MCP SERVER | NO CODE | CLOUD HOSTED

Together AI MCP

Power Multi-Modal AI with Open Source Models

Together AI connects your AI agent to over 100 open-source models, giving you a unified platform for everything from text chat and image creation to audio transcription and model fine-tuning. It powers advanced generative AI applications without requiring you to manage any cloud infrastructure.

A+ Quality Score 100/100

llm | generative-ai | llama-3 | image-generation | fine-tuning

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

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

27 tools available
Cloud-hosted on Vinkius

You can connect this MCP to your agent to access the world's fastest inference cloud for open-source models. This connector gives you a complete toolkit for generative AI, handling everything from basic text chat and creating stunning images to processing audio files or training custom model checkpoints. Need to build complex search features? You generate vector embeddings and rerank documents using specialized tools. Plus, if your application needs constant performance, you can create dedicated endpoints with predictable scaling. Whether you're building an app that talks, draws pictures, or analyzes voice recordings, this MCP keeps all the power running through a single connection point via Vinkius.

Core Capabilities

01 — Generate Text and Chat Responses

Your agent can generate high-quality text responses for conversations using various open-source models.

02 — Create Visual Media

The MCP handles generating realistic images or full videos based on simple text prompts.

03 — Process Audio Files

You can convert spoken words into written transcripts, or turn plain text into natural-sounding speech for voiceovers.

04 — Build Knowledge Retrieval Systems

It generates vector embeddings from documents and reranks results so your agent finds the most relevant information quickly.

05 — Manage Model Training

You can run fine-tuning jobs, upload data files, and manage dedicated endpoints for reliable performance.

One Click on Vinkius — From Prompt to Execution

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

01 First, subscribe to this MCP and provide your Together AI API Key.

02 Next, your agent uses the connection to call specific tools—for example, telling it to create a video or generate embeddings.

03 Finally, you get back the resulting data payload, like an image URL or a transcript file path.

The bottom line is that you use your AI client to trigger advanced generative tasks without having to worry about managing underlying model servers.

Built For

This MCP serves developers who build consumer-facing AI tools, data scientists needing custom models, and product managers trying to rapidly prototype multi-modal features. If your job involves integrating more than one type of AI output (text, image, audio), this is for you.

AI Application Developer

They use the MCP to connect text generation alongside media tools, building a single conversational agent that can both talk and draw.

Data Scientist

They leverage the model management capabilities, uploading custom datasets for fine-tuning and creating vector embeddings for complex retrieval tasks.

Product Manager

They use this to prototype features that require multiple inputs and outputs—like turning user text into a video mockup or an audio ad copy.

What Changes When You Connect

01 You don't worry about infrastructure. By connecting this MCP, you get immediate access to over 100 open-source models for text, image, and audio tasks.

02 When performance matters, you create dedicated endpoints using `create_endpoint`, ensuring your app never slows down due to model throttling.

03 Need custom intelligence? You upload data and use `create_fine_tune` to train a specialized model on your unique business vocabulary.

04 Build advanced search. Instead of simple keyword matching, you generate embeddings with `create_embeddings` and then refine results using `create_rerank` for better accuracy.

05 Handle large-scale workloads easily. Use the batch tools (`create_batch`, `list_batches`) to process thousands of items asynchronously without timing out your agent.

Real-World Applications

Building a Podcast Summary Generator

A user uploads an hour-long interview audio file. The agent first runs `create_audio_transcription` to get the text transcript, then uses `create_chat_completion` on that text to draft five key bullet points, and finally sends those points via a messaging tool.

Creating Marketing Assets for a Product Launch

The product team inputs a core feature description. The agent uses `create_image_generation` to generate several visual concepts and then runs `create_video_generation` on the best image, all within one workflow.

Implementing Advanced Internal Knowledge Search

Instead of just searching a database, the agent takes user questions, uses `create_embeddings` to convert them and the documents into vectors, and then runs `create_rerank` to pull back the absolute most relevant internal policy document.

Automating Customer Service Voice Guides

The system takes a support article written by an expert. It uses `create_audio_speech` to convert that text into a professional voice guide, ready for immediate deployment.

Patterns to Avoid

Using generic LLM APIs

✗ AVOID

The developer just calls the basic chat completion tool and assumes it has enough context or specialized knowledge for complex tasks like audio analysis.

✓ INSTEAD

For specific, multi-modal needs, don't rely on a single endpoint. You must use `create_audio_transcription` first to extract text, then feed that structured data into `create_chat_completion`.

Running tasks synchronously

✗ AVOID

The developer tries to process 500 documents in a single API call because it's faster to code.

✓ INSTEAD

For anything over a few dozen items, you have to use the batch system. Start by calling `create_batch` and then monitor progress using `get_batch`.

Ignoring performance needs

✗ AVOID

The application fails or slows down during peak usage hours because it's relying on shared, general-purpose resources.

✓ INSTEAD

Set up stability first. Use `create_endpoint` to secure a dedicated resource for your critical model calls, guaranteeing predictable speed.

The Right Fit

Use this MCP if your application needs to handle multiple types of AI output—for instance, generating an image *and* writing the accompanying alt text, or transcribing audio *and* summarizing it. If you're building a system that requires specialized data handling like embedding generation or fine-tuning on private documents, you need its advanced model operations. Don't use this if your only goal is simple API calls to a single general chat model; in those cases, a simpler text completion tool might suffice. But if the complexity involves media (audio/video), structured knowledge retrieval (`create_embeddings`), or reliable scaling (`create_endpoint`), then you need the power of this entire catalog.

