MCP SERVER | NO CODE | CLOUD HOSTED

Together AI MCP

Run open-source LLMs and ML tools in one place

Together AI connects your agent to hundreds of open-source LLMs for real-time inference, image generation, and model training. Use this MCP to generate vectors, run complex chats, or fine-tune models like Llama and Mixtral directly from any compatible client.

A+ Quality Score 100/100

llm | model-inference | fine-tuning | open-source-ai | machine-learning | api-deployment

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

Together AI MCP

7 tools available
Cloud-hosted on Vinkius

Need to get bleeding-edge AI models into your daily workflow? This Together AI MCP connects your agent to an entire library of open-source LLMs. You can query powerful models—like Llama, Mixtral, and others—to run chats or perform basic text completions without leaving your chat environment. It's built for developers who need world-class inference speed right now. Beyond just chatting, you can generate rich vector embeddings instantly from raw text logs to populate any analytical database. Need visuals? Instruct the MCP to create images using detailed descriptions. You can also provision and track custom fine-tuning jobs by pointing to a base model and a dataset file. Once connected via Vinkius, your agent gains access to this full suite of capabilities, letting you manage everything from basic text generation to complex model training cycles.

Core Capabilities

01 — Run Advanced Conversations

Your agent handles multi-turn conversations using powerful open-source models by providing a simple chat history and requesting completion.

02 — Generate Text Content

You can execute basic text generation tasks, giving the MCP a model ID and a prompt to get immediate textual output.

03 — Create Image Assets

The MCP generates original images when you supply a detailed physical description (prompt) for an external diffusion model to use.

04 — Prepare Data Embeddings

You can convert raw input texts into rich vector embeddings, which are ready to index in your analytical databases.

05 — Manage Model Training

The MCP creates custom fine-tuning jobs using a base model and a specific dataset file, and you can track the status of those jobs.

06 — Discover Available Models

You list all models available on the Together network to find the best engine for your NLP or vision task.

One Click on Vinkius — From Prompt to Execution

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

01 Sign up for this integration and fetch a developer API key from the api.together.xyz control panel.

02 Plug that key into your agent, specifying which models you need access to.

03 Your AI client then executes sub-second serverless inference directly inside your command interface.

The bottom line is, it lets your agent use advanced LLMs and ML tools without needing to switch environments or write complex boilerplate code.

Built For

This MCP is for the AI Developer who needs production-grade model access right now. It's for engineers tired of juggling multiple APIs, switching between a chat window and a separate ML dashboard just to run one task.

Machine Learning Engineer

Uses the MCP to bulk-generate vectors from raw logs or provisions custom training runs without leaving their main agent interface.

Software Engineer

Tests open-source completions using alternative solutions, like Llama 3, natively in code editors alongside other development tasks.

AI Developer

Orchestrates fine-tuning parameters and launches compute jobs directly from their chat environment instead of needing a separate CLI or dashboard switch.

What Changes When You Connect

01 Stop switching between dashboards. You can generate embeddings or run chat completions using the `generate_embeddings` tool, all from your agent's prompt.

02 Manage entire model lifecycles—from initial testing to production fine-tuning. Use `create_finetune_job` and then check status with `list_finetune_jobs` without leaving your workflow.

03 Need a visual asset? Simply call `generate_image` by providing a detailed prompt; you get an image file back, not just text.

04 Explore the best model for any task. Use `list_available_models` to see hundreds of open-source options before running a single inference.

05 The `chat_completion` tool handles complex conversational flow, making your agent feel much more natural than simple prompt/response cycles.

Real-World Applications

Building a Retrieval System

An engineer needs to index thousands of internal documents. Instead of writing a dedicated script, they ask their agent to use `generate_embeddings` on the raw text chunks and pipe those vectors directly into their vector store.

Creating Content for Marketing

A marketing specialist needs an illustration for a blog post. They prompt their agent, asking it to use `generate_image` with a detailed description (e.g., 'a futuristic cityscape at sunset'), and the image appears instantly.

Updating a Core Model

A machine learning engineer wants to adapt an open-source LLM for internal jargon. They use `create_finetune_job` with their base model ID and dataset, then monitor progress using `list_finetune_jobs`.

Testing Model Alternatives

A developer wants to compare Llama 3 against Mixtral for a chat feature. They use the agent's ability to run completions (`chat_completion`) multiple times in one session, comparing outputs side-by-side.

Patterns to Avoid

Hardcoding API Calls

✗ AVOID

Writing complex code blocks with explicit model endpoints and separate libraries just to run a single chat query or generate vectors.

