JioCloud

AI Inference as a Service

Overview

AI Inference as a Service is JioCloud's managed LLM API platform, available through the JioCloud Marketplace. Subscribe to a large language model, call a standard chat completions API, and let the platform handle hosting, scaling, and security. Six production-grade models are available on demand - from lightweight 70B models to frontier-scale 671B reasoning models.

The service is built for engineering teams that need to embed language intelligence into applications without procuring GPU infrastructure, managing model deployments, or maintaining ML operations pipelines. A single API key, one REST endpoint, and a well-designed system prompt are all that is required to go from idea to working integration.

Key Features

- **Six Foundation Models**
  Access Llama 3.1 70B, Llama 3.1 405B, Qwen 2.5 72B, Qwen 3 235B, DeepSeek V3 671B, and DeepSeek R1 70B for diverse AI workloads.
- **Interactive Playground**
  Test prompts, validate model behavior, adjust parameters, and track token usage before deployment.
- **API Key Lifecycle Management**
  Create, rotate, and revoke API keys with automated rotation and zero downtime.
- **Enterprise Security and Compliance**
  SOC 2 Type II, ISO 27001, GDPR, and HIPAA compliant with TLS 1.2+ and AES-256 encryption.
- **Instant Provisioning**
  Provision a live, API-ready inference instance in minutes through the Marketplace.
- **Standard Chat Completions API**
  OpenAI-compatible Chat Completions API with standard tooling and client libraries.
- **Tier-Based Fixed Pricing**
  Choose Free, Standard, Professional, or Enterprise tiers with fixed token allocations.
- **Token Usage Monitoring**
  Monitor real-time token usage and capacity against tier limits.

JioCloud

Benefits

- **Zero Infrastructure Overhead**
  No GPU procurement, model deployment, or ML operations pipeline required. The platform manages everything.
- **Test Before You Ship**
  Playground lets teams validate prompts and tune parameters before committing to production code.
- **Enterprise Security Built-In**
  SOC 2, ISO 27001, GDPR, HIPAA compliance; end-to-end encryption; no data persistence after each request.
- **Production-Ready in Minutes**
  Marketplace order to working API endpoint in under 5 minutes - the fastest path from use case to integration.
- **Standard API, No Lock-In**
  OpenAI-compatible chat completions format - portable integration pattern that works with existing HTTP tooling.
- **Predictable Cost Model**
  Fixed tier pricing with Free Tier available for development and evaluation. No surprise infrastructure costs.

Supported Integration Clients

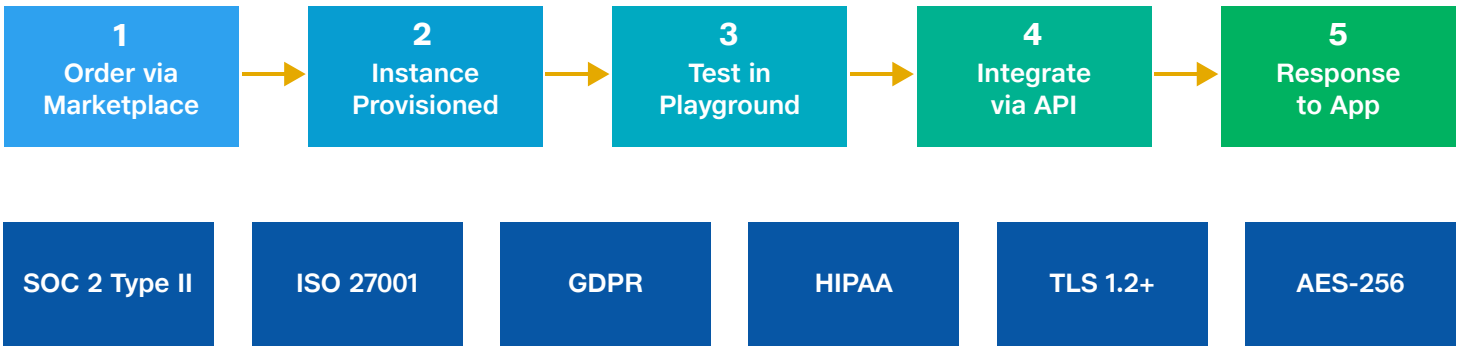
REST / cURL | Python | Node.js / JS | Java / JVM | GO | .NET / C#

Specification

Specification	Details
Deployment Model	Fully managed - JioCloud CloudXP Platform
API Endpoint	POST /v1/chat/completions
Authentication	API Key (header: apikey: <KEY>)
Available Models	Llama 3.1 70B · Llama 3.1 405B · Qwen 2.5 72B · Qwen 3 235B
Response Format	JSON - standard chat completions schema
Pricing Tiers	Free · Standard · Professional · Enterprise
Rate Limiting	Per-tier token-per-minute limits; enforced by the platform
Data Persistence	None - request and response data not stored after each call
Compliance	SOC 2 Type II · ISO 27001 · GDPR · HIPAA
Encryption	TLS 1.2+ in transit · AES-256 at rest
API Key Operations	Create · Rotate · Revoke (console-managed)
Access Interface	JioCloud Console + REST API

JioCloud

Architecture Diagram



Use Cases

- **IT and Customer Operations**
  Integrate AI Inference as a Service into document processing pipelines, support ticket triage systems, and customer communication workflows. A single API call performs multi-task analysis - classification, extraction, summarization - returning structured output that feeds directly into operational systems. Reduces manual processing time by 10× to 100× at scale.

- **Supply Chain and Business Intelligence**
  Deploy as the analytical engine in monitoring and intelligence platforms. The service processes news feeds, regulatory filings, and supplier communications at machine speed - classifying risk signals, generating plain-language narratives, and surfacing the items that require human attention. Procurement and operations teams see structured, actionable output without reading raw source text.

- **Application Development and Rapid Prototyping**
  Engineering teams embed language intelligence into products using the standard chat completions API. No ML infrastructure required - provision a Free Tier instance, prototype in the Playground, and graduate to a production tier when ready. From use case idea to working prototype in hours, not quarters.