MCP SERVER | NO CODE | CLOUD HOSTED

LocalAI MCP

Run Multimodal AI on Your Hardware.

LocalAI lets you run powerful AI models—including text chat, image generation, audio transcription, and face analysis—entirely on your own hardware. It provides a standard API endpoint compatible with OpenAI and Anthropic protocols, letting any client connect to private local models without sending sensitive data to the cloud.

A+ Quality Score 100/100

self-hosted | llm-inference | image-generation | audio-processing | openai-compatible | local-models

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

SCAN TO VIEW

The infrastructure that powers AI agents in the real world.

Vinkius connects AI to the world's software through secure, enterprise-grade infrastructure — enabling real-world execution at scale, built on the Model Context Protocol (MCP).

Your AI Connections Run Through Vinkius Cloud

The world's largest managed MCP catalog

Vinkius is the cloud infrastructure where AI agents connect 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.

LocalAI MCP

19 tools available
Cloud-hosted on Vinkius

This MCP lets you bring advanced artificial intelligence capabilities right into your local environment. Instead of relying on third-party services for every single task, you can run powerful multimodal models directly from your own infrastructure. This means keeping all your sensitive data private while still accessing top-tier AI performance.

Whether you need to generate complex images from text prompts, convert recorded speech into searchable text, or analyze faces for identity verification, this connector handles it locally. You connect your preferred agent through Vinkius and gain access to a comprehensive set of tools that span everything from basic chat completions using `chat_completions` to advanced functions like generating vector embeddings with `create_embeddings`. It's about giving you full control over where the AI processing happens, ensuring speed and privacy are always priorities.

Core Capabilities

01 — Run Chat and Text Generation

You generate text responses for chat or completions using local language models that support both OpenAI and Anthropic standards.

02 — Create Visual Media

You prompt the system to synthesize unique images from scratch, even allowing you to define negative prompts to exclude unwanted elements.

03 — Process Audio Files

You convert spoken audio into written text using transcription or generate natural-sounding speech files from plain text.

04 — Identify and Analyze Faces

You verify a person's identity by comparing faces one-to-one, enroll new individuals, or detect objects within an image for analysis.

05 — Improve Data Retrieval

You generate vector embeddings to index text and use those vectors to improve search results based on a specific query.

One Click on Vinkius — From Prompt to Execution

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

01 Subscribe to this MCP, providing your LocalAI Base URL (e.g., http://localhost:8080) and an optional API Key.

02 Your AI client connects using the provided credentials, establishing a secure link to your local models.

03 You interact with the system through your agent, triggering actions like text generation or image synthesis as if it were any other online service.

The bottom line is that you treat your private, locally hosted AI instance exactly like a cloud API endpoint from anywhere in the Vinkius catalog.

Built For

This MCP is for developers and researchers who cannot send proprietary or sensitive data to third-party servers. It's perfect for internal tools, compliance departments, and anyone building complex AI pipelines that demand absolute privacy.

Privacy-Conscious Developer

You build client-facing applications handling personal data (like biometric information or private communications) and need to ensure the LLM processing never leaves the local machine.

ML Researcher

You test out multiple open-source models for chat, vision, or audio tasks. This MCP lets you easily swap between different local model versions without changing your core code.

DevOps Engineer

You integrate AI capabilities into internal automation pipelines, needing a stable, self-hosted endpoint that doesn't rely on external cloud service uptime or cost limits.

What Changes When You Connect

01 Data Privacy: By running everything locally, you eliminate the risk of sending proprietary or sensitive data to any third-party cloud vendor. This is non-negotiable for compliance and internal tools.

02 Control Over Models: You maintain full control over which AI model runs your workflows. Need to test a new open-source LLM? Just apply it locally with `apply_model` and start using it immediately.

03 Full Media Pipeline: This MCP covers the whole stack. Generate images with `generate_image`, transcribe audio with `transcribe_audio`, and then convert summaries back into voice using `text_to_speech`—all without an internet dependency.

04 Advanced Search: Go beyond basic keyword searches. Use `create_embeddings` to index your documents, and then use `rerank_documents` to guarantee the most contextually relevant answers for RAG workflows.

05 Biometric Capabilities: Handle identity management securely. You can run specific tools like `face_register` or `face_verify` to process sensitive biometric data entirely on private hardware.

Real-World Applications

Compliance Auditing for Biometrics

An HR department needs a tool that verifies employee identities using photos taken at different sites. Instead of sending images offsite, they connect the MCP and use `face_verify` to perform 1:1 biometric checks entirely within their private network.

Creating Localized Marketing Assets

A marketing team needs dozens of unique product mockups for a campaign. They send a text description to the agent, which then uses `generate_image` to output high-res visuals without incurring massive cloud API costs.

Building Internal Call Summaries

A sales team records client calls on internal VoIP systems. They connect the MCP and use `transcribe_audio` immediately, then pass the resulting text to `chat_completions` to generate structured follow-up summaries for CRM entry.

Improving Knowledge Base Search

A legal firm has thousands of documents. Instead of just searching by keyword, they use `create_embeddings` across their entire corpus and then employ `rerank_documents` to ensure the agent retrieves the single most contextually relevant passage for a query.

Patterns to Avoid

Using it only for basic chat

✗ AVOID

Thinking that since you can use `chat_completions`, you don't need to worry about data privacy. You might send your company's most sensitive documents through a general-purpose endpoint.

✓ INSTEAD

If the primary concern is just chatting, ensure the connection is local via this MCP. But remember, for anything involving media or biometrics, you must use dedicated tools like `detect_objects` and `face_verify` to keep the process contained.

Ignoring audio source requirements

✗ AVOID

Attempting to process a live microphone stream directly through the API endpoint. The system expects files or paths, not continuous streams.

