GenAI operates probabilistically, where the same input can yield different outputs depending on various factors, including the model's state, training data, and even the phrasing of the prompt.

This unpredictability leads to various issues in application development:

- **Uncontrolled Outputs:** GenAI models can generate unexpected, inappropriate, or even harmful outputs, necessitating robust filtering and moderation mechanisms.
- **Inconsistent User Experiences:** Variability in AI-generated content can lead to inconsistent user experiences, potentially damaging brand reputation and eroding user trust.
- **Scalability Challenges:** As demand for AI-driven features grows, managing the scalability of these services becomes increasingly complex, especially in cloud-native environments.

The Impact on Developers, Decision-Makers, and Organizations

For **developers,** the introduction of GenAI can both enhance and complicate the development process. When properly implemented, LLMs can boost productivity and accelerate code velocity, helping enterprises shorten time-to-market and improve service uptime. However, developers must now navigate not only the integration of AI-driven features but also the risks associated with unpredictable outputs. This often demands additional layers of logic, validation, and monitoring, which increases codebase complexity and extends the development lifecycle.

Decision-makers face a different set of challenges. While the innovation potential is vast—with benefits like improved application performance and resource efficiency—they must carefully weigh these advantages against the need for governance, control, and compliance. The rapid adoption of GenAI can outpace an organization's ability to manage associated risks, potentially leading to security breaches, regulatory non-compliance, and operational failures.

For **organizations,** the stakes are high. Poorly managed GenAI implementations can result in significant financial losses and reputational damage due to flawed outputs. However, those who successfully harness the power of AI while mitigating risks can gain a substantial competitive edge by proactively navigating the complexities of LLMs.

The GenAI Opportunity

Despite the challenges, GenAI presents vast opportunities with an exponentially growing spend within IT. A recent PWC survey[1] found that nearly **100% of business leaders** say their company is **prioritizing at least one AI-related initiative** in the near term. By 2027, Gartner[2] believes the market for artificial intelligence services, including AI and GenAI, **will reach $822 billion.**

Organizations can leverage AI to automate mundane tasks, enhance customer interactions, personalize experiences, and even create entirely new products and services. AI's ability to generate content, code, and decisions at scale enables businesses to operate more efficiently and innovate faster.

For instance, AI-driven chatbots can manage a high volume of customer inquiries, providing instant responses and freeing human agents to focus on more complex issues. In software development, AI can assist in writing code, identifying bugs, and optimizing performance, reducing the time-to-market for new features.

Moreover, AI can help organizations make data-driven decisions by analyzing large datasets and identifying patterns that would be impossible for humans to detect. This can lead to better business outcomes, such as improved product recommendations, optimized pricing strategies, and more effective marketing campaigns.

However, to fully capitalize on these opportunities, organizations must address the technical challenges and risks associated with GenAI which we explore further in this guide.

[1] PricewaterhouseCoopers. (n.d.). *Managing the risks of generative AI.* PwC. https://www.pwc.com/us/en/tech-effect/ai-analytics/managing-generative-ai-risks.html

[2] Gartner, Crossing the Chasm: Tech Provider Plans for Generative AI in 2024., McDonald, M., Lovelock, J.-D., & Andrews, W. (2024, February 23)

