MCP SERVER | NO CODE | CLOUD HOSTED

Gladia Speech AI MCP

Turn any spoken word into structured text data.

Gladia Speech AI provides enterprise-grade speech recognition and analysis, turning any audio or video stream into actionable data. This MCP handles everything from basic transcription to complex tasks like speaker diarization, multi-language translation across 100+ languages, and applying custom large language model prompts directly to the spoken content. It supports processing pre-recorded files via uploads and managing secure WebSocket connections for real-time live streaming.

A+ Quality Score 100/100

speech-to-text | transcription | audio-analysis | speaker-diarization | translation | natural-language-processing

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

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.

Gladia (Speech AI) MCP

6 tools available
Cloud-hosted on Vinkius

You can feed any audio source—a podcast recording, a length meeting call, or even a live broadcast—into this MCP and get structured text back out. Forget listening to hours of raw audio just to find three action items; your agent handles the heavy lifting. It doesn't just transcribe what was said; it figures out who spoke each line, translates segments into dozens of languages, and can summarize the entire discussion based on specific prompts you give it. When you connect this MCP through Vinkius, your AI client treats it like a natural extension of conversation. Instead of juggling separate services for file uploads, job status checks, and final analysis, you ask one question, and the system executes the entire workflow, delivering clean, ready-to-use text data.

Core Capabilities

01 — Process uploaded audio files

Upload an audio file to start a secure job that transcribes and analyzes the spoken content.

02 — Manage live streaming sessions

Initialize continuous, real-time transcription streams for ongoing meetings or broadcasts over WebSocket connections.

03 — Extract specific data from audio

Apply custom prompts to the transcribed text to pull out structured insights, like names, dates, and action items.

04 — Handle job status tracking

Check the progress or retrieve the final results of any transcription job you've started.

One Click on Vinkius — From Prompt to Execution

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

01 Subscribe to this MCP and enter your Gladia API key.

02 Instruct your AI client to either upload a file for batch processing or start an initial live session link.

03 The system runs the job, and you retrieve the status and final text results directly through conversation.

The bottom line is that it converts raw, unstructured audio into clean, actionable data points without you needing to manage complex API calls.

Built For

This MCP is for content creators and knowledge workers who regularly deal with high volumes of spoken word. If your job requires turning recordings or live conversations into searchable text, this tool saves hours of manual cleanup.

Podcast Editor

Upload final episode audio files to automatically generate transcripts and summaries for show notes.

Business Analyst

Run meeting recordings through the system to identify key decisions, action items, and who was responsible for them.

Technical Content Writer

Transcribe technical interviews or demos live, then use the generated text to draft articles piece by piece.

What Changes When You Connect

01 Instead of manually transcribing hours of video, simply use `init_transcription` to upload a file and get the full text transcript ready for editing. The system handles speaker diarization automatically.

02 For real-time work, you can initialize secure WebSocket connections using `init_live_session`. This means your agent transcribes meetings as they happen, eliminating transcription lag.

03 The analysis goes way beyond simple spelling out words. You apply custom LLM prompts to the audio data to extract specific insights or structure unstructured notes into JSON format.

04 If you need to know what jobs are running or finished, use `list_transcriptions` to pull up a history of all your work in one query. Then, check the results with `get_transcription`.

05 The multi-language support is massive; you can initiate transcription and translation across over 100 languages, making global content creation straightforward.

Real-World Applications

Cleaning up a recorded client interview

A marketing manager needs to analyze an hour-long Zoom call. Instead of manually listening for key quotes, they ask their agent to use `upload_audio_file` and run the transcription with diarization. The resulting text immediately tells them which speaker said what, making follow-up action items easy.

Covering a live panel discussion

A journalist needs real-time notes from a conference panel. They connect their agent to the MCP using `init_live_session`. The transcript streams in instantly, allowing them to capture quotes and speaker shifts without missing a beat.

Processing international podcast archives

A global content team has recorded interviews in six different languages. They use the MCP's advanced transcription features to upload files, enabling simultaneous translation and summarization for all regional markets.

Debugging a failed audio job

An engineer uploaded an audio file but isn't sure if the job finished correctly. They use `list_transcriptions` first to find the Job ID, then call `get_transcription` to confirm the status and retrieve any error logs.

Patterns to Avoid

Trying to transcribe audio via manual API calls

✗ AVOID

A user tries to manually manage file URLs, job IDs, and multiple endpoints just to start a transcription. This is slow, brittle, and requires writing complex code for simple tasks.

✓ INSTEAD

The right way is to let your agent use the MCP's tools. First, call `upload_audio_file` to get it into the system, then tell your agent to run `init_transcription`. The conversation handles the complexity.

