

# Technical Manual and Comprehensive Research Report: Context7 MCP Server

## 1. Introduction and Strategic Context

The integration of Large Language Models (LLMs) into the software development lifecycle has fundamentally altered the velocity of code production. However, this integration faces a persistent, structural bottleneck: the "knowledge cutoff." Modern software ecosystems—encompassing frameworks like Next.js, cloud infrastructures like Cloudflare Workers, and utility libraries like Tailwind CSS—iterate at a cadence that far outstrips the training cycles of foundational models. A model trained on data from 2023 is functionally illiterate regarding the API changes introduced in late 2024 or 2025. This temporal disconnect results in hallucinations, where the model confidently generates syntactically correct but operationally defunct code, citing deprecated methods or phantom APIs.

The Context7 Model Context Protocol (MCP) Server emerges as a critical infrastructure component designed to bridge this temporal gap. By leveraging the standardized Model Context Protocol, Context7 provides a mechanism for dynamic, real-time documentation retrieval that is semantically integrated into the LLM's cognitive workflow.<sup>1</sup> Unlike generic web search tools, which return unstructured HTML noise, Context7 functions as a specialized Retrieval-Augmented Generation (RAG) system optimized specifically for technical documentation. It parses, indexes, and serves "semantic chunks" of version-specific code and API references directly into the context window of Integrated Development Environments (IDEs) such as Cursor, Windsurf, and VS Code.<sup>3</sup>

This report offers an exhaustive technical analysis of the Context7 MCP Server. It explores the system's evolution from a client-heavy token consumer to a server-optimized reranking engine, details the intricacies of its installation across the fragmented IDE landscape, and provides a comparative economic analysis against competitors like Ref.tools. Furthermore, it synthesizes community insights and troubleshooting protocols to serve as a definitive manual for enterprise architects and senior engineers seeking to operationalize this tool within high-velocity development teams.

---

## 2. Technical Architecture and Implementation

To understand the operational efficacy of Context7, one must analyze its underlying architecture, which has undergone significant evolution to address the inherent limitations of context-limited LLMs. The system operates not merely as a data fetcher but as a sophisticated middleware that negotiates between the user's intent and the vast,

unstructured corpus of open-source documentation.

## 2.1 The Model Context Protocol (MCP) Framework

Context7 is architected upon the Model Context Protocol (MCP), an open standard introduced by Anthropic to standardize the connection between AI assistants (hosts) and external data systems (servers). Before MCP, integrating a documentation source with an editor like Cursor required bespoke API plumbing. MCP normalizes this interaction through a client-host-server topology.<sup>3</sup>

- **The Host (Client):** This is the user's environment—applications like Claude Desktop, Cursor, Windsurf, or a CLI-based agent. The host is responsible for maintaining the conversation state and the LLM context window.
- **The Server:** The Context7 MCP executable (typically running as a local Node.js process via npx). It exposes specific capabilities—resources, prompts, and tools—to the host.
- **The Protocol:** Communication occurs via JSON-RPC 2.0 messages. The host queries the server for its capabilities ("What tools do you have?"), and the server responds with definitions ("I can resolve library IDs and fetch documentation").

This standardization allows Context7 to be "write once, run anywhere," instantly compatible with any tool that implements the MCP specification.

## 2.2 Transport Mechanisms

The reliability and latency of the interaction depend heavily on the chosen transport layer. Context7 supports two primary modes, each with distinct trade-offs regarding security and complexity.

Transport Protocol	Implementation Detail	Operational Context	Advantages	Disadvantages
<b>stdio</b> (Standard Input/Output)	JSON-RPC messages are piped directly via stdin and stdout between the IDE and the Node.js process.	Default for local development (Cursor, VS Code).	<b>Security:</b> Sandbox containment; no network ports are opened, reducing the attack surface.  <b>Simplicity:</b> Zero configuration	<b>Debugging:</b> Operational logs are often interleaved with transport messages, complicating troubleshooting.  <b>Runtime:</b> Requires a

			regarding firewalls or port forwarding.	local Node.js environment.
<b>HTTP / SSE</b> (Server-Sent Events)	JSON-RPC messages are sent via HTTP POST; asynchronous server updates are pushed via SSE.	Remote server hosting, multi-agent orchestration, dockerized deployments.	<b>Scalability:</b> Decouples the server from the IDE, allowing centralized hosting.  <b>Observability:</b> Traffic is easily inspected via standard network monitoring tools.	<b>Security:</b> Requires robust authentication (OAuth/API Keys) and encryption (TLS).  <b>Latency:</b> Introduces network overhead compared to in-memory pipes.

