MCP SERVER | NO CODE | CLOUD HOSTED

Eden AI MCP for AI Agents

Monitor LLM API Costs and Usage Across Providers

Eden AI connects your agent to an entire ecosystem of large language model providers. It gives you one central place to manage all your automation workflows, track API usage across services like OCR and image generation, and monitor costs instantly from multiple sources (OpenAI, Google, AWS, etc.). Stop checking dozens of dashboards; get unified control over your AI stack right where you work.

A+ Quality Score 100/100

unified-api | llm-orchestration | ai-providers | model-management | api-usage-tracking | workflow-automation

MAINTAINED BY VINKIUS CLOUD FOR AI AGENTS · JUNE 2026
vinkius.com/mcp/eden-ai

SCAN TO VIEW

The connectivity layer between AI and the world's software.

Vinkius sits between AI and every application. All communication passes through Vinkius Cloud via the Model Context Protocol (MCP) — with governance, observability, and security at every layer.

Your AI Connections Run Through Vinkius Cloud

The world's largest managed MCP catalog

Vinkius is the connectivity layer where AI connects to the software your business already runs. We handle the hosting, the security, the credentials, the uptime — you get agents that actually do things.

We operate the world's largest managed MCP catalog. Major SaaS platforms, CRMs, databases, and cloud providers — running, monitored, production-ready. This MCP server is hosted and maintained by the Vinkius Cloud for AI Agents.

> *The agent doesn't manage credentials, doesn't manage uptime, doesn't manage security. Vinkius does.*
>
> — Architecture principle

Four Pillars of the Vinkius Runtime

01 — Security by design

Credentials stay encrypted at rest via AES-256. The AI agent never touches raw keys — they're injected into a sandboxed V8 isolate at runtime. Actions are logged, and connections have an emergency kill switch.

02 — Built on MCP Fusion

This MCP server was built with **MCP Fusion**, the open-source framework (Apache 2.0) that powers the entire Vinkius catalog. Schema-as-firewall strips undeclared fields, compiled PII redaction runs at zero overhead, and cryptographic lockfiles produce git-diffable audit trails.

03 — Deterministic observability

Eight immutable metrics per endpoint: request volume, p95 latency, error rate, active connections, cost attribution. A live payload feed logs every tool call with mutation detection.

04 — Autonomous operations

Servers are deployed, monitored, and patched autonomously. New capabilities and security patches ship weekly. Zero-downtime deployments ensure continuous availability across all managed MCP servers.

AES-256	Ed25519	24h TTL	V8 Isolate
Encryption at rest	PKI vault signatures	Ephemeral session keys	Sandboxed execution

One Token. Instant Access.

Every MCP server on Vinkius is accessed through a **Connection Token**. Tokens are generated in the cloud dashboard and produce a unique MCP endpoint URL. Paste this URL into any MCP-compatible client — no SDK required.

A single token can serve **multiple AI clients simultaneously**, or you can issue separate tokens per client for granular access control. Each token tracks its own request count, last activity timestamp, and can be individually enabled or revoked.

MCP ENDPOINT

```
https://edge.vinkius.com/{token}/mcp
```

Claude · Cursor · VS Code · Windsurf · Grok · Gemini

Security Is the Architecture

Security in Vinkius is not a feature — it's the foundation of the runtime. The gateway enforces multiple independent protection layers between AI agents and third-party APIs.

01 — Ed25519 PKI Vault

Every workspace has an Ed25519 Master Key. Session keys are generated ephemerally (24h TTL) and signed by the Master Key. Credentials never leave the vault boundary.

02 — V8 Isolate Sandboxing

Tool code runs inside isolated-vm V8 isolates with 64 MB memory caps and per-request timeouts. No filesystem access, no network access except through the SSRF-guarded fetch bridge.

03 — SSRF Guard

All outbound HTTP requests are DNS-resolved and validated before execution. Private IP ranges (10.x, 172.16-31.x, 192.168.x, AWS metadata 169.254.x) are blocked at the network layer.

04 — DLP & PII Redaction

A ResponseGuard pipeline intercepts every tool response. Configurable redaction patterns strip sensitive fields (emails, SSNs, card numbers) before data reaches the AI agent.

