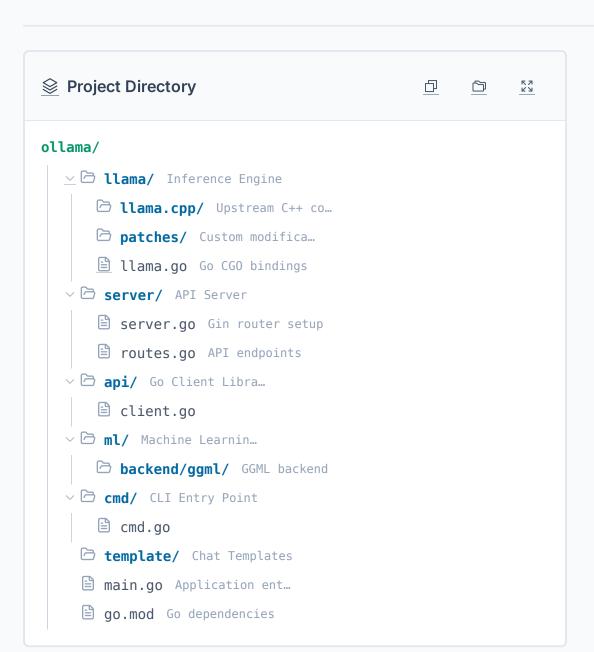
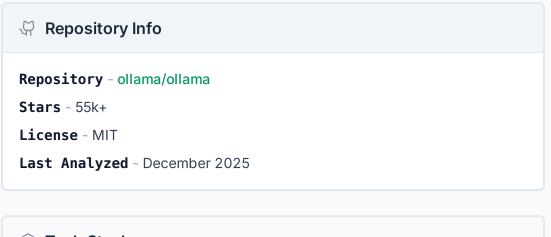


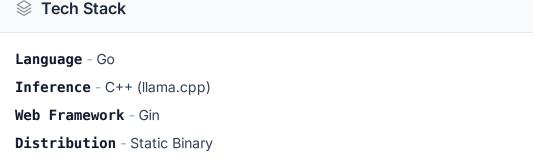
## **Ollama Project Structure**

The easiest way to get up and running with large language models locally. A Go application that wraps and manages the C++ inference engine. Updated 2025-12-30

☑ PNG PDF PDF ☐ Copy </> Prompt #ollama #go #cpp #llm #ai #inference #llama.cpp







## Architecture Notes

Ollama is a Go wrapper around the llama.cpp library. It uses CGO to call into the C++ code for model inference. The Go layer handles the API server (using Gin), model management (pulling from registry, verifying hashes), and the CLI interface. It essentially turns raw model weights into a usable REST API.

## 

llama/ - Contains the C++ code for running LLMs. It embeds llama.cpp and applies custom patches to support specific hardware or features.

**server/** - The HTTP server implementation. It accepts JSON requests from clients and translates them into calls to the inference engine.

ml/ - Abstracts the machine learning backend details, allowing Ollama to potentially support other backends in the future.

## Why This Structure?

Ollama is the standard for local LLM inference. Its architecture prioritizes ease of use: a single binary that handles everything from downloading models to running them on your GPU.