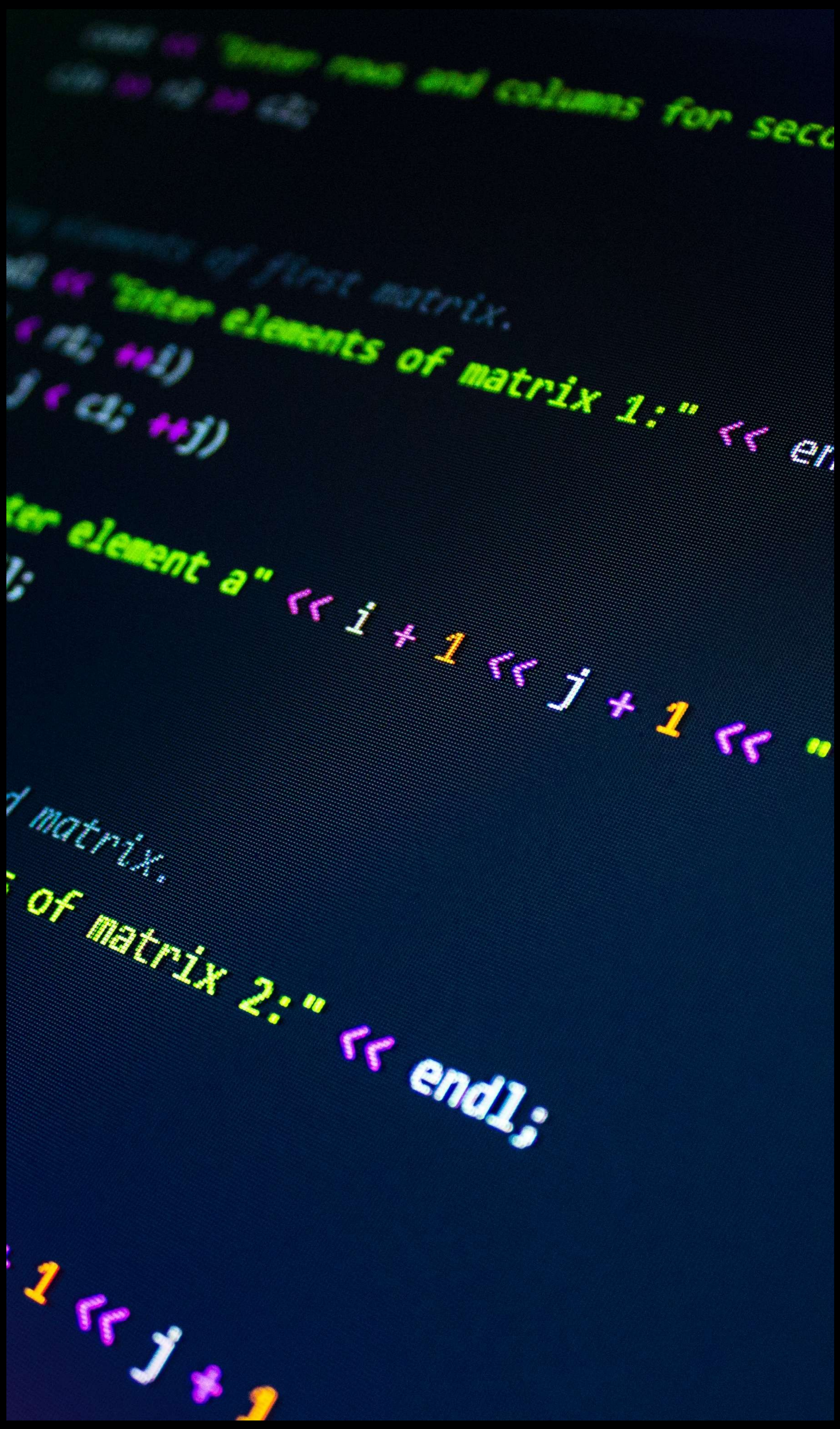
Unified LLM Deployment Platform

All-in-one local deployment for LLMs

for WHO?

for EVERYONE !



WHAT it is

Package the LLM model, frontend UI, and reasoning logic into a **single Docker containe**r.

HOW it works

Ollama – Lightweight local LLM
LangChain – Logical orchestration and API routing
Gradio – Visual interface
FastAPI – Backend bridge
Docker – Unified container

WHY choose us

Unified Deployment

All components in one container

Local Privacy Protection

Fully deployable on local machines

Low-Latency Inference

Fast response time

Extensible Structure

Easily expandable with LangChain plugins; customizable UI and LLM model selection

Education-Friendly

Ideal for LLM-based teaching, experimentation

Check our video below for more information!
https://youtu.be/ZCEMsY0pTZk