MCP SERVER | NO CODE | CLOUD HOSTED

Mistral AI MCP

Run sophisticated LLM tasks with simple conversation.

Mistral AI connects your agent to European LLMs for complex tasks like chat completions and content moderation. You can generate vector embeddings for semantic search, process massive data sets with batch jobs, or check user-generated text safety before it hits production. Use this MCP when you need reliable access to Mistral's models without switching APIs.

A+ Quality Score 100/100

llm | natural-language-processing | embeddings | model-inference | ai-agents | batch-processing

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

Mistral AI MCP

10 tools available
Cloud-hosted on Vinkius

Mistral AI lets your agent talk to powerful European language models directly through conversation. Instead of writing complex API calls every time, you just tell your agent what you want—like drafting a response or checking text for safety. The MCP handles the rest. Need to index thousands of documents? You can set up batch jobs to process them asynchronously and track their progress until they're done. For data retrieval, simply generate vector embeddings; this turns raw text into searchable numerical representations perfect for any custom knowledge base. If you're building a complex system, Vinkius makes it easy: you connect your preferred agent client once and get access to Mistral's full suite of tools right in the chat window.

Core Capabilities

01 — Chat with various models

Send conversations to different Mistral model sizes, from highly capable large models to efficient small ones, receiving formatted responses directly.

02 — Generate vector embeddings

Convert chunks of text into numerical vectors suitable for semantic search and similarity comparisons in a database.

03 — Moderate content safety

Check any text input against predefined categories, returning specific safety scores to flag dangerous or inappropriate material.

04 — Manage large batch processing

Create and track jobs that process huge volumes of data over time, letting you run compute-intensive tasks without timing out.

05 — Discover available models

List all Mistral AI models and their specific IDs, capabilities, and context window sizes so you know which one to use for the job.

One Click on Vinkius — From Prompt to Execution

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

01 Subscribe to this MCP in Vinkius and enter your unique Mistral API key.

02 Your AI client uses the stored credentials to authenticate requests when you call a function, like generating embeddings or starting a chat.

03 The MCP sends the structured request to Mistral AI, receives the processed data, and relays the final output back to your agent conversation.

The bottom line is that it turns complex API interactions into simple conversational commands for your agent.

Built For

ML Engineers or Data Scientists who are tired of writing boilerplate code for every new LLM feature. Content Operations teams struggling to review and moderate massive streams of user-generated content before deployment.

Machine Learning Engineer

Uses the `list_models` tool to compare Mistral's different model sizes (large vs small) and then uses `create_batch` to run performance tests across hundreds of prompts.

Content Operations Manager

Runs a routine check using the `moderate` tool on all new user submissions, automatically flagging anything with high scores in violence or hate categories before it hits the live site.

What Changes When You Connect

01 Stop writing repetitive API calls. You can use the `chat` tool to talk to Mistral's models directly through your agent, making complex interactions feel like a natural chat session.

02 Move beyond keyword searches. By calling `embeddings`, you generate vector representations that allow your agent to search based on meaning and context, not just exact matches.

03 Process massive datasets reliably. The MCP handles batch jobs, so if you need to process ten thousand documents, you use `create_batch` and then check status with `list_batches` —you don't wait for a timeout.

04 Automate content review. Before accepting user input, you can run the `moderate` tool to instantly get safety scores, blocking harmful or violating text before it reaches your database.

05 Maintain clarity across models. The `list_models` tool lets you see all available Mistral options at a glance, ensuring you pick the right model (like 'codestral-latest') for the specific task at hand.

Real-World Applications

Building an internal knowledge retriever

A company needs to build a Q&A system that answers questions based on thousands of private documents. The agent uses `list_files` to upload the PDFs, then calls `embeddings` on chunks of text to create vectors, finally letting your AI client query those vectors for highly relevant answers.

Real-time user input safety checks

A messaging app needs to prevent abuse. Every message submitted is first passed through the `moderate` tool. If the score for 'hate' exceeds 0.5, the system automatically rejects the message and alerts a human moderator.

Analyzing large log files

An ML team needs to process millions of historical chat logs for sentiment analysis. They use `create_batch` with an endpoint designed for classification, allowing them to run the job overnight and check progress using `get_batch` in the morning.

Comparing model performance

A developer wants to know if Mistral's small model is fast enough. They use `list_models` to identify both 'mistral-small-latest' and 'mistral-large-latest', then send identical prompts via the `chat` tool to compare response time and quality metrics.

Patterns to Avoid

Treating text analysis as a single API call

✗ AVOID

A developer tries to handle both embedding creation and content moderation using only the `chat` tool, resulting in an unclear prompt and failure to receive specific safety scores.

✓ INSTEAD

Manually deleting files after use

✗ AVOID

After a batch job finishes, the user forgets that sensitive input data remains in the system's file storage, creating unnecessary security risk or clutter.