It is important to note that while the SSE protocol has been widely used, recent documentation indicates a deprecation notice for the specific SSE endpoint in favor of standard HTTP transport methods for future releases, aimed at simplifying the connection lifecycle.<sup>5</sup>

## 2.3 The Ingestion and Indexing Pipeline

The "magic" of Context7 lies in its backend pipeline, which transforms raw documentation into retrievable intelligence. This process is distinct from the open-source client code and resides on Context7's private infrastructure.<sup>1</sup>

### 2.3.1 Parsing and the llms.txt Standard

The pipeline begins with ingestion. Context7 targets documentation repositories, primarily on GitHub. The parsing engine supports a variety of formats including Markdown (.md, .mdx), Plain Text (.txt), ReStructuredText (.rst), and Jupyter Notebooks (.ipynb).<sup>6</sup> Crucially, the parser creates and utilizes llms.txt files. This emerging standard functions similarly to robots.txt but for LLMs. Instead of directives for crawlers, llms.txt provides a high-density, pre-processed summary of a library's documentation, stripped of HTML boilerplate, navigation elements, and non-technical filler.<sup>7</sup> This file serves as a map, allowing the system to identify the most relevant sections of a library without ingesting the entire site structure, significantly reducing

noise.

### 2.3.2 Enrichment and Vectorization

Raw code snippets are often semantically opaque. A function named `processData` might be ambiguous. To resolve this, Context7 employs an "Enrich" step where auxiliary LLMs generate metadata and concise explanations for each snippet.<sup>7</sup> A snippet is tagged with concepts like "authentication," "middleware," or "streaming," enabling semantic retrieval. These enriched chunks are then embedded into high-dimensional vectors and stored in a vector database (upstash/vector), enabling the system to match natural language queries ("how to secure routes") with code implementations.

### 2.3.3 The 2026 Architectural Shift: Server-Side Reranking

In early 2026, Context7 underwent a major architectural overhaul to address "context bloat." In the legacy system (v1), the server performed a standard vector search and returned a broad set of results—often 10,000 tokens or more—relying on the user's LLM to filter the relevance.<sup>2</sup> This was inefficient and expensive.

The updated architecture introduces a **Server-Side Reranking** step.

1. **Retrieval:** The system fetches a wide pool of candidate chunks via vector search.
2. **Reranking:** A specialized, high-speed model on the server scores these candidates against the user's specific query logic.
3. **Delivery:** Only the top-scoring, highly relevant chunks are sent to the client.

This shift resulted in dramatic efficiency gains: a **65% reduction** in average context tokens (from ~9.7k to ~3.3k) and a **38% reduction** in latency.<sup>2</sup> By moving the cognitive load of filtering from the client (the expensive LLM) to the server (the optimized reranker), Context7 significantly improved the signal-to-noise ratio of its output.

## 2.4 Caching and Rate Limiting

To maintain system stability and performance, Context7 implements rigorous caching and rate-limiting protocols.

- **Infrastructure Caching:** Responses are cached using Redis (Upstash). Since documentation for a specific library version changes infrequently, repeated queries for "Next.js 14 Middleware" can be served instantly from the cache, bypassing the vector search entirely.<sup>1</sup>
- **Rate Limits:** Access is tiered. Anonymous users (no API key) are subject to strict IP-based limits (approx. 100 requests/day). Authenticated users (Pro/Enterprise) access higher tiers (up to 10,000+ requests). The system returns standard HTTP 429 errors when limits are breached, often including a `Retry-After` header to guide client backoff strategies.<sup>8</sup>

---

## 3. Installation and Environment Configuration

The versatility of the MCP standard allows Context7 to be deployed across a diverse array of development environments. While the core execution relies on Node.js (requiring version 18.0.0 or higher), the configuration syntax varies significantly between clients.

