

MCP SERVER

NO CODE

CLOUD HOSTED

Diffbot MCP for AI Agents

Structured Web Data Extraction for Research and Content Analysis

Diffbot lets your AI agent automatically extract structured data from any website. It processes complex web pages—whether they're news articles, e-commerce product listings, or forum discussions—and converts the messy content into clean JSON. You just point it at a URL, and your agent handles everything else.

A+ Quality Score 100/100

web-scraping

data-extraction

computer-vision

structured-data

e-commerce-intelligence

sentiment-analysis



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.

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.

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.

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

Encryption at rest

Ed25519

PKI vault signatures

24h TTL

Ephemeral session keys

V8 Isolate

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.

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.

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.

06 — Honeypot Trap System

Phantom credentials are injected into isolated environments. If a honeypot 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.

FinOps

Cost tracking per tool, payload compression savings, budget optimization signals, and consumption trends.

Firewall & DLP

PII redaction activity, sensitive data protection counters, and security event timeline.

Agent Activity

Which AI clients are connecting, how often, and what they're doing — real-time session tracking.

Tool Health

Slowest and most error-prone tools, with actionable root-cause insights and performance baselines.

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.

Diffbot MCP

10 tools available

Cloud-hosted on Vinkius

Diffbot gives your AI client direct access to structured web data extraction. Instead of having to write complex scrapers or manually copy key details from dozens of sites, you ask your agent for what you need—and Diffbot retrieves it. The system analyzes the page type first; is it a product? An article? A list of search results? It figures it out and extracts the relevant data automatically.

This means whether you're tracking competitor pricing across multiple e-commerce sites or pulling clean, readable text from academic journals, your agent handles the dirty work. You can even analyze forum threads to gauge public sentiment or pull job market trends by gathering structured details like salary ranges and employer names. Because this MCP is available on Vinkius, you connect once with Claude, Cursor, or any compatible client, giving yourself a massive toolkit for turning raw web pages into actionable data.

Core Capabilities

01 — Classify Page Content

Automatically determines if a webpage is an article, product, list, image gallery, or job posting.

03 — Capture E-commerce Details

Retrieves structured product information, including SKUs, specific pricing, brand names, and technical specifications.

05 — Scrape Search Results or Directories

Identifies structured lists on a page, pulling out arrays of titles and direct links for batch processing.

02 — Extract Article Text

Pulls clean text and HTML from news or blog posts while identifying the author and publication date.

04 — Analyze Discussions & Reviews

Gathers content from forum threads or reviews, allowing you to analyze the overall sentiment of user feedback.

One Click on Vinkius — From Prompt to Execution

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

- 01 Subscribe to this MCP and enter your Diffbot Developer Token into your AI client.
- 02 Tell your agent the URL you want data from, along with what specific information you need (e.g., 'What is the price and SKU for this product?').
- 03 Your agent invokes the appropriate tool, and Diffbot returns a clean JSON object containing only the structured data.

The bottom line is: your AI client turns raw URLs into reliable, usable data structures without you needing to write any scraping code.

Built For

Anyone who works with web content but hates manual copy-pasting. Data Analysts need structured inputs for reporting; Market Researchers track competitor pricing across dozens of sites; and Content Marketers need fast ways to summarize articles or monitor brand sentiment from forums.

Data Analyst

Using Diffbot, you feed the tool URLs for thousands of websites and extract specific data points into structured JSON arrays for immediate database loading.

Market Researcher

You monitor competitors by feeding your agent a list of product page URLs, getting back standardized pricing, brand mappings, and specifications for comparison reports.

Content Marketer

Instead of reading every forum thread manually, you feed the tool review pages to automatically aggregate user sentiment scores and extract common topics discussed.

What Changes When You Connect

-
- 01 Get precise e-commerce data, including SKU numbers and brand mappings. The `extract_product` tool makes it possible to scrape critical product details in one go.

 - 02 Stop guessing what a page is. Use the general classification tool (`analyze_page`) to instantly determine if you're looking at an article, list, or job posting before running any extraction.

