AI Engineer Project Blueprint: AI Code Review Bot

Goal

Build an open-source, intelligent code review system that uses local or hosted LLMs to suggest improvements to code in GitHub pull requests - making you a practitioner in LLM serving, prompt engineering, retrieval, model orchestration, and infrastructure.

Tech Stack Breakdown

LLM: Mistral, CodeLlama, StarCoder2 - Local code-focused language model

Serving: text-generation-webui, vLLM, FastAPI - Runs model + exposes an API

Orchestration: LangChain or LlamaIndex - Prompts, chains, tools

Embedding Store: FAISS or Chroma - Store code snippets, PR history

Frontend/API: FastAPI, GitHub API, Pydantic - Handles webhooks and API logic

Containerization: Docker - Local or cloud deploy

(Optional) Fine-tuning: LoRA + PEFT - Optional customization

Minimum Viable Version

[x] GitHub PR webhook

[x] Extract code diff + metadata

[x] Run prompt with LLM locally (or via Hugging Face)

[x] Post suggestions to PR

Intermediate Goals

[ ] Embed past PRs + codebase context (FAISS + LangChain Retriever)

[ ] Tool-enhanced agent (e.g., pylint, black)

[ ] Model selection (GPT-4 vs Mistral)

AI Engineering Extensions

[ ] Fine-tune on code review history using LoRA

[ ] Measure hallucination rate / suggestion quality

[ ] Log responses for evaluation (LangSmith or Langfuse)

Folder Structure

ai-code-review-bot/

app/

main.py

review_engine.py

review_agent.py

vector_store.py

github_client.py

prompts/

review_prompt.txt

models/

mistral/

Dockerfile

requirements.txt

README.md

Learning Resources

Local LLMs:

- Mistral/LLaMA: github.com/oobabooga/text-generation-webui

- Fast inference: github.com/vllm-project/vllm

LangChain/LlamaIndex:

- LangChain docs: docs.langchain.com/docs

- LlamaIndex indexing: docs.llamaindex.ai

Fine-Tuning:

- PEFT/LoRA: huggingface.co/blog/peft

- QLoRA: github.com/artidoro/qlora

Prompt Engineering:

- dair-ai Prompt Guide: github.com/dair-ai/Prompt-Engineering-Guide

Why This Project Gets You Hired

Real-world LLM application

Full-stack AI dev: model, logic, infra

Shows prompt design, vector store, RAG, agent use

Clear portfolio value, demo-ready

Extensible (CI/CD hooks, dashboards, eval)