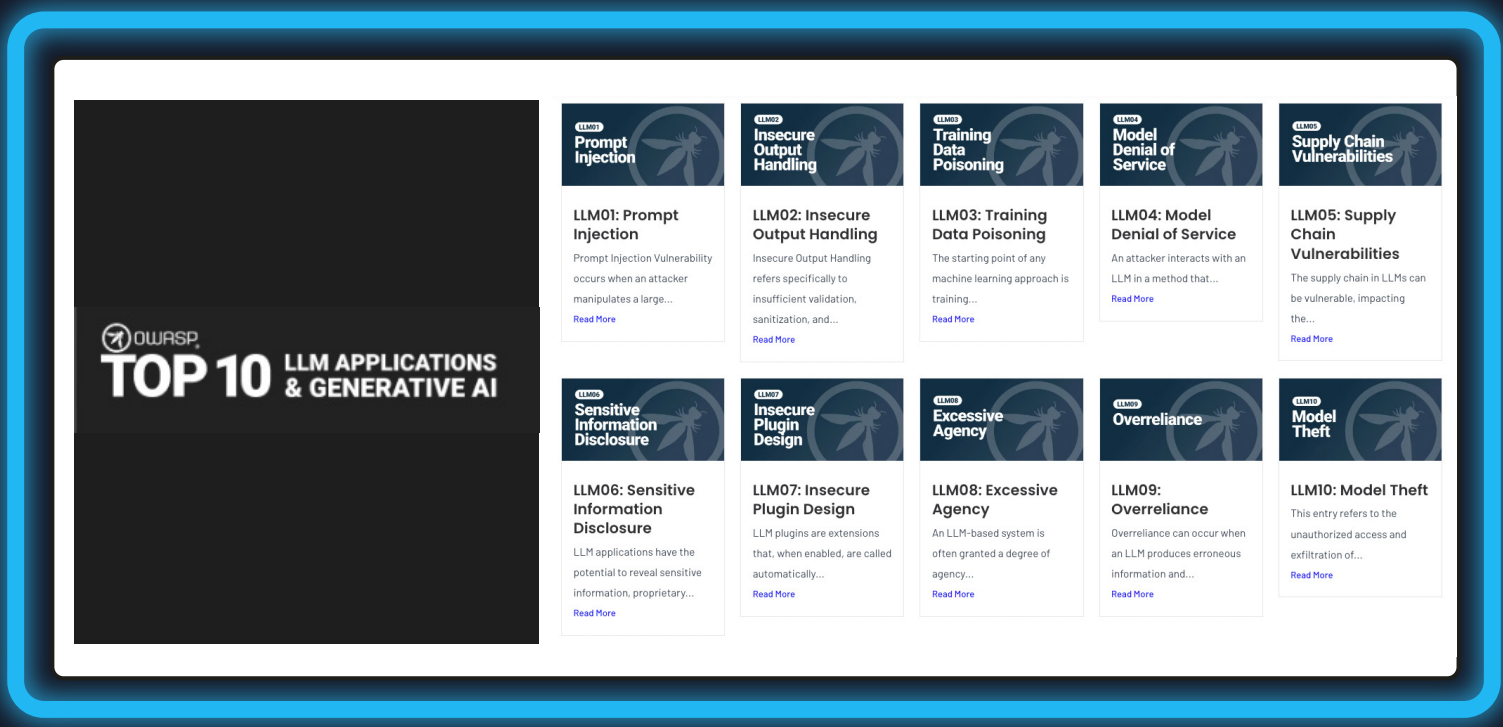
Technical Challenges and Risks of GenAI

Conducting a PoC with GenAI is easy; getting it to production without the right governance is impossible.

Let's review the most important challenges where a scalable, policy-driven approach is required.

Security and Compliance

Security and compliance are critical concerns when integrating GenAI into applications. In a survey conducted by Salesforce.com in 2023[3], 71% of executives believe GenAI introduces new security risks to their data systems, highlighting security as a key challenge in GenAI adoption. The Open Worldwide Application Security Project (OWASP) Foundation has been instrumental in helping organizations understand application security risks, publishing an insightful Top 10 list of risks for GenAI[4] applications. Developing a comprehensive, enterprise-wide strategy is crucial, as piecemeal, single-application solutions are neither scalable nor cost-effective for innovation-focused enterprises.



In many early AI implementations, individual applications managed their own LLM credentials, creating unnecessary operational overhead for development teams. This fragmented approach also complicates cost management, as enterprises struggle to gain a unified view of AI-related expenses as integration with AI models is scaled.

[3] Salesforce.com. (2023, March 6). IT Leaders Call Generative AI a 'Game Changer' but Seek Progress on Ethics and Trust. https://www.salesforce.com/news/stories/generative-ai-research

[4] OWASP Foundation. (n.d.). OWASP top 10 for large language models (LLMs). OWASP. https://genai.owasp.org/llm-top-10/

