Franz Ayestaran - AI-Augmented Software Engineer

franz@ayestaran.dev • https://franz.ayestaran.dev • https://linkedin.com/in/ayestaran • Tel: 07940901877

SUMMARY

AI-Augmented Software Engineer with a 30-year engineering trajectory, now specialising in **LLM training systems, interpretability tooling, agent orchestration, and closed-loop AI ecosystems**. I design and build **full-stack AI platforms** — from dataset ingestion to GPU-accelerated training, from 3D interpretability to cloud inference — with a focus on **reproducibility, transparency, and developer-centric tooling**.

My work spans **Apple Silicon (MLX), NVIDIA CUDA**, and bespoke Linux GPU servers. I build systems that make advanced AI *usable*, *inspectable*, and *deployable* — the kind of infrastructure that accelerates teams, not just models.

"The LLM Training Dashboard is a tool I built to make modern artificial intelligence easier to understand, experiment with, and improve."

CORE AI COMPETENCIES

- **LLM Training & Fine-Tuning:** LoRA, QLoRA, full-fine-tuning, GGUF export, tokenizer pipelines, dataset validation.
- **GPU Compute:** MLX (Apple Silicon), CUDA (NVIDIA), RTX 5060 Ti (Vast.ai), VRAM-aware training optimisation.
- **AI Infrastructure:** Ollama cloud inference, Caddy routing, Cloudflare tunnels, Terraform, containerised GPU workloads.
- **Interpretability & Visualisation:** Embedding Galaxy, Brain Atlas, tensor heatmaps, attention flow, neuron concept discovery.
- **Agent Systems:** Real-time task orchestration, skill routing, multi-agent pipelines, introspection dashboards.
- **Full-Stack Engineering:** Python, PyTorch, FastAPI, Flask, Node.js, Next.js, SQLite, Redis, Docker, Nginx.
- **Mobile AI Integration:** iOS, Android, Wear OS, WatchOS, Objective-C, Swift, Kotlin.

FLAGSHIP AI PROJECTS

LLM Training Dashboard — Full AI Training & Interpretability Platform

Creator & Architect • 2026–Present • https://dashboard.llmtraining.dev/

A complete end-to-end LLM training environment supporting MLX and CUDA backends. **Impact & Highlights:**

- Built a **full training pipeline**: dataset ingestion → preprocessing → tokenisation → fine-tuning → GGUF export.
- Designed **9-layer interpretability stack**, including Embedding Galaxy, Brain Atlas, attention flow, and neuron concept discovery.
- Implemented **real-time telemetry** (CPU/GPU/RAM/VRAM, I/O, network, training metrics).
- Added **Hugging Face sync**, LoRA import, auto-quantisation, and reproducible training profiles.
- Created **3D interactive visualisations** enabling non-experts to understand model internals.
- Runs on **Apple Silicon M1 Max** and **NVIDIA RTX 5060 Ti** GPU server.

This project demonstrates the ability to design and build a complete AI system from scratch — combining software engineering, machine learning, data handling, and user-centred design.

Machine & Deep Learning Console — Multi-Phase ML/DL Training Environment

Creator • 2026–Present • https://machinedeeplearning.llmtraining.dev

A unified platform for classical ML, deep learning, and transformer-based training.

- Automatic dataset detection (CSV, JSON, TXT, vision folders).
- Supports MicroLLaMA, MiniLLaMA, ViT, CNNs, MLPs, SVMs, Decision Trees.
- Real-time GPU/CPU telemetry, loss curves, accuracy, throughput.
- ONNX export, GGUF conversion, inference runners.