### 3.1 Cursor: The Primary Deployment Target

Cursor, a VS Code fork with native AI capabilities, represents the most common use case for Context7.

#### Configuration Strategy:

While Cursor offers a UI for adding MCP servers, direct manipulation of the configuration file is recommended for persistence and reproducibility.

- **File Path:** ~/.cursor/mcp.json (Global) or .cursor/mcp.json (Project-specific).
- **Configuration Block:**

```
JSON
{
  "mcpServers": {
    "context7": {
      "command": "npx",
      "args": [
        "-y",
        "@upstash/context7-mcp"
      ],
      "env": {
        "CONTEXT7_API_KEY": "YOUR_API_KEY_HERE"
      }
    }
  }
}
```

- **Note on Arguments:** The -y flag is critical as it suppresses the interactive prompt ("Need to install the following packages:...") which would otherwise cause the stdio connection to hang and time out.<sup>1</sup>

### 3.2 Claude Code: The CLI Agent

Claude Code, Anthropic's terminal-based coding tool, manages MCP servers via its own internal registry.

- **Installation Command:**

Bash

```
claude mcp add context7 -- npx -y @upstash/context7-mcp --api-key YOUR_KEY
```

This command abstracts the JSON configuration, automatically registering the server in the Claude Code environment. For users requiring advanced configuration (like timeouts), editing the config.toml file in ~/.claude/ is necessary.<sup>10</sup>

### 3.3 VS Code and Windsurf

For standard VS Code (using the MCP extension) and Windsurf (Codeium's editor):

- **File Path:** typically %APPDATA%/Code/User/globalStorage/mcp-servers.json or ~/.vscode/mcp.json.
- **Syntax:** Identical to the Cursor JSON format.
- **Troubleshooting:** VS Code requires strict JSON syntax. A missing comma or trailing comma will prevent the MCP server from starting. Additionally, users must ensure the npx command is in the system PATH visible to VS Code, which can sometimes be an issue on Windows machines.<sup>11</sup>

### 3.4 Specialized Clients: Cline, JetBrains, and Zed

- **Cline:** An autonomous coding agent extension. It features a "MCP Servers" tab in its settings where users can input the command (npx) and args (-y, @upstash/context7-mcp). Cline also supports a cline\_mcp\_settings.json file for manual edits.<sup>12</sup>
- **JetBrains (IntelliJ/PyCharm):** Support is integrated via the AI Assistant plugin. Users navigate to **Settings -> Tools -> AI Assistant -> Model Context Protocol** to add the server definition.<sup>13</sup>
- **Zed:** The high-performance editor supports MCP via extensions. Configuration is added to the settings.json under the context\_servers key.<sup>1</sup>

### 3.5 Advanced Configuration: Environment Variables

For enterprise deployments where security and reliability are paramount, configuration via environment variables is preferred over hardcoding secrets in JSON files.

Variable	Function	Default	Recommended Tuning
CONTEXT7_API_KEY	Authentication token. Takes precedence over CLI flags.	None	Store in .env or system vault.

MCP_TIMEOUT	Startup timeout. Determines how long the client waits for the server to initialize.	10000ms	<b>60000ms.</b> Essential for npx usage, as package download often exceeds 10s.
MCP_TOOL_TIMEOUT	Execution timeout. Time allowed for a tool to return data.	30000ms	<b>60000ms.</b> Deep searches or large doc fetches can spike latency.
https_proxy / http_proxy	Network proxy. Context7 respects standard Node.js proxy configurations.	None	Mandatory for corporate firewalls.

