MCP SERVER | NO CODE | CLOUD HOSTED

Exa AI MCP

Turn Web Searching into Conversation

Exa AI MCP connects your agent to advanced web intelligence for deep research and analysis. It moves beyond basic keyword searches, allowing you to query the web based on meaning and context. You can instantly find sites similar to a target URL, pull clean text from any page, or search and retrieve content all in one go. Exa lets your AI client act like a real-time analyst for deep market research or competitor tracking.

A+ Quality Score 100/100

semantic-search | web-discovery | information-retrieval | content-analysis | url-matching

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

Exa AI MCP

4 tools available
Cloud-hosted on Vinkius

Stop bouncing between browser tabs just to gather context. This MCP gives your agent the power of semantic web discovery. You can tell it what you need—like understanding the trends around 'climate change mitigation'—and it searches the web based on meaning, not just matching keywords. Need to know who else is covering that topic? Your agent finds similar websites instantly. It also pulls clean text from complex pages so your workflow doesn't get bogged down in HTML junk. If you're doing deep market research or auditing sources for a project, this MCP ensures your answers are grounded in real-time, relevant data. Because Vinkius hosts this MCP, you connect once to your preferred AI client and gain access to sophisticated web analysis that turns complex information gathering into simple conversation.

Core Capabilities

01 — Perform semantic searches

Query the open web using natural language, retrieving results based on meaning rather than literal keywords.

02 — Identify related websites

Find websites that share characteristics with a given URL for competitor analysis or source mapping.

03 — Extract structured content

Retrieve clean, usable text and HTML content from specific web pages without manual scraping.

04 — Execute combined search tasks

Run a single request that searches the web and pulls the full page contents simultaneously.

05 — Monitor API usage

Track your research budget by monitoring API calls and crawl statistics within your agent's workflow.

One Click on Vinkius — From Prompt to Execution

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

01 Subscribe to this MCP on Vinkius and provide your Exa API Key.

02 Connect the MCP to your preferred AI client (like Claude or Cursor).

03 Prompt your agent with a request, such as 'Search for X and give me the content of the top result.' The agent executes the search, pulls the data, and presents it.

The bottom line is that you get continuous web analysis piped directly into your conversational workflow.

Built For

This MCP is essential for research analysts who spend hours sifting through search results. It's also critical for AI developers building agents that need real-time, context-aware web data to function correctly.

Research Analyst

Uses the MCP to monitor semantic search results and track competitor links across multiple industry sources without manual cross-referencing.

AI Developer

Implements `search_with_contents` in agentic applications, ensuring that content extraction is verifiable and reliable for downstream tasks.

Content Strategist

Runs the MCP to find similar sources or audit related websites when planning new coverage areas for a client.

What Changes When You Connect

01 Stop wading through messy HTML. Use `get_contents` to pull clean, structured text from any web page directly into your agent's memory.

02 Identify competitors instantly. Run a check for similar sites using the appropriate tool to map out related sources or potential market overlaps.

03 Cut down on requests by combining actions. The `search_with_contents` tool lets you search and retrieve full content in one go, speeding up complex workflows.

04 Keep your research budget tight. Monitor API usage directly through the MCP to maintain strict control over your web discovery spending.

05 Get deeper results with semantic power. Use `search_web` for queries that rely on meaning rather than just matching keywords.

Real-World Applications

Competitor Landscape Audit

A marketing analyst needs to know who else is covering a new industry standard. Instead of manually Googling, they prompt their agent to run `search_web` and then use the similar discovery capability to build a comprehensive list of related publications.

Academic Literature Review

A PhD candidate needs deep context on 'renewable energy policy'. They ask their agent to perform a semantic search, retrieve multiple sources using `search_with_contents`, and analyze the full text against their hypothesis in one step.

Technical Documentation Check

A developer needs to verify a specific piece of documentation online. They use the MCP's content retrieval tools to pull clean, verifiable text from a URL ID, ensuring their agent is working with accurate data, not just snippets.

Rapid Trend Monitoring

An operations lead needs to monitor niche industry trends daily. They set up an automated workflow that runs `search_web` on key phrases and flags any major changes in related websites or coverage.

Patterns to Avoid

Keyword stuffing searches

✗ AVOID

Running simple Google-style searches with long, keyword-stuffed queries that miss the actual context of what you need.

✓ INSTEAD

Use `search_web` to query semantically. This allows your agent to understand the *meaning* behind your query and pull results based on topic relevance, not just keyword matches.

