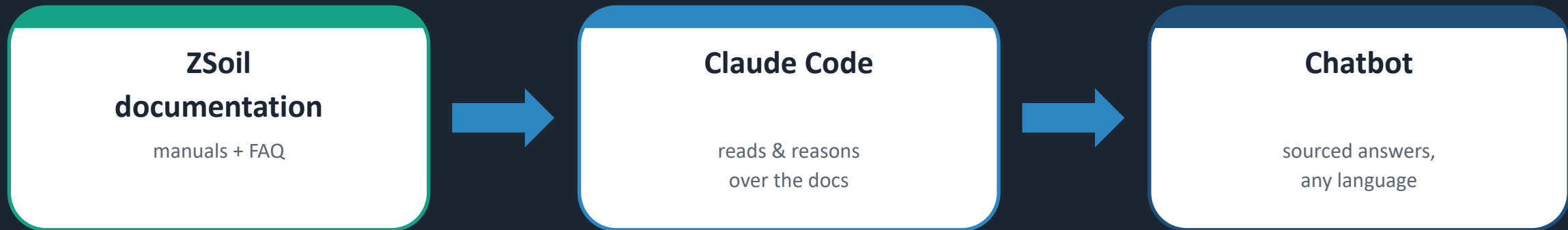


Prototyping AI Assistance for ZSoil

Basile Zimmermann, Thomas K. Zimmermann

Claude Code, informed by the ZSoil docs, answering users through a Chatbot



An engineer asks in plain language — the answer comes back grounded in the manual and **citing its source**.

The documentation indexed so far

PDF

Getting Started

First steps & workflow basics for new ZSoil users.

PDF

Virtual Lab

Calibration and element/material testing guidance.

PDF

HSS model

The Hardening-Soil Small-strain constitutive model.

FAQ

FAQs

The online zsoil.com knowledge base (Q&A).

Extensible by design: adding a document means re-running the local indexer and uploading the new index — no code change, no new language model, no re-training.

Demonstration — the assistant from a cold start

1 · Initial state

empty chatbot, language selector, login done



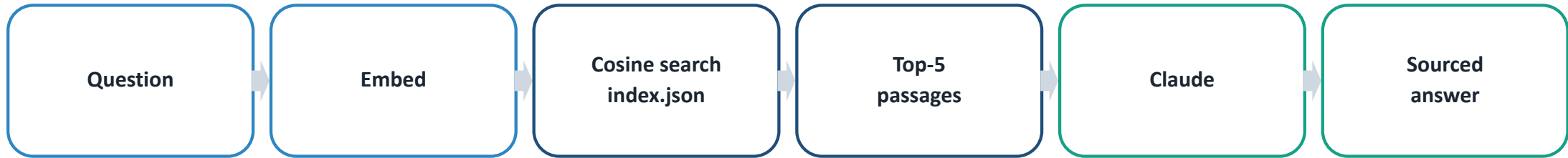
2 · Ask a real question

watch retrieval → sourced answer → cost

“How to run a consolidation analysis?”

Expect: a concise procedure + Sources (document, page / FAQ category) + the cost of the query.

How it works, on one slide



Indexing — local, one-off

- PDFs + FAQ → cleaned, ~500-token chunks
- OpenAI embeddings (3-large, 1536-dim)
- **index.json** — vector + source + kind/page

Runtime — PHP on the server

- Same embedding model & dimension as index
- Cosine → top-5 → Claude (Sonnet 4.6)
- Multilingual translate-for-search trick

Built deliberately small: plain HTML/CSS/JS + PHP · raw HTTP (cURL), no SDK / no Composer / no build · secrets & index live only on the server.

Associated costs — 100% pay-per-use

- **No fixed cost** — you pay only per question asked.
- Per question with **Sonnet 4.6**: ~3 000 input + ~300 output tokens ≈ **1.35 ¢**.
- Cost driven mainly by **answer length** — the prompt enforces concision.
- The model is a **one-line config switch**.

Model	¢ / question	≈ USD / yr	≈ CHF / yr
Haiku 4.5	~0.45	~165	~140
Sonnet 4.6 (current)	~1.35	~490	~420
Opus 5	~2.25	~820	~700

Reference: 100 questions/day ≈ 36 500/year. Multilingual adds ~\$15/yr. July 2026.

Levers to cut cost: cheaper model · fewer retrieved passages (lower top-k) · shorter answers.

Private by design

Access control

Shared secret over TLS, PHP session; ask.php returns 401 unauthenticated. No per-user identity — deliberately light.

Not indexable

robots.txt Disallow: / · X-Robots-Tag: noindex, nofollow, noarchive · Options -Indexes.

Files locked down

.htaccess denies direct GET on *.json, config*.php and app.html; the app is served by readfile() after auth.

Secrets off git

API keys and index.json are git-ignored and FTP-only — never in the repo, never in the deploy artefact.

Data flow: the question and the five retrieved passages leave the server over TLS to OpenAI (embedding) and Anthropic (generation). Neither trains on API data by default, but manual excerpts do transit.

Improvement possibilities

- Index **more documents** (full manuals, theory & reference guides).
- Index **figures & screenshots** (vision models) instead of linking out.
- **Conversation memory** — follow-up questions in context.
- **Feedback loop** — thumbs up/down to tune retrieval.
- **Caching** of frequent questions to cut cost further.
- **More languages**, no re-indexing needed.
- Deeper **ZSoil integration** — link answers to model entities.

Thank you

Questions welcome — and feel free to try the online version.

potiron.ch/zsoil

Password: zace

Works on phone, tablet and laptop — nothing to install.