05 — Cryptographic Audit Trail

Every request is signed into a SHA-256 hash chain with Ed25519 signatures. Events form a tamper-proof, SIEM-exportable forensic record.

06 — Honeytoken Trap System

Phantom credentials are injected into isolated environments. If a honeytoken is used outside Vinkius infrastructure, the server is quarantined instantly.

Emergency Kill Switch

EU AI Act Art. 14(1) Compliant

The kill switch is an **emergency halt** mechanism — not a simple toggle. When triggered, it executes three actions atomically:

01 — Server deactivated

The MCP server is immediately taken offline across the entire cluster.

02 — All tokens revoked

Every connection token is invalidated. Total lockout — reconnection blocked until new tokens are issued.

03 — WebSocket connections killed

Active connections terminated via Redis pubsub broadcast. Propagates to every runtime node in the cluster.

Full Visibility. Zero Guesswork.

The Vinkius cloud dashboard includes a full MCP Governance suite — real-time analytics and security controls for production AI operations.

Control Plane

KPI dashboard with request volume, latency, success rate, token consumption, and AI-generated operational briefings.

Agent Activity

Which AI clients are connecting, how often, and what they're doing — real-time session tracking.

FinOps

Cost tracking per tool, payload compression savings, budget optimization signals, and consumption trends.

Tool Health

Slowest and most error-prone tools, with actionable root-cause insights and performance baselines.

Firewall & DLP

PII redaction activity, sensitive data protection counters, and security event timeline.

Incident Log

Error trends, failure rates, status-code breakdowns, and forensic audit trail access.

> Get started at **cloud.vinkius.com** — connect your AI agent in under 60 seconds.

Eden AI MCP

6 tools available
Cloud-hosted on Vinkius

Building complex applications with LLMs means dealing with a messy web of APIs, billing cycles, and provider limits. Eden AI solves that headache by unifying the entire AI infrastructure into one conversational interface. You can manage multi-step workflows—from generating text to extracting data or creating images—all without leaving your chat window. Instead of building separate pipelines for every service, you tell your agent what job needs doing, and it routes the request to the best provider available. It also gives you a real-time ledger: check credits, monitor costs over time, and see exactly which model is driving up your spending. If you use Vinkius, connecting Eden AI through the catalog makes sure that all these powerful capabilities are accessible instantly via any MCP-compatible client.

Core Capabilities

01 — Monitor API Costs

Get an immediate snapshot of your credit balance and monitor consumption trends across different services.

02 — Execute Specialized AI Tasks

Run complex, multi-modal tasks like translating text, reading images via OCR, or generating artwork using one command.

03 — Generate Embeddings

Create vector embeddings from any piece of text to power search and retrieval systems.

04 — Route Chat Completions

Send prompts and let the agent choose the optimal provider or model for the best chat response.

05 — List Model Capabilities

Check which specific large language models are supported by the platform, along with their feature set.

One Click on Vinkius — From Prompt to Execution

Available at **vinkius.com/mcp/eden-ai** — connect your AI agent in three steps.

01 Connect your AI client to Eden AI on Vinkius and authorize it using your API key.

02 Use natural conversation to define complex operations, such as 'list all active workflows' or 'show me the cost of image generation'.

03 The agent executes the command, providing real-time data, pricing comparisons, and operational status updates.

The bottom line is that you talk to your AI client like a coworker, and it handles all the underlying complexity of routing, billing, and execution for every major AI provider.

Built For

This MCP is built for engineers who are spending too much time manually tracking usage across multiple cloud consoles. It's for anyone building complex, multi-provider LLM applications that need financial oversight and workflow management built in.

AI Developer

You use this MCP to test model availability and feature pricing on the fly before committing to a production stack.

MLOps Engineer

You manage workflow execution status, audit organizational credit balances, and track consumption against budgets.

Product Manager (AI)

You monitor which features are being used most often and compare pricing across providers to advise on product feature costs.

What Changes When You Connect

01 Cost Visibility: Use `monitor_consumption` to track exactly how much every feature costs, eliminating surprise bills.

02 Unified Access: Don't switch consoles. Your agent handles specialized tasks like OCR or image generation using the `universal_ai` tool.