**The Timeout Bottleneck:** A pervasive issue reported in community threads is the "Connection Timeout" error upon first use.<sup>14</sup> This is almost invariably caused by the npx process taking longer to download the @upstash/context7-mcp package than the default 10-second timeout allows. Increasing MCP\_TIMEOUT to 60 seconds is a critical best practice for robust installation.<sup>15</sup>

---

## 4. Operational Tools and API Mechanics

Context7 exposes its functionality through a concise set of tools. Unlike a general-purpose web browser, these tools enforce a structured interaction model that guides the LLM toward precise data retrieval.

### 4.1 resolve-library-id: The Semantic Directory

The entry point for most interactions is resolve-library-id. The LLM typically knows a library by its colloquial name (e.g., "React Query"), but the system requires a precise identifier.

- **Mechanism:** This tool accepts a libraryName and a query. It performs a fuzzy match against the Context7 index.
- **Relevance Ranking:** The query parameter is vital. If a user asks about "routers," the system uses the query context to prioritize react-router over next/router or vue-router based on the surrounding conversation.
- **Output:** It returns a list of candidates including the canonical id (e.g., /tanstack/react-query), a trustScore (indicating source authority), and available versions.<sup>1</sup>

## 4.2 get-library-docs (and query-docs)

Once the ID is resolved, the get-library-docs (or in newer versions, query-docs) tool is invoked.

- **Mechanism:** This triggers the core RAG pipeline. It takes the libraryId and the user's query.
- **The Query Transformation:** The LLM often reformulates the user's high-level request into a specific search query. For instance, "How do I use auth?" becomes "authentication middleware implementation examples."
- **Payload:** The response contains "snippets"—discrete blocks of text and code—along with metadata like source URLs. Due to the 2026 update, this payload is highly filtered, containing only the top-k results deemed relevant by the reranker.<sup>1</sup>

## 4.3 Optimization: The "Slash Syntax"

To optimize token usage and latency, Context7 supports a bypass mechanism known as "slash syntax." If the user or the agent already knows the correct library ID, they can invoke it directly in the prompt:

"How do I configure row level security? use library /supabase/supabase"

- **Benefit:** This short-circuits the resolve-library-id step. The MCP server detects the valid ID pattern and proceeds immediately to documentation retrieval. This saves one round-trip of network latency (approx. 1-2 seconds) and the input/output tokens associated with the resolution tool call.<sup>18</sup>

## 4.4 Configuration for Library Authors: context7.json

Context7 allows library maintainers to control how their documentation is indexed via a context7.json file placed in the repository root. This functions similarly to a website's configuration file.

- **Structure:**

```
JSON
{
  "projectTitle": "Acme SDK",
  "excludeFolders": ["internal", "deprecated", "tests"],
  "branches": ["main", "v2-stable"],
  "rules": ["Always use async/await", "Prefer interface over type"]
}
```

- **Impact:** The excludeFolders directive is crucial for token efficiency, preventing the indexing of massive but irrelevant directories like node\_modules or archived versions. The rules array allows authors to inject "system prompt" style instructions that accompany



any documentation retrieved from their library, guiding the LLM on coding standards specific to that tool.<sup>1</sup>

## 5. Token Economics and Performance Analysis

The economic viability of Context7 is a central theme in its adoption, particularly when compared to the cost of raw token consumption.

### 5.1 The "Context Bloat" Crisis and Resolution

In the legacy era (pre-2026), RAG tools often operated on a "dump and filter" philosophy. A query would retrieve 10-20 "chunks" from a vector database, populating the context window with 10,000+ tokens of documentation.<sup>2</sup>

- **The Cost:** At typical API rates (e.g., \$15/1M tokens for Opus-class models), a single "bloated" query could cost \$0.15. For a developer iterating 50 times a day, this amounted to substantial daily waste.
- **The fix:** The 2026 server-side reranking update reduced this average load to ~3,300 tokens. This represents a **65% reduction** in token consumption per query.
- **Economic Impact:** The reduction saves approximately 6,400 tokens per query. Over a month of heavy usage (1,000 queries), this saves ~6.4 million tokens. Even at cheaper model rates (\$3/1M), this is a direct saving of ~\$20/month, effectively subsidizing the cost of the tool itself.