 - 03 Analyze public sentiment without reading thousands of comments. The `extract_discussion` tool pulls forum threads and prepares them for automated sentiment scoring.

 - 04 Monitor market trends by gathering standardized data. You can use the `extract_job` tool to pull salary vectors and employer names from career sites across different industries.

 - 05 Process content efficiently with `extract_article` . This gives you clean, readable text bodies separated from boilerplate site navigation or ads.
-

Real-World Applications

Competitive Pricing Monitoring

A market researcher needs to track how three competitors change their pricing on key products weekly. Instead of visiting and manually logging data, the agent uses Diffbot's API to gather structured product details from all URLs, giving a clean JSON report of price changes.

Building Job Market Reports

An HR analyst wants to see salary trends for software engineers in a specific city. The agent uses Diffbot's job extraction tool across multiple recruitment sites, providing a consolidated list of explicit salary ranges and employer names.

Curating News Aggregators

A content marketer needs to build a daily summary of industry news. The agent runs the `extract_article` tool on top search results to pull only the clean text and author information, eliminating boilerplate site clutter.

Analyzing Customer Feedback

A product manager wants to understand why customers are leaving 1-star reviews. The agent uses the `extract_discussion` tool on review pages, allowing them to analyze thousands of comments for common themes and sentiment.

Patterns to Avoid

Treating all web content as simple text

X AVOID

Trying to copy-paste a complex product page into an AI prompt, hoping it extracts the SKU or price. You'll just get a wall of raw HTML and unstructured mess.

✓ INSTEAD

Use `extract_product` directly with the URL. This tool is built to identify specific e-commerce fields like brand mappings and SKUs, giving you structured data instead of garbage text.

Overlooking page type classification

X AVOID

Running a general article extraction tool on a search results page will fail because the content structure is wrong. The output will be irrelevant.

✓ INSTEAD

First, run `analyze_page` to confirm if the page is actually a 'list' or 'search result'. If it is, use the `extract_list` tool for accurate title and link arrays.

Ignoring custom logic needs

X AVOID

Needing data from a highly specialized, non-standard website (like an internal government portal) that no standard API covers. The output will be incomplete.

✓ INSTEAD

You must use `extract_custom_api`. This tool bridges your specific rulesets to the raw URL, enabling extraction even when the site structure is unique.

The Right Fit

Use this MCP if your primary goal is converting messy web pages into structured data formats like JSON. If you are dealing with e-commerce sites, always start with `extract_product` or `analyze_page`. Don't use it if you only need to summarize a single block of text that isn't sourced from the open web; for that, a basic text processing tool is fine. You shouldn't use it if your data source is non-web (e.g., local CSV files). Remember, this MCP requires external tokens and relies on network access to the URL you provide.

Diffbot MCP for AI Agents: Capturing Structured Product Details from E-commerce Sites

Right now, getting product information is a nightmare. You open an e-commerce site, scroll through the page, and manually copy the SKU, then click to find the price, then switch tabs to note the brand mapping. This process takes minutes per item and breaks down fast when you have dozens of products to compare.

With this MCP, your agent handles it all in one go. By using tools like `extract_product`, you point at a list of URLs, and the agent returns clean JSON objects containing precise pricing, SKU details, and specifications for every single item. You get a ready-to-use data feed.

Diffbot MCP for AI Agents: Monitoring Web Trends with Structured Data

Before connecting Diffbot, monitoring market trends means running multiple tabs and manually pulling job titles, salary ranges, or competitor article summaries. This is slow, error-prone work that requires dedicated analyst time.

Now, you simply ask your agent to monitor a category of sites. The tool executes the extraction across those pages—whether it's gathering data via `extract_job` for market analysis or summarizing multiple articles using `extract_article`. You get organized intelligence, not just raw links.

10 Tools in the Diffbot MCP for Structured Web Data Extraction

Use these ten tools to extract anything from a webpage—from article text bodies and product specs to job postings and forum reviews—and get it into clean JSON.