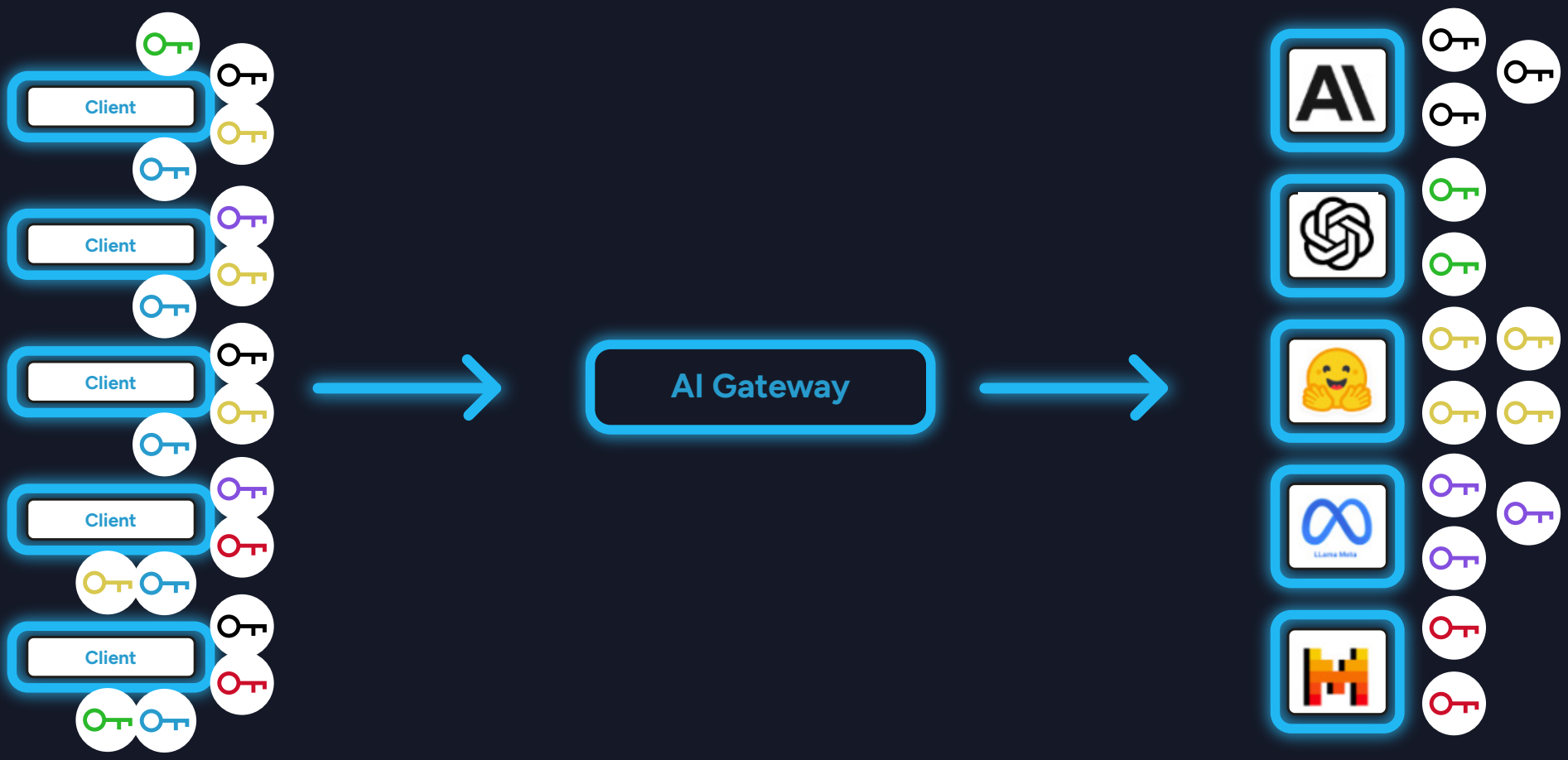
Credential Management

AI models require access to vast amounts of data, some of which may be sensitive or confidential. Ensuring that this data is handled securely and complies with regulations such as GDPR, CCPA, or HIPAA is critical. Additionally, the outputs generated by AI models must also comply with legal and ethical standards, which can be challenging given the unpredictable nature of GenAI.

Without proper controls, AI models could inadvertently generate content that violates regulations, leading to fines, legal action, and damage to the organization's reputation. For example, a GenAI-powered chatbot could generate a response that discloses proprietary or personal information, violating privacy regulations.

Challenge

- API keys and client identity managed individually for each LLM provider and passed differently
- Key management (tracking, revocation, refresh) distributed across all providers and clients
- Identity and access management integration with each LLM provider
- DIY problem for local LLMs



Prompt Management

The way AI models respond to inputs—prompts—can vary widely. This variability poses a challenge for developers, who must ensure that AI-generated content aligns with the organization's standards and goals. Effective prompt management involves designing prompts and applying them automatically to guide the LLM model in producing the desired output while minimizing the risks of exfiltrating proprietary data or generating inappropriate content.

