**🗺 Step 1 — Local‑Only Architecture Blueprint**

We’ll sketch this first so we know exactly how each piece connects:

**Core Components**

* **FastAPI backend** — main API layer, handles requests from Discord bot and any local tools.
* **Cerebras client** — LLM calls, isolated in its own helper/module for easy swapping or upgrades.
* **Chroma** — local vector DB for code/document embeddings and retrieval.
* **Discord bot** — user‑facing interface, sends queries to FastAPI, receives responses.
* **Config layer** — .env + config.py for all environment variables and defaults.

**Flow Example**

Code

Discord User → Discord Bot → FastAPI Endpoint → Cerebras Client

↘

Chroma (retrieval)

**📝 Step 2 — Codebase Comment Mapping**

We’ll go file‑by‑file and:

1. **Add high‑level docstring** at the top of each file explaining its role in the architecture.
2. **Inline comments** for tricky logic, API calls, or config parsing.
3. **TODO markers** for incomplete or placeholder sections.
4. **Refactor notes** where structure could be improved later.

**📋 Step 3 — Roadmap Extraction**

Once comments are in place, we’ll:

* Pull all TODO and NOTE markers into a single **development checklist**.
* Group them by **priority** (must‑fix now vs. nice‑to‑have later).
* Tie them back to the architecture diagram so we see dependencies.