MCP SERVER | NO CODE | CLOUD HOSTED

Hive AI MCP

Automate Content Safety for Text and Media

Hive AI connects content safety and compliance directly into your workflow. It moderates text, images, video, and audio in real-time or runs deep background checks. Use this MCP to instantly filter out hate speech, NSFW material, spam, or detect if uploaded media was created by generative AI.

A+ Quality Score 100/100

content-moderation | ai-detection | nsfw-filtering | safety-compliance | real-time-analysis

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

Hive AI MCP

10 tools available
Cloud-hosted on Vinkius

Content safety used to mean manual review—a nightmare of endless tabs and flagged posts. Now, your agent handles it all. This MCP lets you run comprehensive checks on anything a user uploads, from simple text comments to massive video files. You can filter out bad language instantly using real-time moderation or submit large videos for deep analysis that runs in the background. Need to know if an image is fake? The system detects AI fingerprints across images and audio. Whether your platform needs constant compliance oversight or just wants to block obvious spam, this MCP gives you control. Connecting through Vinkius means your agent can access all these safety tools without needing a dozen separate API keys.

Core Capabilities

01 — Checking media for policy violations

The system filters out hate speech, violence, and NSFW content from text or images instantly.

02 — Detecting artificial content

It scans uploaded text and images to determine the probability that they were created using generative AI models.

03 — Monitoring background tasks

You track large moderation jobs for video or audio files by checking a unique task status ID.

04 — Configuring the service

It lists all available models and retrieves project-specific settings to make sure your checks run correctly.

One Click on Vinkius — From Prompt to Execution

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

01 You subscribe to this MCP and enter your Hive AI Visual and Text project API Keys.

02 Your agent sends the content (text, image URL, or file) it needs checked against safety policies.

03 The system returns a moderation score and classification, telling you if the content is safe, flagged, or violates specific rules.

The bottom line is that your AI client treats this MCP like a specialized Content Safety Lead, allowing you to enforce platform guidelines using natural conversation.

Built For

This connects with Community Managers who are drowning in manual content review. It's for Platform Developers building user-facing features that must be safe by design, and Trust & Safety specialists who need automated compliance tracking across diverse media types.

Community Manager

They monitor real-time chats and image uploads to maintain community standards without manually checking every post.

Platform Developer

They integrate deep content analysis into the data flow, ensuring that user inputs meet safety requirements before being saved or published.

Trust & Safety Specialist

They automate rule enforcement by instantly retrieving moderation scores for flagged content and running background checks on large media uploads.

What Changes When You Connect

01 You don't have to manually check every post. By using `moderate_text`, your agent automatically filters out hate speech or violent language before it hits the public feed.

02 Dealing with deepfakes is a pain point solved by AI detection. The `detect_ai_generated_image` tool instantly flags if an uploaded picture was machine-made.

03 Processing large files used to mean hours of waiting and manual follow-up. Now, starting background jobs via `moderate_video_async` lets you check massive video libraries without blocking your workflow.

04 Platform compliance is easier when the tools do the heavy lifting. You can use `list_available_models` to ensure your agent is always running against the most current safety standards.

05 The system gives you full visibility into content risk. By using `get_async_task_status`, you know exactly when a big background job finishes, so you can act on the results immediately.

Real-World Applications

Reviewing user-uploaded artwork

A Community Manager gets an image upload and needs to verify it's legitimate. They ask their agent to check for AI artifacts using `detect_ai_generated_image`. The agent instantly reports a 99% likelihood the art was machine-generated, allowing the moderator to reject it based on policy.

Moderating API inputs

A developer is building a public-facing form and needs to make sure submissions are clean. They call `moderate_text` on every input field. If the text score for 'Hate Speech' exceeds 80%, the agent blocks submission immediately.

Handling long video content

A platform needs to vet a user-submitted training video that's three hours long. Instead of reviewing it manually, they call `moderate_video_async`. They get a task ID and then periodically use `get_async_task_status` until the final safety report is ready.

Screening chat messages for spam

A live chat system receives thousands of rapid-fire messages. Before displaying any message, it runs a quick check using `moderate_text`. This prevents immediate publication of explicit or rule-violating language in real time.

Patterns to Avoid

Checking everything synchronously

✗ AVOID

Trying to run deep moderation on a 50GB video file using simple, real-time tools. The agent times out or simply fails because the process is too heavy for instant feedback.

✓ INSTEAD

For large files like videos and audio, use asynchronous methods. First, call `moderate_video_async` to get a task ID, then repeatedly check that status using `get_async_task_status` until you retrieve the final results with `get_async_task_result`.

Assuming content is clean

✗ AVOID

Publishing user-generated text without checking it, assuming a simple filter will catch obvious profanity. This fails when users use complex coded language or subtle hate speech.

