



Answers from the vendor's docs, not from memory

A scraper, a hybrid vector and keyword index, and an MCP tool that hands a coding agent cited documentation.



Sytze P.
Solution Engineer

The problem

A coding assistant only knows what was in its training data. Vendor APIs move faster than that, so the model answers with parameters that were real a year ago and are not real now. The failure is quiet. The code looks right until it runs.

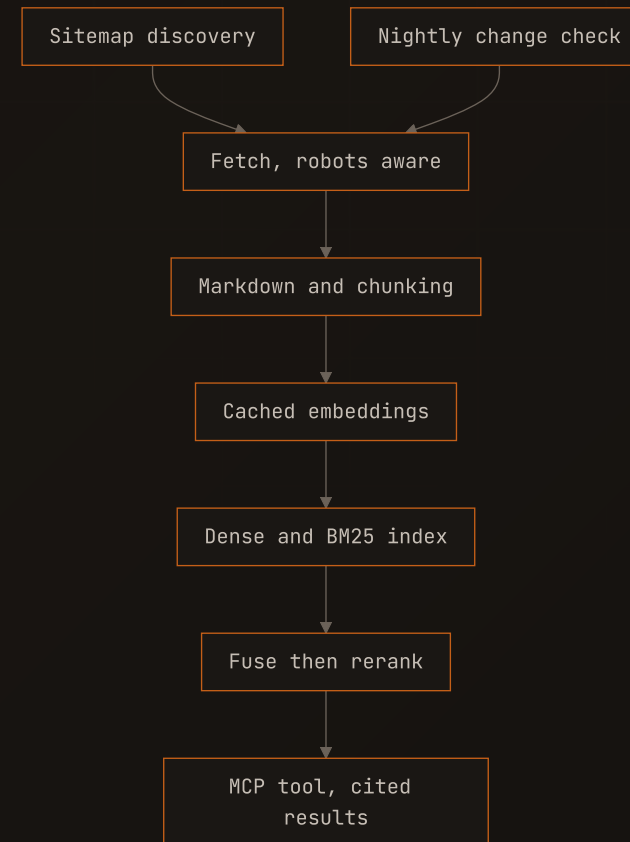
Pasting the correct doc page into the prompt works once. It stops working when the answer is spread across an API reference, a guide and a release note, and it never scales past the person doing the pasting.

- APPROACH

Docs site to agent answer

Pages come from the site's own sitemap, fetched under its robots rules, converted to markdown and split at header boundaries into chunks of about 400 tokens. A code example too long for one chunk is kept whole rather than cut in half. API documentation is half prose and half example, and a chunk that ends mid-example is worse than no chunk at all.

Each chunk is indexed twice, as a dense embedding for meaning and as a BM25 sparse vector for exact parameter names. A query hits both, the two result sets are fused, and a cross-encoder re-scores the top twenty candidates before five are returned.



- HIGHLIGHTS

How it works

- TWO INDEXES, THEN A RE-RANKER

Semantic search finds the concept and BM25 finds the exact parameter name. The fused list is re-scored by a cross-encoder before the top five go back to the agent.

- EVERY RESULT NAMES ITS SOURCE

A result carries its heading trail and the URL of the page it came from, so a claim can be checked against the document instead of taken on trust.

- NIGHTLY REFRESH THAT PROBES BEFORE IT FETCHES

Conditional requests mean only changed pages are re-embedded. Pages that drop out of the sitemap are marked stale and deleted after a week.

Results & impact

cites the source document

heading trail and URL on every result

returns nothing when unsure

confidence gate on the reranked top score

re-indexes only what changed

conditional requests, then a content hash

• STACK

Tech stack

Python

scrape, index and serve

Qdrant

dense and sparse vector store

sentence-transformers

embeddings and cross-encoder

MCP

tool interface for the agent

Playwright

fallback for JS-rendered pages

FastAPI and APScheduler

scheduled refresh service

pytest

unit and integration tests



Sytze P.

Solution Engineer

A scraper, a hybrid vector and keyword index, and an MCP tool that hands a coding agent cited documentation.