Prompt management is especially important in applications where AI interacts directly with end users, such as chatbots or virtual assistants. A poorly designed prompt can lead to off-topic, confusing, or even offensive responses, potentially harming the user experience and damaging the brand's reputation.

Consumption Control

AI models, particularly those running in cloud-native environments, can quickly become resource-intensive. Without proper management, the cost of running these models can quickly spiral out of control. This is particularly true for LLMs, which require substantial computational power and storage.

Organizations must implement mechanisms to control who can access AI services, how much they can consume, and under what conditions. This includes setting usage limits, monitoring consumption patterns, leveraging patterns like semantic caching to reduce response times and cost, and optimizing the deployment of AI models to ensure cost efficiency.

Data Integrity

The integrity of the data used by AI models is critical to the quality of the outputs they generate. If the data is biased, outdated, or compromised, the AI models may produce incorrect or harmful outputs impacting the downstream application's customer experience. Ensuring that data is accurate, up-to-date, and appropriate for its clients while minimizing hallucination is essential to maintaining the integrity of LLM-driven applications.

Data integrity is particularly important in industries where AI is used to make critical decisions, such as healthcare, technology, finance, and law. In these cases, inaccurate or biased data can lead to serious consequences, including incorrect diagnoses, financial losses, service downtime and legal disputes.

Impact of These Technical Challenges and Risks on an Organization

The technical challenges and risks associated with GenAI can have a significant impact on an organization. Failure to address these issues can lead to:

- **Security Breaches:** Sensitive data may be exposed, leading to legal and financial repercussions.
- **Regulatory Non-Compliance:** Organizations may face fines, legal action, and damage to their reputation if AI-generated content violates regulations.
- **Increased Costs:** Uncontrolled consumption of AI services can lead to escalating costs, particularly in cloud-native environments. Lack of observability of these costs can lead to bill shock across teams and entire enterprises.
- **Brand Damage:** Inconsistent or inappropriate AI-generated content can harm the organization's reputation and erode customer trust.

However, organizations that effectively manage these challenges can harness the full potential of GenAI, driving innovation and gaining a competitive advantage.

Role of AI Gateways in LLM workloads

As early as 2019, API calls represented **83% of web traffic,** according to an Akamai State of the Internet report[5]. The emergence of LLM workloads has likely driven that percentage up and will push it even higher in the future.

API Gateways are already critical infrastructure for service management and connectivity. Many of the same characteristics that drove their historical adoption are even more important for enterprises learning to tame the potential chaos of AI-driven workloads.

However, GenAI use cases require more than just an API Gateway.

What is an AI Gateway?

An AI Gateway is a specialized type of API Gateway designed to manage, secure, and optimize interactions with AI services, particularly those that utilize large language models (LLMs) and other AI-driven functionalities. AI Gateways offer a centralized control point for managing AI-specific workloads, providing enhanced capabilities tailored to the unique requirements of AI applications. These gateways are crucial for organizations looking to integrate generative AI into their operations safely and efficiently. AI Gateways handle various tasks, such as managing prompt injections, controlling data flows, ensuring compliance with data governance policies, and optimizing the performance of AI models to reduce latency and resource consumption.