### 5.2 Comparative Economics: Context7 vs. Alternatives

The market offers alternatives like **Ref.tools** and custom-built scraping MCPs. Community benchmarks provide illuminating comparisons.<sup>21</sup>

Metric	Context7 (Post-2026)	Ref.tools	Custom Web Scraper MCP
Token Cost	Moderate (~3.3k). Optimized for <i>contextual breadth</i> .	Low (~500-1k). Optimized for <i>surgical precision</i> .	High (Variable). Often returns raw HTML garbage unless heavily filtered.
Per-Query Cost	~\$0.04 - \$0.05	~\$0.08 (historical avg)	~\$0.02 - \$0.04
Data Quality	High.	High. Live crawling	Low. Often

	Pre-processed, cleaned markdown. Semantically enriched.	with "reader mode" extraction.	contains navbars, ads, and broken formatting.
<b>Update Frequency</b>	Delayed (10-15 day recrawl cycle).	Live (Real-time crawling).	Live (Real-time crawling).

**Analysis:** While custom scrapers are cheaper in terms of *execution* (no subscription), they are often more expensive in *token consumption* due to the lack of pre-processing. Ref.tools competes on precision, claiming up to 85% token savings by finding the *exact* function signature rather than a documentation page. Context7 occupies a middle ground: it provides enough context for the LLM to understand *patterns* and *concepts*, not just syntax, which is often necessary for complex architectural queries.

### 5.3 Pricing Tiers and Limits

Context7 employs a tiered pricing model to manage infrastructure costs.<sup>23</sup>

Tier	Request Limit	Private Repos	Features
<b>Free</b>	~100 / day (variable)	No	Basic public library search. Subject to aggressive rate limits during peak loads.
<b>Pro</b>	~5,000 / seat / month	Yes	Higher rate limits, private repository indexing, email support.
<b>Enterprise</b>	Custom	Yes	SSO, SLA, unlimited indexing, dedicated support.

**Community Friction:** Users have noted that the "Pro" limits can feel restrictive during intense coding sessions, with some reporting hitting caps after heavy refactoring work involving multiple libraries. However, the limits are generally sufficient for standard daily workflows.<sup>25</sup>

---

## 6. Usage Manual and Best Practices

Maximizing the utility of Context7 requires a shift in prompting strategy. Users must treat the tool as a librarian, not a search engine.

### 6.1 Prompt Engineering for RAG

- **Specificity is Currency:** Vague prompts yield vague documentation.
  - *Weak Prompt:* "Help me with Supabase." -> Returns generic "Getting Started" guide.
  - *Strong Prompt:* "How do I implement row-level security policies for a 'profiles' table in Supabase using the Javascript client? use context7" -> Returns specific API definitions for createPolicy and RLS syntax.
- **Version Pinning:** The single most common failure mode is **Version Hallucination**. If a library has undergone a major breaking change (e.g., Next.js Pages vs. App Router), the LLM might guess the wrong version.
  - *Best Practice:* Explicitly state the version in the prompt. "Using Next.js 14 App Router conventions..." forces the reranker to downrank legacy v12 documentation.<sup>1</sup>

### 6.2 Multi-Library Workflows

Complex applications rarely rely on a single library. Orchestrating Context7 for multi-library tasks requires care to avoid context overflow.

- **Sequential Loading:** Instead of asking for "Docs for React, Tailwind, and Supabase" in one prompt (which floods the context), adopt a sequential strategy.
  1. "First, find the docs for a Supabase Auth hook." (Wait for response).
  2. "Now, finding the docs for Tailwind forms, style a login component using that hook."
- **The "One-at-a-Time" Rule:** Limit tool calls to one major library per turn. This allows the LLM to synthesize the information before moving to the next layer of complexity.<sup>1</sup>

### 6.3 Workflow: Private Repository Indexing

For enterprise users, the ability to index private code is a killer feature.