03 Model Selection: Need a specific chat style? The `chat_completion` tool lets you route requests to different providers and models instantly.

04 Budget Control: Keep track of your operational spending anytime by calling `check_credits` for an immediate balance report.

05 Data Readiness: Use the `create_embedding` function to generate high-quality vector data without leaving the chat interface.

Real-World Applications

Auditing Quarterly AI Spending

An MLOps engineer needs a full picture of spending. They ask their agent to use `monitor_consumption` for the last quarter, getting a report that breaks down costs by feature and provider.

Building a Multi-Modal Chatbot

A developer needs a chatbot that can not only answer questions but also read receipts from uploaded images. They use `universal_ai` to integrate OCR alongside standard chat completion calls.

Selecting the Best Model for a New Feature

A Product Manager wants to know if Google or OpenAI is cheaper for sentiment analysis. The agent uses provider intelligence (via `list_embedding_models` and pricing data) to compare costs before implementation.

Ensuring Sufficient Funds for a Launch

An operations team lead needs to know if the current budget covers the deployment. They run a quick check using `check_credits` before initiating any large-scale testing.

Patterns to Avoid

Juggling multiple provider dashboards

✗ AVOID

Spending an hour logging into OpenAI, then Google Cloud Console, and then AWS to compare costs for a simple embedding task.

✓ INSTEAD

Instead, just ask your agent to run `list_embedding_models` or use the `chat_completion` tool. It pulls all the necessary data points from one place.

Ignoring usage tracking

✗ AVOID

Deploying a new workflow that suddenly spikes costs, but no one knows which specific model caused it because monitoring was never set up.

✓ INSTEAD

Run `monitor_consumption` regularly. It provides historical data and cost breakdowns so you catch spikes the moment they happen.

Forgetting specialized capabilities

✗ AVOID

Thinking standard chat completion is enough when your workflow requires image analysis or document reading.

✓ INSTEAD

Use the `universal_ai` tool. It handles OCR and other complex media inputs, letting you treat it like a simple text prompt.

The Right Fit

Use this MCP if your primary pain point is cost transparency or managing multi-provider complexity. You need to know *if* a feature works across different AI providers (e.g., comparing OpenAI vs. Google pricing) and you're building workflows that involve more than just text generation, like images or documents. Don't use it if your only goal is simple, single-provider chat completion; in that case, connecting directly to a specific provider might be overkill. But if you need to audit credits via `check_credits` or track usage with `monitor_consumption`, this is exactly what you need.

Eden AI MCP for LLM API Cost Management

Right now, managing an application that uses multiple large language models feels like juggling chainsaws. You're checking provider A's dashboard for cost, then logging into provider B to see if they support the same feature, and finally running a separate spreadsheet just to tally up your total spending. It takes hours of manual cross-checking.

With this MCP, you simply ask your agent what you need. If you want pricing intelligence or usage reports, it handles the complexity for you. You get real-time cost data in conversation—no spreadsheets, no logins required.

Eden AI MCP for Unified Workflow Orchestration

The old way meant building separate API calls for every function: one script for text, another library for image generation, and a third service just to handle the embedding. You had to manage three different connection types and error handlers.

Now, you define the workflow once. The agent takes that single request and executes all necessary steps—text completion, specialized media processing, and data storage—seamlessly in one conversational flow.

Explore 6 Eden AI Tools for LLM Orchestration & Costs

Use these tools to check balances, run specialized tasks like OCR, generate embeddings, or monitor costs across multiple AI providers.

#	TOOL	DESCRIPTION
01	`check_credits`	Quickly checks your current credit balance within the Eden AI platform.
02	`universal_ai`	Runs specialized AI tasks like OCR, translation, or image generation across providers.
03	`list_embedding_models`	Lists all available embedding models provided by Eden AI.
04	`chat_completion`	Generates a chat response using the best multi-provider routing, letting you select or specify the model used.
05	`monitor_consumption`	Provides detailed reports on your API usage and associated costs over specified time periods.
06	`create_embedding`	Generates text embeddings using the Eden AI platform for vector databases.

See It in Action

Real prompts you can use once this MCP is connected to your AI agent through Vinkius Cloud.

U What are my remaining credits, and how much did I spend on embeddings today?

