

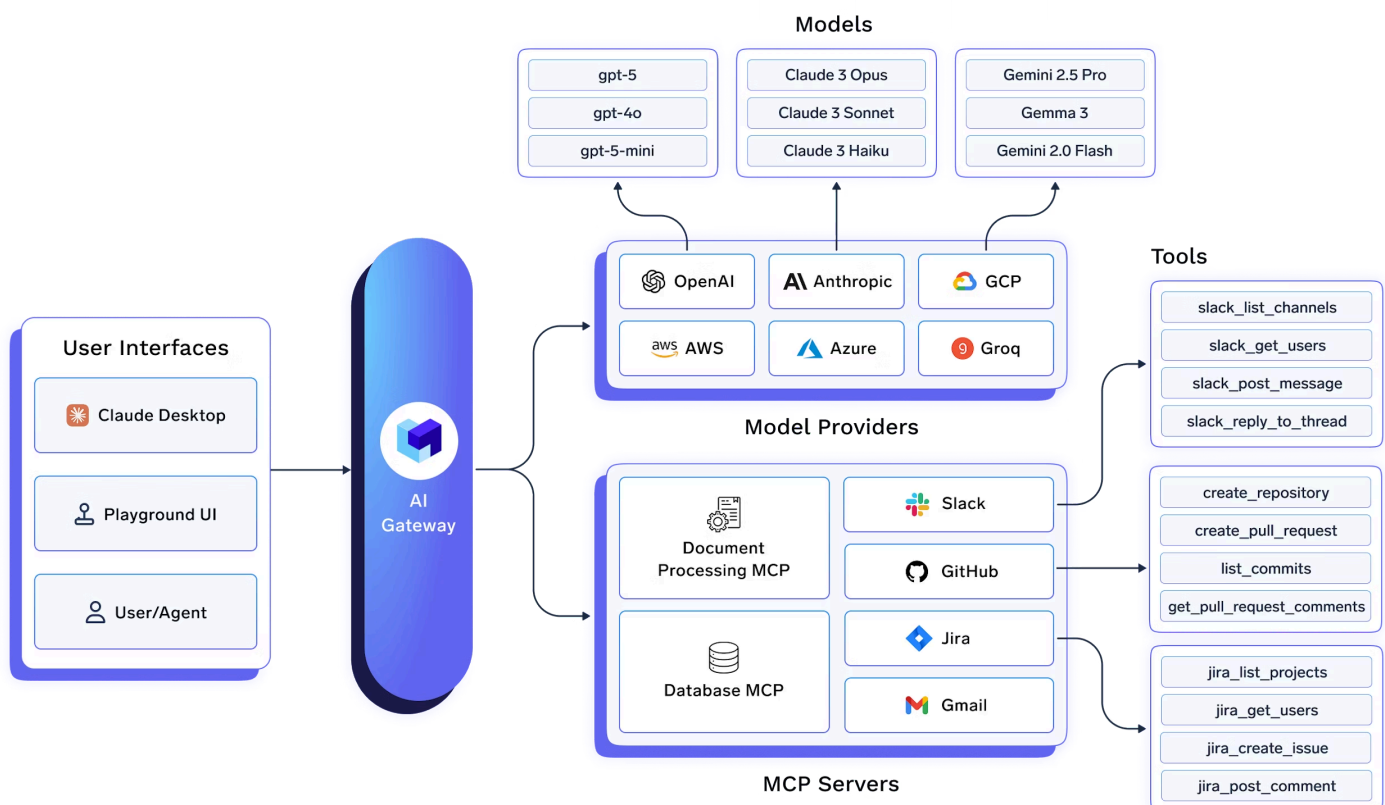
Get Started

Introduction to AI Gateway

 Copy page

TrueFoundry AI Gateway: A unified interface for accessing 1000+ LLMs with enterprise-grade security, observability, and governance

TrueFoundry AI Gateway is the proxy layer that sits between your applications and the LLM providers and MCP Servers. It is an enterprise-grade platform that enables users to access 1000+ LLMs using a unified interface while taking care of observability and governance.