1. **Preparation:** Ensure the private repository contains a context7.json file in the root. This file *must* contain a projectTitle to aid the indexer.
  2. **Authorization:** Through the Context7 dashboard, link the GitHub account and authorize access to the specific repo.
  3. **Claiming:** Use the dashboard to "Claim" the library. This triggers the ingestion pipeline.
  4. **Verification:** If verification fails (a common error), ensure the API key used in the IDE matches the account that claimed the repo. The system validates ownership to prevent unauthorized indexing of proprietary code.<sup>5</sup>
-

## 7. Troubleshooting and Maintenance

Operationalizing Context7 involves navigating a series of common failure modes.

### 7.1 Error 429: Too Many Requests

This indicates the user has hit the rate limit.

- **Symptom:** The IDE shows "Error: Too many requests" or the tool returns an empty result.
- **Resolution:**
  1. **Check API Key:** Ensure a valid key is configured. Anonymous usage is extremely limited.
  2. **Backoff Strategy:** If using an automated agent, implement exponential backoff (wait 1s, then 2s, then 4s) when this error is encountered.<sup>8</sup>
  3. **Dashboard Check:** Verify if the monthly quota has been exhausted.

### 7.2 Connection Timeouts and ENOENT

- **Symptom:** "Error spawn node ENOENT" or "Connection timeout" during startup.
- **Cause:** The IDE cannot locate the Node.js binary, or npx is timing out.
- **Resolution:**
  1. **Absolute Paths:** In mcp.json, replace "npx" with the full path to the executable (e.g., /usr/local/bin/npx).
  2. **Increase Timeout:** Set MCP\_TIMEOUT=60000 in the environment variables.
  3. **Pre-install:** Run npm install -g @upstash/context7-mcp globally. Then change the command in mcp.json to node and args to the global path of the script. This bypasses the npx download step entirely.<sup>16</sup>

### 7.3 "Fetch is not defined"

- **Symptom:** A reference error occurring within the MCP server process.
- **Cause:** This typically occurs when running on an outdated version of Node.js (pre-v18) where the global fetch API was not yet standard.
- **Resolution:** Upgrade Node.js to v20 or v22 (LTS). Verify the version by running node -v in the terminal used by the IDE.<sup>29</sup>

---

## 8. Comparative Ecosystem Analysis and Conclusion

### 8.1 Context7 in the RAG Landscape

The landscape of AI documentation tools is bifurcated.

- **Web Search (Brave/Google):** Good for "What is X?" questions. Poor for coding because it lacks depth.
- **Ref.tools:** Excellent for "surgical" fixes and live debugging of brand-new releases. It is

the sniper rifle of documentation.

- **Context7:** Excellent for "architectural" understanding and implementing features using established libraries. It is the encyclopedic reference.

Community sentiment on platforms like Reddit suggests a division: power users who prefer total control often gravitate toward custom-built RAG pipelines (using tools like agent-fusion), while enterprise developers favor the "plug-and-play" reliability of Context7.<sup>30</sup>

## 8.2 The Future: lms.txt and Beyond

Context7's heavy investment in the lms.txt format positions it strategically for the future. As more library authors adopt this standard, the need for expensive vector search and reranking may diminish. The system could evolve into a specialized browser that simply "reads" the declared lms.txt of a project, providing 100% accurate, author-sanctioned context with near-zero latency.<sup>7</sup>

## 8.3 Conclusion

The Context7 MCP Server is more than a convenience; it is a necessary adaptation to the rapid pace of software evolution. By decoupling the LLM's knowledge base from its training date, it allows developers to utilize state-of-the-art models with state-of-the-art libraries. While it requires careful configuration—particularly regarding timeouts and version specificity—and a disciplined approach to prompt engineering to manage token costs, its ability to inject verified, version-correct code examples into the IDE makes it an indispensable tool for the modern software engineer. The 2026 architectural updates have successfully addressed the critical flaw of context bloat, maturing the tool from an experimental utility into a robust enterprise asset.