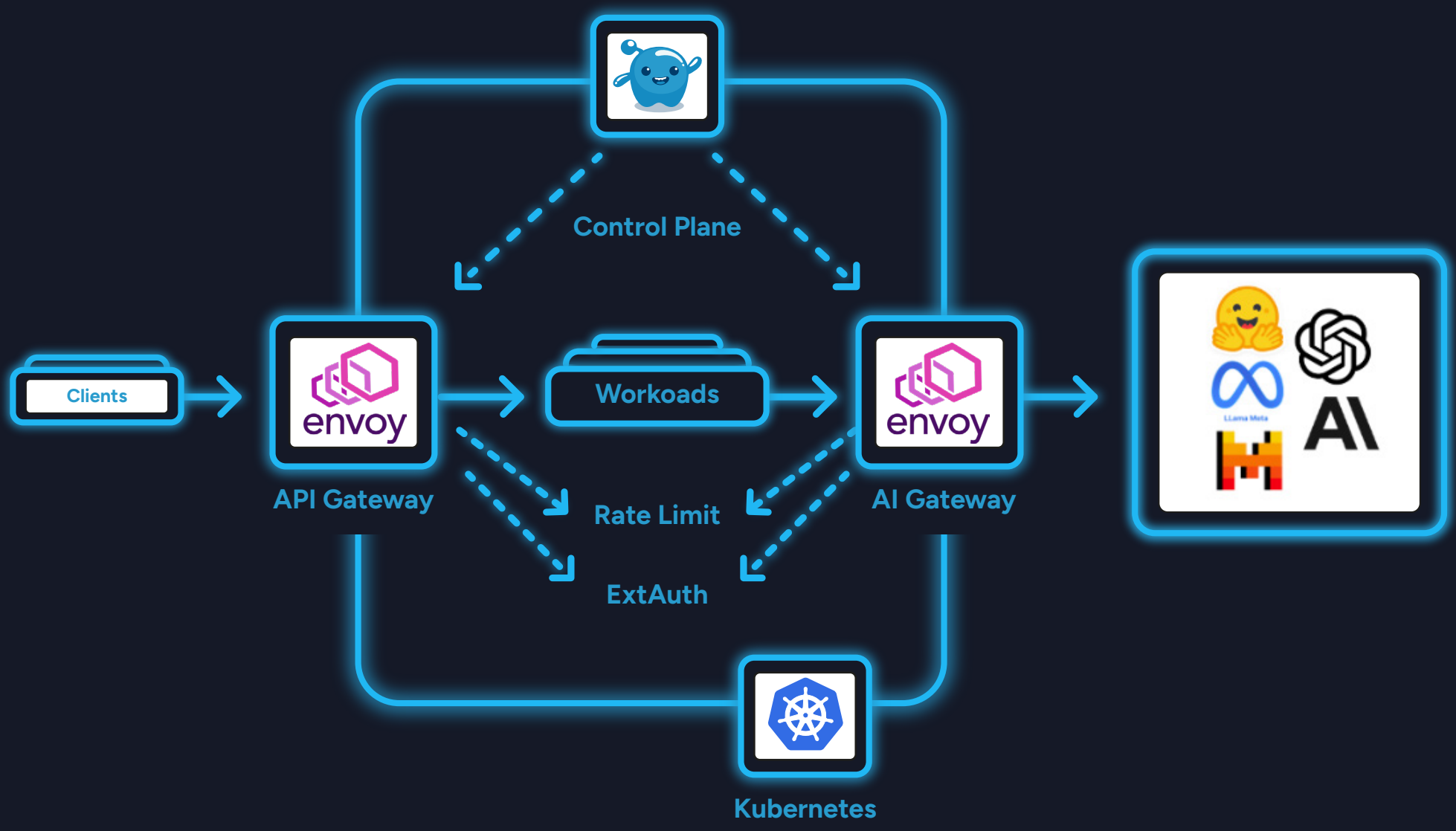
How Does an AI Gateway Differ from a Traditional API Gateway?

While traditional API Gateways and AI Gateways share some core functionalities—such as traffic management, security enforcement, and monitoring—AI Gateways are specifically designed to handle the unique challenges posed by AI workloads. Traditional API Gateways focus on managing RESTful or gRPC API calls between microservices in cloud-native environments, primarily ensuring secure, reliable, and scalable service-to-service communication. In contrast, AI Gateways incorporate additional capabilities to address the unpredictability and complexity of AI models.

[5] Akamai. (n.d.). The state of the Internet: Security research & threat reports. Akamai. https://www.akamai.com/security-research/the-state-of-the-internet

AI Gateways often feature enhanced observability and monitoring tools tailored to AI workloads, such as detailed metrics on model performance, cost management, and output analysis to mitigate the risks of unpredictable AI behavior. They also support more sophisticated request handling, including prompt management and modification, semantic caching, and advanced rate limiting, to ensure that AI models are used responsibly and effectively.

AI Gateways are often distinguished from their older API brothers by where they typically sit in the request flow. As shown in the diagram below, API Gateways commonly manage inbound requests to an organization's application network, allowing declarative configuration to drive behaviors like securing communication channels, traffic routing, authentication and authorization, and rate limiting. In contrast, AI Gateways typically manage outbound requests from an enterprise application network. They typically proxy traffic between apps and their LLM service providers, providing services like credential management, security, caching, and prompt management.