Key Features

**Unified API Interface**

Call 1000+ LLMs using a single endpoint with unified API interface

API Keys Management

Generate and manage API keys for users/applications

Multimodal Inputs

Support for text, image, and audio inputs across compatible models

Access Control

Fine-grained access control and permissions management

Rate Limiting

Control Models Usage with flexible rate limiting policies per user/model/application

Load Balancing

Distribute requests across multiple model instances based on weight, latency or cost metrics.

Budget Limiting

Control spending and enforce cost limits for users, teams, and models

Guardrails

Content filtering and safety checks to ensure

Observability & Metrics

Opentelemetry compliant metrics and logging for all requests.

Prompt Playground

Centralized prompt playground with versioning and management system

Batch Predictions

Process multiple requests efficiently with batch processing

MCP Registry

Deploy and manage your own MCP servers with TrueFoundry AI Gateway.



Centralized Authn/Authz for all MCP Servers

One API key to access all MCP servers and their tools.

Virtual MCP Servers

Create virtual MCP servers combining specific tools from multiple MCP servers.

Agent Playground

Test Agents by adding tools and models from Playground

Build Agents with unified API for all MCP servers

Connect to MCP Servers with a single API in the gateway.

Rate Limiting and Observability for Tools

Coming Soon

Supported Model Providers

We integrate with 1000+ LLMs through the following providers.

💡 If you don't see the provider you need, there is a high chance it will just work as self hosted models or OpenAI provider. Please reach out to us at support@truefoundry.com and we will be happy to guide you.



3/20/26, 10:16 AM

Introduction to AI Gateway - TrueFoundry Docs



Databricks



truefoundry



AI21



AI21



Anthropic



Together AI



xAI



DeepInfra



Perplexity AI



Mistral AI



Groq



Self Hosted



OpenRouter



SambaNova



Cerebras

Supported APIs

The following accordions show which features are supported for each provider across different endpoints:



Legend:



Supported by Provider and Truefoundry



Provided by provider, but not by Truefoundry

-

Provider does not support this feature

Chat Completion (/chat/completions)

Provider	Stream	Non Stream	Tools	JSON I
OpenAI				

<https://www.truefoundry.com/docs/ai-gateway/intro-to-llm-gateway#introduction-to-ai-gateway>

4/8

Provider	Stream	Non Stream	Tools	JSON I
 truefoundry Azure OpenAI	✓	✓	✓	✓
Anthropic	✓	✓	✓	-
Bedrock	✓	✓	✓	-
Vertex	✓	✓	✓	-
Cohere	✓	✓	✓	✓
Gemini	✓	✓	✓	✓
Groq	✓	✓	✓	✓
AI21	✓	✓	-	✓
Cerebras	✓	✓	-	✓
SambaNova	✓	✓	-	✓
Perplexity-AI	✓	✓	-	✓
Together-AI	✓	✓	✓	✓
xAI	✓	✓	✓	✓
DeepInfra	✓	✓	✓	✓
Embedding (/embeddings)				
Image Generation (/images/generations)				
Image Edit (/images/edits)				
Image Variation (/images/variations)				
Audio Transcription (/audio/transcriptions)				
Audio Translation (/audio/translations)				
Text To Speech (/audio/speech)				



Batch (/batches)

Fine Tune

Files (/files)

Moderation (/moderations)

Model Response (/responses)

Completion (/completions)

Live / Realtime API

Compaction API

Messages API

Ecosystem & Integrations

Discover how TrueFoundry connects with your favorite AI frameworks and tools to streamline your ML development workflow.

[AI Frameworks](#) [Coding Assistants](#) [Agent Builder](#) [Guardrails](#) [Others](#)

 **CrewAI**

 **OpenAI Swarm**

 **OpenAI Agents
SDK**

 **Phidata**

 **Pydantic AI**

 **LangChain**



>

Deployment Options

The Truefoundry AI Gateway can either be used as a SaaS offering or deployed on-premise.

- **SaaS Offering:** You can directly use the gateway as a SaaS offering by signing up on our website, you can find the instructions [here](#).
- **Enterprise Deployment** for enterprise security and control. You can deploy the gateway in your cloud or on-premise. You can find the architecture and deployment instructions [here](#).

Frequently Asked Questions

What's the performance impact of using the gateway?

Can I deploy the gateway on-premise?

How do I integrate my self-hosted models?

Can I use the gateway without the full MLOps platform?

Was this page helpful?

 Yes

 No

Quick Start >



Powered by [mintlify](#)

>