## Bibliografia

1. Context7 MCP Server -- Up-to-date code documentation for LLMs and AI code editors - GitHub, accesso eseguito il giorno febbraio 1, 2026, <https://github.com/upstash/context7>
2. Context7 Without Context Bloat | Upstash Blog, accesso eseguito il giorno febbraio 1, 2026, <https://upstash.com/blog/new-context7>
3. Context7 MCP: Up-to-Date Docs for Any Cursor Prompt | Upstash Blog, accesso eseguito il giorno febbraio 1, 2026, <https://upstash.com/blog/context7-mcp>
4. Claude Code Just Fixed MCP's Biggest Problem, accesso eseguito il giorno febbraio 1, 2026, <https://www.youtube.com/watch?v=itS3f1Y52t0>
5. Context7 MCP - Up-to-date Code Docs For Any Prompt - LobeHub, accesso eseguito il giorno febbraio 1, 2026, <https://lobehub.com/mcp/upstash-context7>
6. upstash/context7-legacy: Instant LLM Context for Agents and Developers - GitHub, accesso eseguito il giorno febbraio 1, 2026, <https://github.com/upstash/context7-legacy>
7. Introducing Context7: Up-to-Date Docs for LLMs and AI Code ..., accesso

- eseguito il giorno febbraio 1, 2026,  
<https://upstash.com/blog/context7-llmtxt-cursor>
8. How to avoid HTTP error 429 (Too Many Requests) python - Stack Overflow, accesso eseguito il giorno febbraio 1, 2026,  
<https://stackoverflow.com/questions/22786068/how-to-avoid-http-error-429-too-many-requests-python>
  9. [Bug]: Being rate limited on every request · Issue #808 · upstash/context7 - GitHub, accesso eseguito il giorno febbraio 1, 2026,  
<https://github.com/upstash/context7/issues/808>
  10. Model Context Protocol - OpenAI for developers, accesso eseguito il giorno febbraio 1, 2026, <https://developers.openai.com/codex/mcp/>
  11. Windsurf: Adding local Context7 MCP via npx breaks MCP (Marketplace refresh loop; all MCP servers stop responding) on Windows 11 #829 - GitHub, accesso eseguito il giorno febbraio 1, 2026,  
<https://github.com/upstash/context7/issues/829>
  12. Configuring MCP Servers - Cline Docs, accesso eseguito il giorno febbraio 1, 2026, <https://docs.cline.bot/mcp/configuring-mcp-servers>
  13. Claude Code + Context7 MCP Server Is a GAME CHANGER for AI Coding - YouTube, accesso eseguito il giorno febbraio 1, 2026,  
<https://www.youtube.com/watch?v=BJX6uJHlz5U>
  14. Mastering Time Awareness in AI Agents: A Deep Dive into the Time MCP Server, accesso eseguito il giorno febbraio 1, 2026,  
<https://skywork.ai/skypage/en/Mastering-Time-Awareness-in-AI-Agents:-A-Deep-Dive-into-the-Time-MCP-Server/1972171295157227520>
  15. Connect Claude Code to tools via MCP, accesso eseguito il giorno febbraio 1, 2026, <https://code.claude.com/docs/en/mcp>
  16. Claude Code failed to connect Serena in Idea CE 2025.2 but works in VS Code #635, accesso eseguito il giorno febbraio 1, 2026,  
<https://github.com/oraios/serena/issues/635>
  17. Deep Dive: I dug and dug and finally found out how the Context7 MCP works under-the-hood : r/ClaudeAI - Reddit, accesso eseguito il giorno febbraio 1, 2026,  
[https://www.reddit.com/r/ClaudeAI/comments/1muoes4/deep\\_dive\\_i\\_dug\\_and\\_dug\\_and\\_finally\\_found\\_out\\_how/](https://www.reddit.com/r/ClaudeAI/comments/1muoes4/deep_dive_i_dug_and_dug_and_finally_found_out_how/)
  18. How to Supercharge Your AI Coding Assistant with Context7 - DeployHQ, accesso eseguito il giorno febbraio 1, 2026,  
<https://www.deployhq.com/guides/context7>
  19. @upstash/context7-mcp - npm, accesso eseguito il giorno febbraio 1, 2026,  
<https://www.npmjs.com/package/@upstash/context7-mcp>
  20. The Hidden Cost of MCPs and Custom Instructions on Your Context Window, accesso eseguito il giorno febbraio 1, 2026,  
<https://selfservicebi.co.uk/analytics%20edge/improve%20the%20experience/2025/11/23/the-hidden-cost-of-mcps-and-custom-instructions-on-your-context-window.html>
  21. I built a Context7 alternative that costs 40% less with similar code quality - here are my test results : r/modelcontextprotocol - Reddit, accesso eseguito il giorno