✓ INSTEAD

Ignoring processing status updates

✗ AVOID

The user runs `create_batch` on 10GB of data and simply waits. They don't check the progress, resulting in a timeout error because they didn't use `list_batches` or `get_batch`.

✓ INSTEAD

Using general LLMs for structured tasks

✗ AVOID

Asking an AI agent to classify data and simultaneously generate embeddings using only text chat, which cannot reliably produce the required vector format.

✓ INSTEAD

The Right Fit

Use this MCP if your process requires multiple distinct steps: first, you need to analyze content (moderation or chat); second, you need to turn that content into a searchable format (embeddings); and third, you have volumes of data that require background processing (batch jobs). This is the right choice when reliability and structured output are critical. Don't use it if your only goal is basic text summarization—a simple `chat` call might suffice. If your primary need is just to access a single LLM endpoint without complex workflow management, consider alternatives that provide a simpler

chat-only interface. However, for building robust, multi-stage applications, the combination of `embeddings`, `moderate`, and batch processing makes this MCP essential.

Dealing with content moderation is a headache.

When user submissions flood in—comments, forum posts, or chat logs—you currently have to build complex systems that check every piece of text against multiple rules. You might run separate API calls for profanity filters, then another call for hate speech detection, and finally manually review the results before allowing content through. It's a slow, expensive mess.

With this MCP, you simply ask your agent to moderate the input using the dedicated tool. The system runs all safety checks in one step, returning consolidated scores right away. You get clean data with explicit pass/fail metrics, letting you build guardrails into your workflow without writing complex filtering code.

Generate vector embeddings for semantic search using the `embeddings` tool.

Before this MCP, creating a knowledge base meant you had to write custom indexing scripts that pulled text and manually calculated vectors in an external service. The process was brittle, requiring constant maintenance whenever the input format changed.

Now, generating embeddings is a single, conversation-driven step. Your agent handles the complexity: you pass the text chunk, it calls `embeddings`, and you get structured vector data ready to plug right into your database. It makes building semantic retrieval systems straightforward.

Mistral AI MCP: 10 Tools Available

These tools give you programmatic access to every core function of Mistral AI—from chatting with LLMs to managing massive data pipelines.

#	TOOL	DESCRIPTION
01	`cancel_batch`	Stops a running batch job using its unique ID if processing needs to be halted early.
02	`chat`	Sends conversational messages to specified Mistral models, receiving the assistant's textual response.
03	`create_batch`	Starts a large processing job by pointing it to an input file and a specific API endpoint.
04	`delete_file`	Permanently removes a previously uploaded data file used for batch or document AI processing.
05	`embeddings`	Generates vector embeddings from text input, which are necessary for semantic search and database storage.
06	`get_batch`	Retrieves the detailed status and results of a specific batch processing job using its ID.
07	`list_batches`	Provides an overview of all past and current batch jobs, showing their status and file IDs.
08	`list_files`	Lists every data file uploaded to the MCP, including its ID, size, and purpose.
09	`list_models`	Shows a list of all available Mistral AI models with their IDs and technical capabilities.
10	`moderate`	Checks input text against safety guidelines, returning detailed scores for categories like hate or violence.

See It in Action

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

U Send a message to Mistral Large asking 'What is the capital of France?'

Mistral responded: 'The capital of France is Paris. It is the most populous city in France and serves as the country's political, economic, and cultural center.'

U List all available Mistral models.

Available models: mistral-large-latest (most capable, 128K context), mistral-small-latest (efficient, 32K context), codestral-latest (code specialist, 32K context), mistral-embed (embeddings, 8K context).

U Moderate this text: 'I want to learn about AI safety and content filtering.'

The text passed moderation with all safety scores below thresholds. Categories checked: sexual (0.01), hate (0.00), violence (0.00), self-harm (0.00), criminal (0.00). Content is safe to process.

Frequently Asked Questions

01 **How do I use Mistral AI MCP for chat completions?**

You use the `chat` tool by providing the desired model ID and the message array in your prompts. This lets you send conversations to various models like 'mistral-large-latest' while keeping everything within your agent flow.

02 **What is the difference between `embeddings` and `chat`?**

Chat is for back-and-forth conversation, returning natural language answers. Embeddings are for data storage; they convert text into numbers (vectors) so you can programmatically compare meaning across documents.

03 **Can I process millions of records using Mistral AI MCP?**

Yes. For large volumes, use the `create_batch` tool to set up a job. You then track its progress over time with `list_batches` and `get_batch`, ensuring stability and managing costs.

04 **What if I submit a batch job by mistake?**

You can use the `cancel_batch` tool immediately. Just provide the specific batch ID, and it stops all further processing for that job.

05 **Does Mistral AI MCP handle file management?**

Yes. You can manage files used by the service using `list_files` to see what's uploaded, or use `delete_file` when you are done with a dataset.

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 `"mistral-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

Mistral 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 Mistral 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	Mistral AI MCP
Server ID	`019d845a-2d74-7353-97bf-558e1150b6cc`
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/mistral-ai.