PersonaForge Architecture

Frontend

Backend

Gradio UI
Auth | Create | Chat
Upload | AI-to-AI | Export

LLM Models
llava | gemma3
llama3.2 |
gemma:2b
tinyllama

HTTP Requests

Inference

Auth Module
FastAPI Users
JWT Tokens
Access Control

Auth

FastAPI Backend
API Endpoints
/auth | /persona
/ask | /upload
/chat | /export

Persona Mgmt

Persona Engine
Prompt Generation
Storage
Versioning

DB Access

RAG

RAG Engine
Document Upload
Embedding
Retrieval

SQLite DB
Users
Personas

Vector DB
RAG Data

Vector Storage

Logs

Logs
Chat History
JSONL

Storage