

Groundwork — AI Solution Architecture Advisor

An architecture advisor that interviews you first, then recommends with citations from real technical books.

BUILT WITH

RAG

AWS Bedrock

Pinecone

FastAPI

Next.js

Solution Architecture

Tool Calling

6

steps from interview to cited recommendation

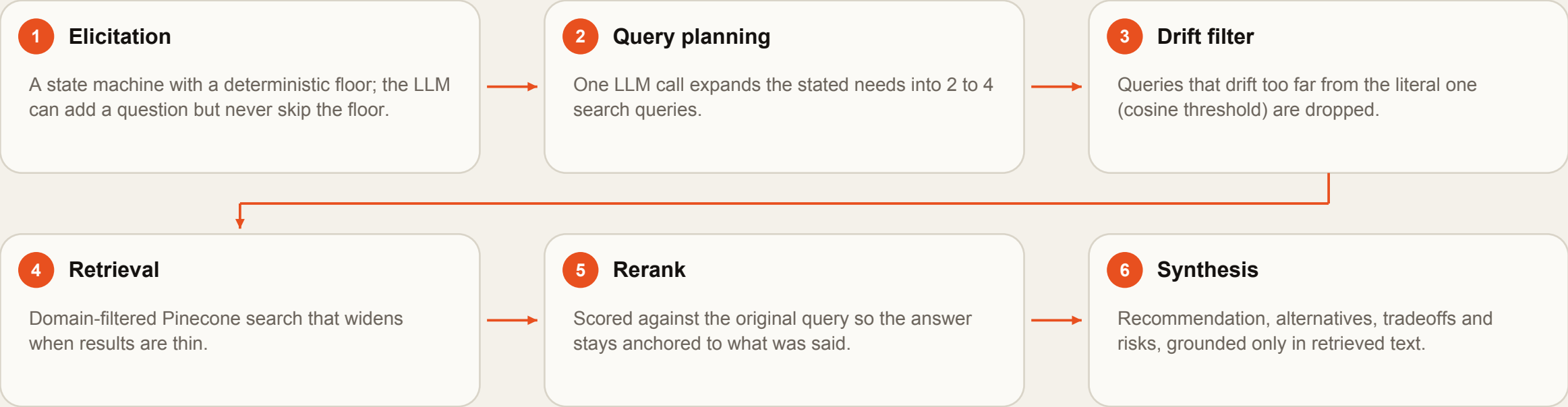
2

technical books per knowledge domain

2 weeks

built solo, live on Render and Vercel

How it works



WHAT EACH PART DOES

Elicitation state machine	Asks questions until enough constraints are known. A hard minimum and maximum decide when to stop.
Query planner	Turns the stated needs into 2-4 search queries, including concepts implied but never named.
Drift filter	Drops any generated query whose meaning strays too far from what the user said.
Retrieval and rerank	Pinecone search, then a rerank against the user's original wording.
Synthesizer	Writes the recommendation, alternatives, tradeoffs and risks, with citations to book and chunk.

Key decisions

01

A fixed floor decides when to stop asking

The language model can ask one more question but can never skip the minimum number of turns and populated fields. An "I have enough" on a nearly empty state is overridden.

02

Expand the search, then check for drift

A scenario that needs updates within a second never says "streaming", so a literal search misses the right content. Generated queries are kept only if they stay close in meaning to the literal one.

03

One tagged index, with a fallback

All content sits in one Pinecone index filtered by domain. If a filtered search returns too little, it widens to the whole corpus instead of showing "no matches".

04

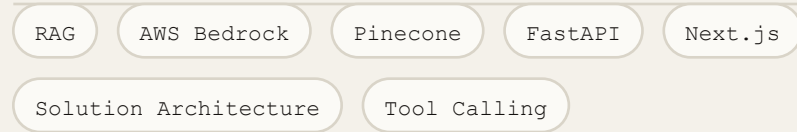
Rerank against the original wording

Expansion widens what is found; reranking keeps the final answer anchored to what the user actually said.



Stack, status and what is next

BUILT WITH



KNOWN LIMITS

- Only 2 of 5 architecture domains (ai-ml, data) have real book content so far.
- No automatic hallucination check; citations give traceability but nothing flags a wrong claim.
- Free-tier hosting means cold starts after inactivity and no uptime guarantee.
- Tested by the builder against real infrastructure, not yet by end users.