With an advanced AI Gateway like Gloo, a single configuration store can be used to program Envoy proxies in the same network that provide both ingress and egress services. (Or, you can combine both ingress and egress routing into a single proxy instance.)

To summarize, all AI Gateways are API Gateways. But not all API Gateways are AI Gateways. The latter provide specialized LLM-driven services that are not available in the former.

Why is an AI Gateway Necessary?

Given the complexity and unpredictability of GenAI, an AI Gateway is necessary to maintain control over how LLM services are accessed and consumed. By routing AI service requests through an AI Gateway, organizations can monitor, throttle, and log interactions with LLMs. They can also ensure that only authorized apps and users can access them and that their consumption is controlled. Without an AI Gateway, organizations may struggle to enforce security, compliance, and governance policies, leading to increased risk and operational inefficiencies.

AI Gateways provide a centralized point of control, allowing organizations to manage LLM interactions consistently across their entire IT environment. This is particularly important in cloud-native environments, where services are distributed and can scale rapidly.

How Do AI Gateways Address The Risks?

AI Gateways play a crucial role in managing the risks associated with GenAI. They act as a centralized control point for all LLM interactions, providing a layer of security, compliance, and governance. By routing all requests to AI services through an AI Gateway, organizations can enforce policies, monitor usage, and ensure that AI-powered applications align with their standards.

An AI Gateway can help mitigate risks in several ways:

- **Security:** AI Gateways can enforce authentication, authorization, and encryption for all API requests, ensuring that only authorized apps and users can access LLM services and that data is transmitted securely. They provide an array of tools to systematically address the chaos threatened by the security risks including prompt injection, insecure output handling, and sensitive information disclosure, to name a few.

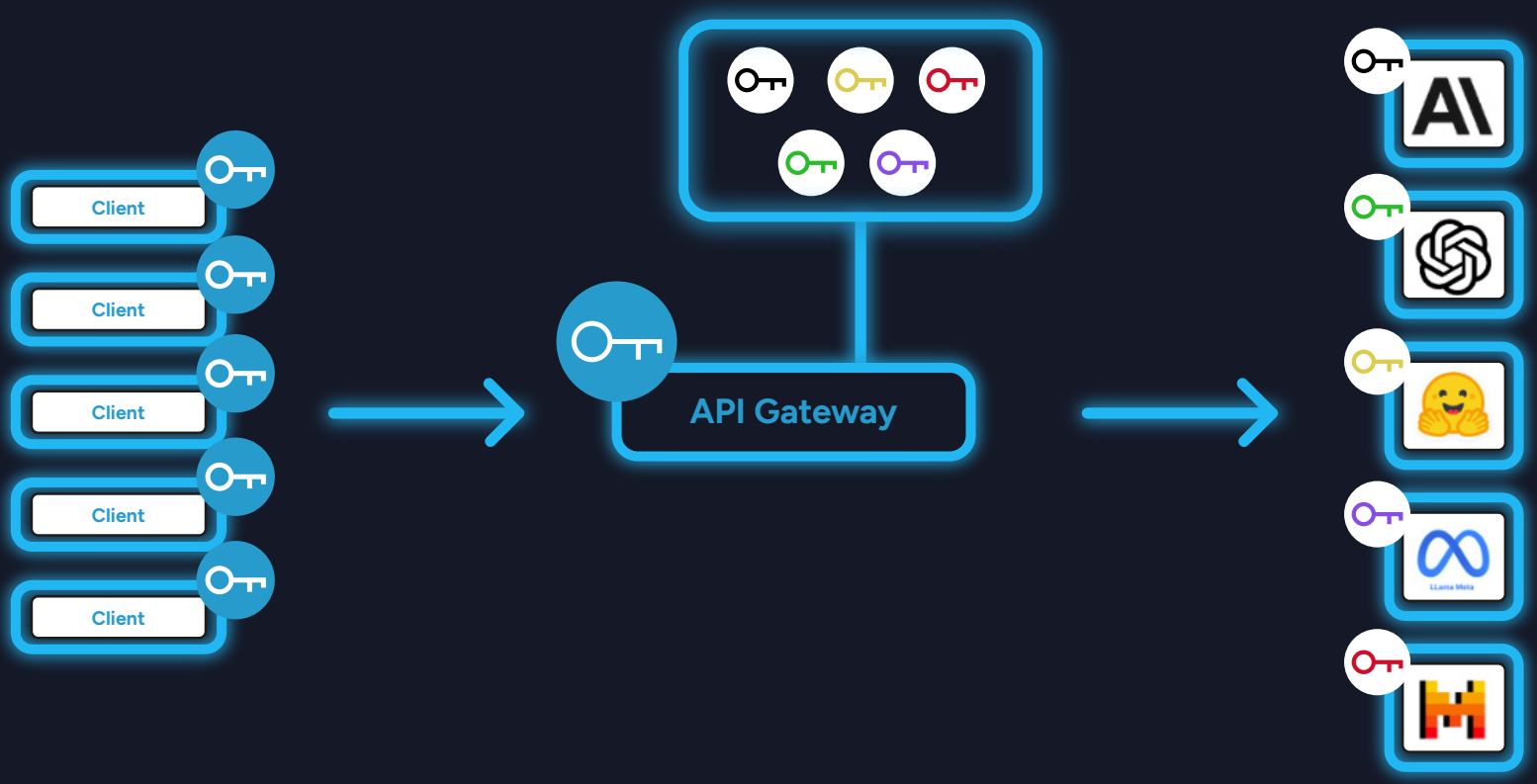
- **Credential Management:** With Gloo AI Gateway, you can centrally secure and store the API keys for accessing your AI provider in a Kubernetes secret in the cluster. The gateway proxy uses these credentials to authenticate with the AI provider and consume AI services. To further secure access to the AI credentials, you use fine-grained RBAC controls.