Copy-pasting snippets

✗ AVOID

Having an AI generate a search result summary but then having to manually jump into another tab to verify the full text or context.

✓ INSTEAD

Use `search_with_contents` or `get_contents`. This retrieves the actual, cleaned page content directly to your agent's output, giving you the complete source material.

Ignoring deprecated tools

✗ AVOID

Relying on outdated endpoints like `find_similar` because they were listed previously.

✓ INSTEAD

Always check for modern capabilities. For finding similar sites, use the designated semantic search methods available via the MCP to ensure reliable, up-to-date results.

The Right Fit

Use this MCP if your research requires verifiable, deep context from external web sources or when you need to map out competitor coverage. You should connect here if your goal is 'What does this topic mean across the entire internet?' If, however, you only need general information retrieval and are confident that a simple keyword search against an indexed knowledge base will suffice, then stick with a standard RAG (Retrieval Augmented Generation) tool. Don't use this MCP if you just want to know the definition of a word; it's overkill. Use it when you need to prove what someone *actually* said or wrote online.

The web is always changing, and context is everything.

Today, deep research means opening dozens of tabs, running multiple searches with slightly different keywords, copying snippets of text into a spreadsheet, and cross-referencing data points just to build a coherent picture. It's time-consuming and you always risk missing the full context or pulling in messy junk HTML.

With this MCP, your agent handles that whole process. You ask it about a topic, and it doesn't just give you five links; it executes sophisticated semantic searches, finds related sources, and pulls clean, readable text for all of them. The result? A fully contextualized data packet ready for analysis.

Exa AI MCP delivers verifiable web context.

Manual processes involve searching with keywords, then running another search to find similar sites, and finally opening each site to copy the text. This is a multi-step, error-prone sequence that requires constant human intervention and switching between tools.

Now you instruct your agent to perform a full discovery task—combining `search_web` with content retrieval in one go. You get the intelligence immediately, verified by clean data pulled through `get_contents`. Your focus shifts from gathering data to analyzing it.

Exa AI: 4 Web Discovery Tools

These tools allow your agent to perform everything from basic semantic searches to complex, combined content retrieval for deep web analysis.

#	TOOL	DESCRIPTION
01	`find_similar`	Finds websites that are similar to a given URL; note this endpoint is deprecated and should not be used.
02	`get_contents`	Retrieves the cleaned text content specifically for provided page IDs.
03	`search_with_contents`	Performs a search across the web and pulls both the results and their full page contents in one request.
04	`search_web`	Runs a semantic query against the open web, returning relevant URLs, titles, and metadata.

See It in Action

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

U Find 5 high-quality research papers about 'Climate Change mitigation' using Exa.

I've retrieved 5 highly relevant research papers. Notable results include studies on carbon capture and renewable energy adoption. Would you like the links and summaries for each?

U Find sites similar to https://techcrunch.com.

I've found 10 sites similar to TechCrunch. Top matches include VentureBeat, The Verge, and Wired. I can provide the relevancy scores for each if you'd like.

U Search for 'future of AI robotics' and give me the full content of the best result.

Search complete! The top result is a detailed analysis from MIT. I've retrieved the cleaned content, which discusses embodied AI and humanoid developments. Would you like a summary?

Frequently Asked Questions

01 **How does Exa AI MCP perform semantic searching?**

Exa AI performs semantic searches using meaning-based embeddings rather than simple keyword matching. This means your agent understands the intent of your query and finds results based on topic relevance, even if keywords aren't used.

02 **Can I use Exa AI MCP to monitor my usage?**

Yes, you can track your research budget. The MCP provides monitoring tools so your agent workflow keeps a count of API calls and crawl statistics, keeping you in control of costs.

03 **What is the difference between `search_web` and `search_with_contents`?**

`search_web` returns metadata—just the URLs, titles, and summaries. However, `search_with_contents` performs both actions in one request, giving you the full, clean text content of the best results instantly.

04 **Does Exa AI MCP help with competitor analysis?**

Absolutely. You can run a query to find sites similar to a known URL, helping you map out competitors or related sources for market research purposes.

05 **Is the `find_similar` tool still good to use in Exa AI MCP?**

No. The listing notes that `find_similar` is deprecated by Exa. You should rely on the semantic search capabilities of the other tools for accurate site similarity checks.

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 `"exa-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

Exa 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 Exa 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	Exa AI MCP
Server ID	`019d8437-a7b1-73c7-abcc-df6541b6870c`
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/exa-ai.