✓ INSTEAD

Use the agent's built-in tools. Simply tell your agent to `chat_completion` after specifying the desired model ID, letting the MCP handle the boilerplate API connection.

Ignoring Model Variety

✗ AVOID

Assuming one powerful LLM is good enough for everything—using a single endpoint for chat, image generation, and embedding creation.

✓ INSTEAD

Use `list_available_models` to select the best specialized tool. For instance, use `generate_embeddings` instead of asking your main chat agent to do vector math.

Manual Job Tracking

✗ AVOID

After submitting a fine-tuning job, having to log into a separate web console every few minutes just to see if the process succeeded or failed.

✓ INSTEAD

Use `list_finetune_jobs` inside your agent. You submit the job with `create_finetune_job`, and then check its status within the same conversational thread.

The Right Fit

Use this MCP if your primary need is accessing a wide, current selection of open-source models for diverse tasks—chatting, embedding, image creation, or training. It's ideal when you need to prototype quickly and test multiple model architectures in one place. Don't use it just because you want an LLM; the value here is in its breadth (the many available tools). If your goal is only simple text generation with a single provider and no other ML needs, another dedicated completion tool might suffice. But if you are building anything that requires data preparation (embeddings), visual assets (`generate_image`), or model customization (`create_finetune_job`), this MCP is necessary.

The Headache of Model Tooling

Today, integrating different AI functions means a lot of manual work. You run the chat in one window, then copy text out to a separate vector database UI to generate embeddings, and if you need an image, you have to switch over to an art generator's web portal. It's a constant cycle of copying, pasting, and switching tabs just to get one feature working.

With this MCP, that manual handoff disappears. You tell your agent what you want—whether it's generating vectors using `generate_embeddings` or getting text completions via `text_completion` — and the model runs everything internally. The result appears right where you asked for it.

Together AI: Model Operations

The specific manual steps that vanish include setting up separate API keys for different models and manually tracking job states across multiple vendor dashboards. You also stop having to decide if the model you are using is right for the task.

Now, your agent handles all of it. You simply ask the MCP to manage the workflow—for example, running `chat_completion` first, then asking it to summarize the output and generate embeddings with `generate_embeddings`. It's one continuous flow.

Together AI with 7 Tools

These tools let you run model inference for chatting, text generation, image creation, embedding vectorization, and managing custom model training jobs.

#	TOOL	DESCRIPTION
01	chat_completion	Executes a multi-turn conversation using specified Together AI models and message history.
02	text_completion	Performs standard text generation by receiving only a model ID and an initial prompt.
03	create_finetune_job	Initiates a new model fine-tuning job using a specified base model and training dataset file.
04	generate_embeddings	Converts an array of input texts into rich vector embeddings for use in databases.
05	generate_image	Creates a visual image by translating a detailed descriptive prompt into a picture using external diffusion models.
06	list_finetune_jobs	Retrieves and shows the current status of all fine-tuning jobs you've created.
07	list_available_models	Lists every AI model currently supported on the Together AI network for your review.

See It in Action

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

U List all the models currently available on Together AI.

I've fetched 132 available models. Here are the top chat models:
- meta-llama/Llama-2-70b-chat-hf
- mistralai/Mixtral-8×7B-Instruct-v0.1
- google/gemma-7b-it
Ask if you want the embedding or image models only.

U Generate an embedding array using model `togethercomputer/m2-bert-80M-8k-retrieval` for the sentence 'The cat sat on the mat'.

Embeddings generated successfully. Dimensions: 768. Sample values:
[-0.0124, 0.0411, 0.0812, ... -0.0123]

Frequently Asked Questions

01 **How do I know what models are available using the Together AI MCP?**

You use the `list_available_models` tool. This instantly provides a list of all supported LLMs, letting you pick the best one for your chat or embedding task.

02 **Can I fine-tune my own model with Together AI MCP?**

Yes. You start by calling `create_finetune_job`, providing a base model and your training data file, and then monitor the progress using `list_finetune_jobs`.

03 **What is the difference between chat_completion and text_completion?**

Use `chat_completion` when you need multi-turn conversations that require a history of messages. Use `text_completion` for simple, single-shot prompts.

04 Does Together AI MCP handle image generation?

Yes, it handles images using the `generate_image` tool. Just give it a detailed text description and receive an image asset back.

05 Is this only for coding tasks?

No. While great for developers, you can use this MCP for anything that needs complex AI: data vectorization (`generate_embeddings`), content creation, or model training.

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 `"together-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

Together 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 Together 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	Together AI MCP
Server ID	`019d7613-8fef-713a-ac52-03cbd6e1202c`
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/together-ai.