
AI ENGINEER / RESEARCHER / BUILDER

Miguel Penha Reis

AI from weights to systems.

I research and engineer language models, latent reasoning architectures and local AI systems, then build the software that turns those ideas into usable products. My work moves between model research, agent infrastructure, desktop software, realtime systems and local-first tooling.

LANGUAGE MODELS

Fine-tuning · quantization · local inference

AGENT SYSTEMS

Isolation · policy · interoperability

LOCAL AI

GPU-aware workstation software

RESEARCH

Latent reasoning · recurrent compute

[GITHUB](#)

[HUGGING FACE](#)

[VELUM AI](#)

Projects

LOCAL CODING MODEL

VELUM-Coder

9B coding-focused fine-tune released for local inference with GGUF quantizations and a domain-specific importance matrix.

GGUF · llama.cpp · Imatrix · LoRA

OPEN PROJECT ↗

PRIVATE AGENT RUNTIME

Veil Agent

Windows agent runtime that separates model reasoning from host authority using Hyper-V isolation, deterministic policy, auditing and privacy-first networking.

Python · PyQt6 · Hyper-V · WFP

OPEN PROJECT ↗

LOCAL API COMPATIBILITY

OmniClaw

Local bridge that exposes OpenAI Chat Completions, Responses and Anthropic Messages APIs for Claude Code, Codex, OpenCode and compatible clients.

Node.js · Streaming · Tool Calls

OPEN PROJECT ↗

EMULATION PLATFORM

PKG_DoW

Unified Windows launcher for compatible PS4/PS5 content with modular cores, PKG import, Vulkan tuning, diagnostics and frame-time-focused AutoBoost policies.

Windows · Vulkan · Desktop UX

OPEN PROJECT ↗

REALTIME COMMUNITY CLIENT

Open DC

Open-source community client with accounts, servers, DMs, realtime chat, WebRTC calls, screen sharing, forums, bots and webhooks.

Flask · SQLite · Socket.IO · WebRTC

OPEN PROJECT ↗

LOCAL GENERATIVE AI

Forge3D

Local image-to-3D generation engine producing refined, textured GLB assets through a GPU-aware pipeline.

TRELLIS · UltraShape · FastAPI · Three.js

OPEN PROJECT ↗

Projects

AI RESEARCH RUNTIME

LeBRON

Experimental recurrent latent computation for frozen decoder models through a structured J-Space adapter, competing branches, verification and adaptive halting.

PyTorch · Transformers · Latent Reasoning

OPEN PROJECT ↗

AI EDUCATION PLATFORM

Projeto Lunar

Educational ecosystem connecting students, teachers and families through AI-assisted learning tools, dashboards, media and school workflows.

Python · Flask · Gemini · Socket.IO

OPEN PROJECT ↗

CREATIVE AI

Lyra Engine

Local AI music-production environment for generation, songwriting assistance and GPU-aware model orchestration.

ACE-Step · Ollama · Flask · Audio AI

OPEN PROJECT ↗

DESKTOP SOFTWARE

AGNER Browser

Experimental desktop browser with isolated profiles, persistent sessions, local data and custom browser UX.

Python · PyQt6 · QtWebEngine · SQLite

OPEN PROJECT ↗

RETRIEVAL-AUGMENTED AI

IA RAG Nexa

Local knowledge assistant combining semantic retrieval, BM25/FTS, rank fusion and traceable source references.

Flask · Ollama · Chroma · BM25 · SQLite

OPEN PROJECT ↗

LANGUAGE MODELS

Nexora Models

Compact coding and multimodal models focused on debugging, tool use, structured reasoning and efficient local inference.

Qwen · Gemma · GGUF · llama.cpp

OPEN PROJECT ↗

Build the model. **Build the system.** **Understand both.**

The portfolio intentionally crosses research and product engineering. Understanding the model matters, but so do the runtime, interface, deployment constraints, privacy boundaries and user experience around it.

Miguel Penha Reis

AI ENGINEER · RESEARCHER · BUILDER

[**github.com/guell11**](https://github.com/guell11)

[**huggingface.co/guell00**](https://huggingface.co/guell00)