Challenge

- API keys and client identity managed individually for each LLM provider and passed differently
- Key management (tracking, revocation, refresh) distributed across all providers and clients
- Identity and access management integration with each LLM provider
- DIY problem for local LLMs

Solution

- Gateway manages LLM provider keys in protected store
- API Keys created by gateway can map to multiple providers, simplifying client development across LLMs
- Leverage advanced authN/Z in gateway for controlling access via JWT, OPA, etc.
- Centralized point of key management (tracking, revocation, refresh)



- **Compliance:** AI Gateways can log all LLM interactions, providing an audit trail that helps organizations demonstrate compliance with regulations. They can also enforce policies that prevent the generation of non-compliant content.
- **Prompt Management:** AI Gateways can enforce rules around prompt management, ensuring that AI models are only fed with well-structured prompts that guide them toward generating appropriate outputs.
- **Consumption Control:** AI Gateways allow enterprises to better observe and throttle LLM requests, preventing excessive consumption of AI services and helping organizations control costs.

The Right API Gateway Foundation Is Critical

Selecting the right API Gateway technology is crucial.

Multiple companies in the API gateway marketplace today are built on technology foundations that are 15+ years old and were architected with a set of assumptions that pre-date the cloud era and not fit for today's cloud native, GenAI workloads.

Solo's Gloo AI Gateway is built on **Envoy**, a modern, open-source proxy built for the cloud and designed with principles like declarative configuration in mind from the beginning. Choosing the right foundation for an
AI Gateway must include consideration of the underlying API Gateway and proxy technology.

Gloo AI Gateway from Solo.io is an ideal choice for managing GenAI services in cloud-native environments. It has been designed to deliver critical functionality to AI workloads addressing the challenges around; security and access management, observability and monitoring, performance and efficiency and delivering LLM enhancements and customizations comprehensively to AI workloads.

You can learn more about the importance of AI Gateway technology foundations in this guide **here.**

Start Implementing GenAI with Confidence

Containing the complexity of GenAI isn't easy. Doing it in a way that doesn't stifle innovation is even more difficult.

An effective AI Gateway strategy built on the right API Gateway foundation is the place to start. As GenAI evolves, so too will the challenges and opportunities it presents. Organizations that effectively manage the risks associated with GenAI while leveraging its capabilities will be well-positioned to lead in their respective industries.

AI Gateways like Gloo AI Gateway are essential tools designed to help bring order to the chaos and complexity that come with implementing GenAI. By providing a centralized, control point for managing AI services, Gloo AI Gateway helps organizations maintain security, compliance, performance, and control whilst enabling innovation and growth.

To learn more about the best practices of implementing AI Gateways, download our **Best Practices to AI Gateway guide.**

SOLO.IO

The Gateway to AI Innovation

www.solo.io