#	TOOL	DESCRIPTION
01	<code>analyze_page</code>	Automatically classifies any web page and extracts structured data like articles, products, or events in a single pass.
02	<code>extract_article</code>	Extracts clean content from news sites, identifying the title, author, date, and body text of an article.
03	<code>extract_custom_api</code>	Allows you to pull data using specific extraction rules that you define in your own Diffbot dashboard setup.
04	<code>extract_discussion</code>	Gathers comments and reviews from forum threads, allowing analysis of user-generated conversations.
05	<code>extract_event</code>	Pulls schedules and details for events, giving you organized information about dates and locations.
06	<code>extract_image</code>	Retrieves the main images from a page so you can build galleries or identify key visuals.
07	<code>extract_job</code>	Extracts specific job details, including titles, employer names, and salary ranges, from career pages.
08	<code>extract_list</code>	Identifies bounded search results or directory listings on a page to extract arrays of links and item titles.
09	<code>extract_product</code>	Extracts comprehensive e-commerce data points like pricing, brand names, SKUs, and product descriptions.
10	<code>extract_video</code>	Gathers video metadata and content details from a webpage so you can track media assets.

See It in Action

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

U What are the core specs and pricing for this phone model: <https://techsite.com/iphone16?>



Product Details Report

- **Name:** iPhone 16 Pro Max
- **Brand:** Apple
- **SKU:** IPH-16PM-256
- **Price:** \$1,199.00 (In Stock)
- **Availability:** Limited stock.

Specifications:

- Display: 6.7 inches, ProMotion
- Storage options: 128GB to 1TB
- Camera: 48MP Main Lens

Need help comparing this against last year's model?

U Find the main topics and sentiment from these customer reviews: [Review List URL]



Customer Feedback Analysis

Based on 450 reviews, here are the top three themes:

1. **Battery Life (Positive):** Most users report a significant improvement in longevity. **(Sentiment Score: +0.8)**
2. **Camera Quality (Neutral/Negative):** While sharp, many complained about low-light performance compared to previous models. **(Sentiment Score: -0.3)**
3. **User Interface (Positive):** The new UI is simple and intuitive.

I pulled this data using the discussion extraction tool.

U Can you summarize the key takeaways from this industry report article? [Article URL]



Key Takeaways: Global AI Adoption (2024)

- **Title:** The Future of Enterprise AI
- **Author:** Dr. Evelyn Reed
- **Date Published:** June 1, 2024

The report highlights three major shifts:

- ✓ **Decentralization:** Companies are moving away from single vendor solutions toward hybrid models.
- ✓ **Edge Computing:** Processing power is shifting closer to the user for real-time data handling.
- ✓ **Regulatory Focus:** New global standards require greater transparency in AI model training.

Would you like me to generate a bulleted summary of the executive abstract?

Frequently Asked Questions

01 How does Diffbot MCP for AI Agents help with web scraping when I don't know the HTML structure?

It doesn't matter if you know the code. The MCP uses advanced classification to understand what content is—whether it's a price, an article title, or a user comment. It gives you structured data automatically.

02 Can I use Diffbot MCP for AI Agents to track competitor pricing across multiple product pages?

Yes. You can feed the agent a list of URLs and ask it to pull standardized fields like SKU, price, and brand mapping from every page into one report.

03 Is Diffbot MCP for AI Agents better than just using my AI client's native web browsing feature?

Yes. Native browsing gives you raw text; this MCP gives you machine-readable, structured JSON data. This means your agent can reliably use the data in subsequent steps without errors.

04 What kind of websites can Diffbot MCP for AI Agents handle? Is it limited to news sites?

It handles almost anything: e-commerce, job boards, academic articles, forum discussions, and even specialized directories. The tool adapts to the page type.

05 I want to analyze customer reviews; what specific data can Diffbot MCP for AI Agents extract?







It pulls out individual comments from discussion threads, allowing your agent to run automated sentiment scoring and group common feedback themes across thousands of entries.

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 <code>"diffbot": { "url": "..." }</code>
 Windsurf	MCP Settings → <code>mcp_settings.json</code> → 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

Diffbot 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](https://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 Diffbot. 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	Diffbot MCP
Server ID	019d7585-b446-73ab-8d43-a10a1d1a1eb2
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/diffbot.