ClawForgeAI — Real-Time Agent Orchestration Dashboard

Creator • 2026–Present • https://clawforgeai.llmtraining.dev

A server-hosted multi-agent runtime with:

- Live task orchestration and queueing
- Agent/model routing transparency
- Skill introspection with latency/error metrics
- Persistent task history for reproducibility
- Extensible plugin/skill framework

Ollama Cloud Inference Layer — Secure, Reproducible AI Infrastructure

Architect • 2026–Present

Built a production-grade inference layer using:

- Linode + Vast.ai GPU nodes
- Caddy routing + Cloudflare tunnels
- GGUF model hosting
- Cross-platform Apple Silicon ↔ CUDA compatibility
- Low-latency distributed inference

AI Village — 2D Game Engine with AI-Driven NPCs

Co-Creator • 2025–Present • https://aivillage.ayestaran.dev

A core component of the closed AI ecosystem, providing real-time behavioural evaluation for models trained in the LLM Training Dashboard.

Key Capabilities:

- **LLM-Driven NPCs:** Autonomous agents with reasoning, memory, emotional state modelling, and context-aware behaviour.
- **Closed-Loop Pipeline:** Direct integration with the LLM Training Dashboard → GGUF export → Ollama Cloud Inference → in-world evaluation.
- **Interactive Simulation:** 2D game environment for testing alignment, emergent behaviour, and human-AI interaction.
- **Secure & Reproducible:** Fully self-hosted on Debian 12 (Nginx, Node 20, PHP 8.2) with no external API dependencies.
- **Research-Grade Testing:** Supports controlled experiments on LoRA adapters, model updates, and interpretability findings.

TokenCalc — Dataset Token Estimator for LLM Training

Creator • 2026–Present • https://tokencalc.llmtraining.dev

Uploads text/JSON/JSONL, estimates token volume, multiplies across epochs, compares against OpenAI/Anthropic pricing.

AI-AUGMENTED ENGINEERING PHILOSOPHY

You explicitly describe your workflow as:

"I integrate advanced AI tools directly into my software engineering workflow — not to replace engineering judgment, but to amplify it."

This positions you as the kind of engineer who builds **AI-accelerated engineering systems**, not just AI-powered features — a distinction that leaders like Altman care deeply about.

Published iOS Applications (Apple App Store)

https://apps.apple.com/gb/developer/zonk-technology/id726984885

https://secure.zonktechnology.com/itracker

Independent Developer — Zonk Technology Active Apple Developer Since 2009

You have **six** shipped and maintained iOS apps — all designed, developed, and published independently. These demonstrate long-term product ownership, UI/UX capability, and full-stack mobile engineering.

Published Apps:

- **iTracker (2013)** — Personal tracking utility
- **iWebSearch (2013)** — Customisable web search interface
- **iWeatherMap (2014)** — Weather visualisation and mapping
- **iColourPicker (2014)** — Colour selection and palette tool
- **iPicSolve® (2014)** — Image-based puzzle solver
- **iTime Table (2019)** — Scheduling and timetable management

SELECTED TECHNICAL SKILLS

AI/ML: PyTorch, Transformers, MLX, CUDA, GGUF, LoRA/QLoRA, ONNX, PCA, RAG, MCP **Backend:** Python, FastAPI, Flask, Node.js, PHP, C#, SQL Server, MySQL, SQLite **DevOps:** Docker, Terraform, Nginx, Caddy, Cloudflare, Linode, Vast.ai **Frontend:** Next.js, React, Tailwind, HTML/CSS/JS **Mobile:** Swift, Objective-C, Kotlin, Android Studio, Xcode **Legacy Expertise:** Pascal, Delphi, VB6, MS-DOS systems, embedded electronics

EARLY CAREER

Your early career shows a rare depth in **systems programming, embedded development, and low-level engineering** — from PIC microcontrollers to DOS utilities to Windows CE enterprise tools. This gives you a foundation most modern AI engineers simply don't have.

Examples include:

- **GPS Telemetry Recorder** (VB6, SQL Server, NMEA parsing)
- **PDA Digital Jobsheet** (Windows CE, Embedded VB)
- **Tooldisk / Burn-In / DOSIMAGE** (Turbo Pascal, 8086 Assembly)
- **BBC Micro Spritewise** (1985, BASIC)

This history signals **longevity, adaptability, and deep technical roots** — traits highly valued in frontier-model environments.

EDUCATION

MSc Artificial Intelligence (In Progress) University of Hertfordshire • 2026–2028 Unconditional offer accepted.