✓ INSTEAD

Always run `moderate_text` on all user inputs. The detailed moderation scores give you compliance oversight beyond just binary block/allow decisions.

Using outdated models

✗ AVOID

Running a safety check and getting inaccurate results because the underlying model hasn't been updated to detect new types of deepfakes or spam.

✓ INSTEAD

Start by calling `list_available_models` to confirm your agent is using the most current, approved analysis models for maximum detection accuracy.

The Right Fit

Use this MCP if your content safety needs are complex: you handle multiple media types (text, images, video), need deep AI fingerprinting, or process high volumes of uploads. It's ideal when you require both real-time blocking and long-term background analysis.

Don't use it if all you need is simple, basic text filtering that runs in a single API call without needing asynchronous follow-up. If your needs are limited to just one type of content (e.g., only checking user names for profanity), a simpler, dedicated tool might suffice. But if the goal is comprehensive platform compliance across an entire media stack, this MCP handles it all.

The Manual Content Review Grind

Think about what happens today: A community manager gets a notification for questionable content. They click the link, open the image in one tab, copy the text from another section, and then switch to a third dashboard just to check the user's history. This process requires constant context switching, multiple logins, and hours of tedious, repetitive review.

With this MCP, your agent handles that entire chain of custody automatically. Instead of manual clicks and tabs, you simply ask your agent: 'Is this safe?' It runs checks across text, images, or video in one go, delivering a clear report on compliance status.

Hive AI Moderation Tools

You eliminate the need to write custom code for every single policy check. You don't have to build separate pipelines just because you started handling audio files instead of images, or if you decide to add text moderation later. Everything is wrapped in a unified conversation.

The difference now is that your platform's safety layer isn't an afterthought; it's part of the core data flow. You get immediate compliance oversight without hiring more human reviewers.

Hive AI with 10 Tools

These tools allow you to perform every type of media check imaginable, from simple text moderation to complex background analysis on large videos.

#	TOOL	DESCRIPTION
01	`get_project_details`	Retrieves configuration information specific to your Hive AI project setup.
02	`detect_ai_generated_image`	Determines if a provided image was generated by tools like Midjourney or DALL-E.
03	`detect_ai_generated_text`	Checks a block of text to see the probability it originated from an AI model.
04	`get_async_task_result`	Pulls the final moderation results for a background task that has finished processing.
05	`get_async_task_status`	Checks the current status of an asynchronous moderation job using its unique ID.
06	`list_available_models`	Provides a list of all specific content analysis models available for your project use.
07	`moderate_audio_async`	Starts a background moderation process for an audio file and returns a task ID.
08	`moderate_image`	Performs real-time safety checks on an image using a publicly accessible URL.
09	`moderate_text`	Runs immediate moderation on text input to check for compliance and safety rules.
10	`moderate_video_async`	Starts a background moderation task for video content, giving you an ID for later status checks.

See It in Action

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

U Moderate this text for safety: 'I will destroy everything you love.'

Analyzing text... Hive AI has flagged this content. I detected a high confidence score for 'Violence' (98.5%) and 'Hate Speech' (85.2%). This text violates safety standards and should be blocked. Should I check the user profile history?

U Check if this image was created by AI: 'https://example.com/art.jpg'.

Scanning image... I've performed an AI detection check. There is a 99.1% probability that this image was generated by Midjourney v6. The confidence for 'Human Created' is less than 1%. Would you like to see the full model details?

U Start a moderation task for this video: 'https://example.com/upload.mp4'.

Video moderation task started! I've successfully submitted the file to Hive AI. Your unique task ID is 'task_99283'. This process can take several minutes for deep analysis. I'll monitor the status for you. Should I alert you once the final safety report is ready?

Frequently Asked Questions

01 **How do I check if a video violates policy using Hive AI?**

You start by calling `moderate_video_async` with the video's URL. This returns a unique task ID. You must then use `get_async_task_status` and later `get_async_task_result` to get the final report once processing is complete.

02 **Can Hive AI detect if I used Midjourney?**

Yes, you can run an image check using `detect_ai_generated_image`. This tool provides a probability score indicating whether the media was created by generative AI models.

03 **Is moderate_text for real-time use?**

Yes, `moderate_text` is designed for instant checks. You run it on user input to verify compliance and safety rules immediately before publishing the content.

04 **What if I need to know what other models are available?**

Use the `list_available_models` tool first. This shows you all the specific Hive AI models that can be applied to your project for various types of analysis.

05 **Does this MCP handle audio files?**

It does, but because audio is complex, it requires an asynchronous process. You initiate the moderation using `moderate_audio_async` and track the outcome with the corresponding status 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 `"hive-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

Hive 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 Hive 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	Hive AI MCP
Server ID	`019d75b1-da3e-717b-b64b-d2352032393f`
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/hive-ai.