✓ INSTEAD

For accurate speech processing, you must first capture and save the audio data (a file or path), then pass that specific file reference to `transcribe_audio`.

Thinking it replaces all APIs

✗ AVOID

Assuming this MCP can handle every single API call your organization uses, even those outside of AI, like database lookups or email sending.

✓ INSTEAD

This MCP is specifically for running LLM and media tools locally. For actions outside the scope of text, image, audio, or face analysis, you'll need a different integration.

The Right Fit

Use this if your primary requirement is data sovereignty—if sending data to a third-party cloud provider violates privacy rules or costs too much. This MCP gives you the power of multimodal AI while keeping the processing local. Don't use it if you simply need a quick, one-off test using a publicly available online demo; for those, a

simple public endpoint might be faster. However, if your workflow involves biometrics (`face_verify`), generating high volumes of media (`generate_image`), or processing sensitive audio, this local solution is mandatory. If your job only requires basic text completion without needing to reference private documents, you might just use a standard chat client, but for anything involving data indexing, go with the `create_embeddings` and `rerank_documents` tools here.

Manual media pipelines are slow and expensive.

Today, generating marketing assets means passing text through a web form, downloading an image file, checking the resolution on Photoshop, writing a summary in Notion, and then uploading that document to your shared drive. It's click-by-click, manual copy-pasting that eats up hours of labor every week.

With this MCP, you simply tell your agent what you need—say, 'Generate five images of a futuristic library.' The system handles the generation using `generate_image` , and then it can automatically summarize the findings for your internal wiki. You get results in one controlled flow, without leaving your private network.

Get LocalAI's multimodal power with chat_completions

The biggest time sinks are the data transfers: recording a meeting, uploading it to a service, waiting for transcription, downloading the text file, and then pasting that text into another tool for summarization. It's a chain of manual handoffs.

Now, you pass the audio directly through the MCP using `transcribe_audio` , and your agent gets the clean text instantly. You can feed that output immediately to `chat_completions` for summarizing or even use it in `create_embeddings` for instant indexing. The whole process runs as one continuous, private operation.

LocalAI: 20 Tools for Local AI Inference

These tools allow your agent to perform everything from generating chat responses and creating images to analyzing faces and transcribing audio, all using models running on your private hardware.

#	TOOL	DESCRIPTION
01	anthropic_messages	Generates multi-turn chat messages using local models compatible with Anthropic's API structure.
02	apply_model	Installs a new AI language or media model from the available gallery.
03	chat_completions	Generates conversational text responses using local models compatible with OpenAI's API structure.
04	create_embeddings	Converts blocks of text into numerical vector embeddings for advanced search and indexing.
05	detect_objects	Scans an image and returns a list of identified objects along with their locations.
06	face_analyze	Provides demographic or characteristic analysis on human faces found in images.
07	face_identify	Compares a face to previously registered individuals to determine who the person is (1:N comparison).
08	face_register	Enrolls and securely stores a new individual's facial data for future identification.
09	face_verify	Confirms if an unknown face matches a known identity by comparing it one-to-one.
10	generate_image	Creates entirely new visual content based on your text prompts, supporting negative prompts to filter out undesirable elements.
11	get_auth_status	Checks the current authentication status and lists available identity providers.
12	get_auth_usage	Displays usage metrics for personal API tokens or access keys.
13	get_system_info	Retrieves general operational details and backend information about the local AI instance.
14	get_version	Returns the specific version number of the LocalAI software running on the infrastructure.

#	TOOL	DESCRIPTION
15	`list_models`	Retrieves a list of all AI models that are currently installed and ready for use by your agent.
16	`open_responses`	Generates open-ended, unstructured text responses when specific chat protocols aren't required.
17	`rerank_documents`	Refines search results by reordering documents based on how closely they relate to your specific query.
18	`text_to_speech`	Converts plain text into an audio file using high-quality synthetic voice generation (TTS).
19	`transcribe_audio`	Transcribes recorded speech files or paths, converting the spoken word back into editable text.

See It in Action

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

U List all models available on my LocalAI instance.

I've retrieved the list of models. You have 'llama-3-8b', 'stablediffusion', and 'whisper-1' currently active and ready for use.

U Generate a chat response using the 'llama-3' model about the benefits of local AI.

Using the 'llama-3' model: Local AI offers significant benefits including enhanced data privacy, reduced latency, and the ability to operate without an internet connection...

U Create an image of a futuristic library using the 'stablediffusion' model.

I've initiated the image generation for a 'futuristic library'. The process is complete, and you can view the generated image at the provided local URL.

Frequently Asked Questions

01 **How do I start using LocalAI with chat_completions?**

You first connect your client to this MCP and ensure you have a local LLM installed via `apply_model`. Then, your agent can call the `chat_completions` tool just like it would any other API.

02 **Can I run image generation if my data needs to stay private?**

Yes. By using the MCP, you leverage local models for media creation. You simply call `generate_image`, and the visual content is processed entirely on your own hardware.

03 **What's the difference between face_identify and face_verify?**

Face verification (`face_verify`) confirms if a single unknown face matches a known person (1:1). Face identification (`face_identify`) determines who a person is by comparing their face against many registered identities (1:N).

04 **Does LocalAI help me search my documents better?**

Absolutely. Instead of basic keyword searches, you use `create_embeddings` to build searchable vectors from your documents and then use `rerank_documents` to improve the relevance of retrieved results.

05 **How do I make sure my audio files are processed correctly?**

You must first pass the file path or raw data through the `transcribe_audio` tool. This converts the speech into text, which you can then use with any of the other chat tools.

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 `"localai": { "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

LocalAI 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 LocalAI. 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	LocalAI MCP
Server ID	`019e38ba-2e24-73ee-8a88-40849fef4982`
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/localai.