Eden AI Usage Report

- **Current Credit Balance:** $78.55
- **Today's Consumption:** $3.12
- **Breakdown:**

* `create_embedding`: 1,200 units ($1.80)
* `chat_completion` (Google): 4 calls ($1.32)

U Compare the pricing for OCR and translation between the three main providers.

Here is the service level data I pulled:

FEATURE	OPENAI PRICING	GOOGLE PRICING	AWS PRICING
OCR	$0.01 / 1k chars	$0.009 / 1k chars	N/A
Translation	$0.02 / 1M chars	$0.015 / 1M chars	$0.025 / 1M chars

Google looks like the leader for cost-effective translation right now.

U List all available embedding models and tell me which ones are best for financial data.

I checked the registry. We support six main types:

- `text-general` : Best for standard articles.
- `finance-v2` : **Recommended.** Optimized for transactional data structures.
- `legal-compliance` : Great for policy documents.

The `finance-v2` model is currently available across all three major providers and shows high accuracy rates.

Frequently Asked Questions

01 **How does Eden AI help me compare LLM pricing in real time?**

Eden AI gives you immediate, comparative pricing reports for features like sentiment analysis or image generation. You don't need to check three different vendor websites; the agent pulls and organizes all the data instantly.

02 **Can Eden AI handle workflows that use multiple kinds of AI services?**

Yes, it manages everything from simple text chat completions to complex media tasks like OCR and image generation. You tell your agent the goal, and it executes all necessary steps in sequence.

03 **What if I need to switch providers for a single task?**

The MCP handles provider routing automatically. If you ask for chat completion, it can route that request to OpenAI or Google based on your criteria, letting you test which model performs best.

04 **How do I track my spending across different departments?**

You use the monitoring tool to get a comprehensive audit of API consumption. You can filter usage reports by feature and see exactly which department or project is driving up costs.

05 **Is Eden AI only for text-based AI tasks?**

No, it supports multi-modal inputs. You can use the platform to analyze images (OCR) and generate multimedia content, not just process pure text prompts.

Go Live in 60 Seconds

Get your connection token from **cloud.vinkius.com**, then paste the endpoint URL into any MCP-compatible client.

YOUR MCP ENDPOINT

```
https://edge.vinkius.com/[TOKEN]/mcp
```

CLIENT	WHERE TO CONFIGURE
Claude AI	Profile → Customize → Connectors → "+" → Add custom connector → Paste endpoint
Cursor	Settings → Features → MCP Servers → "+ Add New MCP Server" → Type: SSE → Paste endpoint
VS Code	Ctrl/Cmd+Shift+P → "MCP: Add Server" → add `"eden-ai": { "url": "…" }`
Windsurf	MCP Settings → `mcp_settings.json` → Add endpoint URL
ChatGPT	Settings → Tools & plugins → Add MCP server → Paste endpoint
Gemini	Extensions → Add MCP Server → Paste endpoint URL

ASK AN AI ABOUT THIS

Let your preferred AI explain this MCP server

- Ask ChatGPT
- Ask Claude
- Ask Perplexity
- Ask Gemini
- Ask Grok

READY TO CONNECT

Eden AI is live on Vinkius Cloud.

Get your connection token, paste it into your AI agent, and start building. No SDK. No deployment. Just results.

Start at cloud.vinkius.com →

vinkius.com · support@vinkius.com

INDEPENDENT PLATFORM DISCLAIMER

Vinkius is an independent platform and is not affiliated with, endorsed by, sponsored by, verified by, or otherwise authorized by Eden AI. All third-party trademarks, logos, and brand names are the property of their respective owners. Their use in this document is strictly for informational purposes to identify service compatibility and interoperability.

DOCUMENT INFORMATION

Generated	June 2026
MCP Server	Eden AI MCP
Server ID	`019d758c-d6bf-701d-bbd9-ad0c1df85d1f`
Platform	Vinkius Cloud for AI Agents
Endpoint	`https://edge.vinkius.com/{token}/mcp`

LICENSE & USAGE

This document is generated automatically by the Vinkius PDF Engine. Content reflects the MCP server configuration at the time of generation and may change as updates are deployed. For the most current information, visit vinkius.com/mcp/eden-ai.