febbraio 1, 2026,

[https://www.reddit.com/r/modelcontextprotocol/comments/1m7e797/i\\_built\\_a\\_context7\\_alternative\\_that\\_costs\\_40\\_less/](https://www.reddit.com/r/modelcontextprotocol/comments/1m7e797/i_built_a_context7_alternative_that_costs_40_less/)

22. How Ref takes advantage of MCP to build documentation search that uses the fewest tokens, accesso eseguito il giorno febbraio 1, 2026,  
[https://www.reddit.com/r/mcp/comments/1mc9uvw/how\\_ref\\_takes\\_advantage\\_of\\_mcp\\_to\\_build/](https://www.reddit.com/r/mcp/comments/1mc9uvw/how_ref_takes_advantage_of_mcp_to_build/)
23. Context7 just massively cut free limits : r/ClaudeCode - Reddit, accesso eseguito il giorno febbraio 1, 2026,  
[https://www.reddit.com/r/ClaudeCode/comments/1qblbmr/context7\\_just\\_massively\\_cut\\_free\\_limits/](https://www.reddit.com/r/ClaudeCode/comments/1qblbmr/context7_just_massively_cut_free_limits/)
24. Pricing & Plans | Context7, accesso eseguito il giorno febbraio 1, 2026,  
<https://context7.com/plans>
25. Pro plan is basically unusable : r/ClaudeAI - Reddit, accesso eseguito il giorno febbraio 1, 2026,  
[https://www.reddit.com/r/ClaudeAI/comments/1q9va21/pro\\_plan\\_is\\_basically\\_unusable/](https://www.reddit.com/r/ClaudeAI/comments/1q9va21/pro_plan_is_basically_unusable/)
26. [Bug]: Context7 does not verify ownership of a library even with context7.json in root #1277, accesso eseguito il giorno febbraio 1, 2026,  
<https://github.com/upstash/context7/issues/1277>
27. HTTP Error 429 (Too Many Requests) - How to Fix - Postman Blog, accesso eseguito il giorno febbraio 1, 2026, <https://blog.postman.com/http-error-429/>
28. Bug: MCP server context7 was unable to start successfully. · Issue #590 - GitHub, accesso eseguito il giorno febbraio 1, 2026,  
<https://github.com/upstash/context7/issues/590>
29. Reference error: Fetch is not defined (when calling resolve-library-id tool) · Issue #187 · upstash/context7 - GitHub, accesso eseguito il giorno febbraio 1, 2026,  
<https://github.com/upstash/context7/issues/187>
30. Am I missing something with the Context7 MCP hype? : r/ClaudeAI - Reddit, accesso eseguito il giorno febbraio 1, 2026,  
[https://www.reddit.com/r/ClaudeAI/comments/1oyqi8e/am\\_i\\_missing\\_something\\_with\\_the\\_context7\\_mcp\\_hype/](https://www.reddit.com/r/ClaudeAI/comments/1oyqi8e/am_i_missing_something_with_the_context7_mcp_hype/)
31. Which MCP is currently the best for refactoring or code review? : r/cursor - Reddit, accesso eseguito il giorno febbraio 1, 2026,  
[https://www.reddit.com/r/cursor/comments/1lfdmm4/which\\_mcp\\_is\\_currently\\_the\\_best\\_for\\_refactoring/](https://www.reddit.com/r/cursor/comments/1lfdmm4/which_mcp_is_currently_the_best_for_refactoring/)
32. Agentic Coding Recommendations - Hacker News, accesso eseguito il giorno febbraio 1, 2026, <https://news.ycombinator.com/item?id=44255608>