The headache of piecing together AI features manually.

Today, building a single feature that needs to do three things—like reading an audio file, summarizing it, and then generating promotional art—is a nightmare. You're jumping between the transcription tool, the chat API, and the image generation platform. You copy text from one dashboard into another service, manage keys for multiple providers, and spend hours just stitching the workflow together.

With this MCP, your agent handles the whole sequence inside one connection point. It takes the audio input, runs `create_audio_transcription`, passes that output to generate a summary via chat completion, and finally feeds keywords into `create_image_generation`. You get a fully functional feature without ever leaving your client.

Generating Media with Dedicated Model Operations

The biggest manual step that disappears is the juggling act between different model APIs. You used to have separate documentation and setup steps just for generating an image versus generating a video, forcing you into complex multi-step code blocks.

Now, if your workflow needs visual content, whether it's basic text prompts or full motion video, you call `create_image_generation` or `create_video_generation`. The whole process is contained and controllable from one place.

Together AI: A Powerful Toolset With 27 Tools

These tools let you manage model lifecycle, generate media, process voice and text data, and run large background jobs all through one connection.

#	TOOL	DESCRIPTION
01	create_audio_speech	This tool generates speech from plain text, creating voiceovers for your content.
02	create_audio_transcription	It converts an uploaded audio file into a written transcript using speech-to-text technology.
03	cancel_batch	You can stop any large, running background processing job immediately.
04	create_chat_completion	This tool generates model responses by simulating a full back-and-forth chat conversation.
05	create_batch	It starts a new, large-scale asynchronous job that runs in the background over time.
06	create_endpoint	You can set up a dedicated connection point to ensure your model performance never drops or slows down.
07	create_fine_tune	This initiates the process of training an open-source model on your specific, proprietary dataset.
08	delete_endpoint	It removes a dedicated connection point you previously set up for performance stability.
09	delete_file	This permanently deletes an uploaded file used for training or batch processing.
10	delete_fine_tune	You can cancel a fine-tuning job that you started and no longer need.
11	create_embeddings	It takes any block of text and converts it into numerical vector embeddings for search indexing.
12	get_batch	You can check the current status and results of a specific background job.
13	get_endpoint	This retrieves all the details about a dedicated model endpoint you created.

#	TOOL	DESCRIPTION
14	get_file	It fetches metadata and information about an uploaded file without needing to download it.
15	get_fine_tune	You get the current status and progress report for a specific fine-tuning job.
16	create_image_generation	This tool generates brand new images based on detailed text descriptions or prompts.
17	list_batches	You see a list of all background jobs that have been created using the system.
18	list_endpoints	It lists every dedicated model endpoint currently running or configured for your account.
19	list_files	You get a list of all data files you've uploaded to the system.
20	list_fine_tune_checkpoints	This lists saved versions, or checkpoints, for a fine-tuning job so you can revert if needed.
21	list_fine_tunes	It gives you an overview of all the fine-tuning jobs that have been run previously.
22	list_models	You can see a list of every model available for use through this MCP connection.
23	create_rerank	This tool reorders documents based on how relevant they are to the user's specific query.
24	create_text_completion	It generates extended text content for a simple prompt, ideal for articles or summaries.
25	update_endpoint	You can change the status—like scaling up or down—of an existing dedicated model endpoint.
26	upload_file	It securely uploads a file for use in fine-tuning, evaluation, or batch processing tasks.
27	create_video_generation	This tool creates entire videos from text prompts or by animating an existing image.

See It in Action

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

U Generate a chat completion using meta-llama/Llama-3.3-70B-Instruct-Turbo explaining quantum computing.

I've initiated the request to Llama-3.3-70B. Quantum computing uses qubits to perform calculations that are impossible for classical computers by leveraging superposition and entanglement...

U Create an image of a futuristic laboratory using the black-forest-labs/FLUX.1-schnell model.

Generating image... I've successfully created the image of a futuristic laboratory. You can access the result at the provided URL or as a base64 string.

U List all available models on Together AI.

Fetching model list... You have access to over 100 models, including Llama-3.3-70B, Mixtral-8×7B, Flux.1, and various embedding models. Would you like to filter by type?

Frequently Asked Questions

01 **How do I use the Together AI MCP for document search?**

You run this by first calling `create_embeddings` on your documents to turn them into vectors. Then, when a user asks a question, you use `create_rerank` to find the most relevant chunks of text from those stored embeddings.

02 **Can I make my AI model better using this MCP?**

Yes. You manage custom training jobs by calling `upload_file` and then initiating a job with `create_fine_tune`. This allows you to teach the open-source models your company's specific jargon.

03 **What is the difference between `create_chat_completion` and `create_text_completion`?**

Use `create_chat_completion` when you need the model to remember context from a conversation history. Use `create_text_completion` for single, self-contained text generation tasks like writing an article summary.

04 **Does this MCP help with large data uploads?**

It handles massive jobs using the batch tools. You start a job via `create_batch`, and then you monitor its progress and retrieve results later using `get_batch`.

05 **How do I ensure my model stays fast for production?**

You use `create_endpoint`. This tool establishes a dedicated, stable connection point that isolates your usage from general traffic fluctuations, guaranteeing reliable performance.

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-alternative": { "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	`019e38fc-f902-730c-94c9-64868c3fd057`
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-alternative.