Using generic text analysis tools on audio

✗ AVOID

A user uploads an MP3 file but uses a basic tool that only generates plain, unformatted text without speaker separation or timestamps.

✓ INSTEAD

This MCP provides advanced features. Start by running `init_transcription` and ensure you prompt for 'speaker diarization' to get structured text showing exactly who said what.

Forgetting job status checks

✗ AVOID

A user initiates a long transcription but forgets to check on it, assuming the results are ready instantly. This leads to wasted time and failure points.

✓ INSTEAD

Always follow up after starting a job by calling `get_transcription` with the Job ID. This confirms if the process is running or if it's finished and ready for review.

The Right Fit

Use this MCP if your primary input data is audio (meetings, podcasts, live feeds) but your desired output is highly structured, searchable text. You need more than just a transcript; you need analysis, translation, or specific insights extracted using custom prompts.

Don't use it if all you need to do is store the raw audio file somewhere else, or if you are already generating transcripts in-house and only need basic storage. If you only need simple data extraction from pre-written text documents (PDFs, Word files), look for a document processing MCP instead. This tool excels at turning sound into intelligence.

The Messy Reality of Handling Spoken Content

Right now, if you get an audio recording—say, a client interview or team meeting—you have to do a painful loop. You download the file, upload it somewhere, maybe use one service for transcription and another separate tool for summarization. Then, you copy-paste the text into a third place just to identify key action items. It's fragmented, it takes hours, and every step risks losing data or context.

With this MCP, that whole process collapses into a single conversation. You feed your agent the audio file, tell it what you need—a summary of decisions made, for example—and it handles the entire pipeline: transcription, diarization, summarization, all in one go. What you get is clean, structured text ready to paste directly into an email or report.

Get Insights with Gladia Speech AI MCP

You eliminate the need for manual transcription cleanup and separate analysis tools. You don't have to wait days for a human transcriber; you initiate the job, check its status using `get_transcription`, and retrieve structured text in minutes.

What's different now is that your agent understands the context of the audio. It doesn't just write out words; it analyzes speaker roles, translates languages on demand, and structures the output according to your exact prompts.

Gladia Speech AI (Speech AI) MCP with 6 Tools

These tools let your agent manage the entire audio lifecycle: from uploading files to initiating live sessions, checking status, and deleting old jobs.

#	TOOL	DESCRIPTION
01	`delete_transcription`	Removes a specific transcription job from your Gladia account.
02	`upload_audio_file`	Transfers an audio file to the platform so you can begin processing it.
03	`get_transcription`	Checks the current status and retrieves the final text results for a known job ID.
04	`list_transcriptions`	Retrieves a list of all previously run, pre-recorded transcription jobs.
05	`init_live_session`	Starts and maintains a secure link for real-time transcription during live broadcasts or meetings.
06	`init_transcription`	Begins the processing job for an uploaded audio file to generate a transcript.

See It in Action

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

U List my 5 most recent transcription jobs.

I've retrieved your recent jobs. You have 5 tasks: 'Meeting_Notes.mp3' (Done), 'Interview_01.wav' (Done), and 3 others. Would you like the results for any of these?

U Start a transcription for this audio URL with summarization enabled: https://example.com/audio.mp3

Transcription job initiated! The Job ID is `job_12345`. I've enabled summarization as requested. I'll monitor the status for you.

U I need a WebSocket URL to start a live transcription session in 16000Hz.

I've generated a live session. Here is your secure WebSocket URL: `wss://api.gladia.io/v2/live/...`. The sample rate is set to 16000Hz.

Frequently Asked Questions

01 **How do I transcribe a live meeting with Gladia Speech AI MCP?**

You initiate a real-time session by calling `init_live_session`. This creates a secure WebSocket link that streams the transcription output to your agent as the meeting happens.

02 **Can I translate audio using Gladia Speech AI MCP?**

Yes. The MCP supports multi-language translation. You can run a job and specify both the source language and the target language for the output text.

03 What is speaker diarization with Gladia Speech AI MCP?

Speaker diarization identifies who spoke what during the audio session. The resulting transcript will tag lines to specific speakers, making it easy to track contributions in a meeting.

04 How do I check if my transcription job finished using Gladia Speech AI MCP?

After starting a job with `init_transcription`, you use the `get_transcription` tool, providing the Job ID. This will tell you the status and provide the final results when ready.

05 Does Gladia Speech AI MCP support video files?

While it processes audio content, you must extract the audio stream first. The MCP is designed to handle the resulting audio files for transcription and analysis.

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 `"gladia-speech-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

Gladia (Speech 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 Gladia (Speech 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	Gladia (Speech AI) MCP
Server ID	`019e389f-d25d-7181-8e9c-7853ea348e91`
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/gladia-speech-ai.