

CAPABILITIES & SELECTED WORK

Autonomous AI systems, built and operated.

One senior engineer who designs, ships, and **runs** live autonomous systems in production — multi-agent platforms, real-time trading infrastructure, multimodal AI.

```
$ systemctl status --fleet
```

multi-agent workforce	● running
live trading infrastructure	● running
multimodal & XR AI	● running

Andrei Gireada

— Senior Software Engineer · 26 years

Bucharest, Romania (EU) · solo, senior, fast

andrei@daemonworks.dev · daemonworks.dev · github.com/NrSlionHeart

Build **and** operate. Solo.

26 years shipping software — from game development to distributed AI. I don't hand over a repo and leave: I architect the system, ship it in production-grade code, wire the observability, and **keep it running**. Most of what follows is live in production right now, built and operated by one person.

CORE COMPETENCIES

- AI / LLM systems
- Multi-agent architectures
- Python backend / infra
- Trading & market systems
- Browser & web automation
- Unity / XR development

HOW I ENGAGE

- Fixed-scope system builds

outcome-based
- Operate & improve retainers

monthly
- Fractional AI / systems lead

part-time
- Direct access — no juniors

just me

APPROACH

- Self-hosted where it counts

cost + control
- Self-healing, observable

production-grade
- Rigor over hype

validated

26 yrs shipping software	6+ self-hosted LLMs, 1.7B-32B	~96 supervised services live · self-healing	10+ real-time market venues, one backbone
70+ automated trading strategies built	67k+ companies mapped by autonomous agents	4 modalities: vision · voice · text · XR	1 engineer — build and operate

TECHNICAL STACK

AI / ML

- Ollama · vLLM · SGLang
- LoRA / Unsloth fine-tuning
- GGUF / llama.cpp
- Qwen2.5-VL · Florence-2 · YOLOv8
- Whisper · Piper TTS
- DeBERTa-v3 · Sentence Transformers
- pgvector · Neo4j

AUTOMATION & TRADING

- Playwright · Selenium
- Set-of-Marks vision UI
- MTL5 / MetaTrader
- QuantConnect LEAN
- LightGBM · CatBoost · HMM

BACKEND & DATA

- Python (asyncio)
- FastAPI · Pydantic
- PostgreSQL · MySQL · SQLite
- QuestDB
- WebSocket / Unix sockets
- Docker · systemd

FRONTEND & XR

- React 18 · TypeScript
- Next.js
- WebGL
- Unity · C#
- PICO XR SDK
- Stripe · nginx

Systems that run the whole function.

MULTI-AGENT PLATFORM

● operating

Autonomous AI Agent Workforce

A self-hosted fleet of specialized AI agents that runs an entire B2B data & outreach operation end-to-end — solo-built, solo-operated.

A **three-tier distributed platform**: a real-time CRM/command layer (60+ API endpoints, live WebSocket feeds), a self-healing task dispatcher with heartbeat monitoring and automatic recovery, and a horizontally-scalable fleet of GPU workers. On top — a **vision-driven web-automation agent** that reads pages the way a person does (Set-of-Marks visual grounding + Qwen2.5-VL + Florence-2 + YOLOv8), **fine-tuned language models** (LoRA → GGUF → Ollama, tiered 1.7B–32B), and a **real-time voice pipeline** (streaming STT → intent → dialog state machine → low-latency TTS).

67,000+ companies & 23,500+ people mapped 6+ self-hosted LLMs fine-tuned on 600k+ examples
multi-region GPU fleet · 24/7

Python · FastAPIReact 18 · TypeScriptQwen2.5-VL · Florence-2 · YOLOv8LoRA/Unsloth · Ollama · vLLM · SGLangFaster-Whisper · FastPitchPlaywright · MySQL

GROUNDING LLM SYSTEMS

● operating

Grounded AI Analyst Copilot

Rebuilt a hallucinating chatbot into a grounded, tool-using analyst — on a self-hosted model at zero frontier-API cost.

Diagnosed the failure at source level (seven ranked root causes across every AI-text surface), then engineered the fix: a **four-layer PostgreSQL grounding store** (append-only raw → normalized entities → market snapshots → derived consensus) with a version-keyed answer cache, a **self-hosted 32B reasoning model**, and a **mode-routing orchestrator** that dispatches each query to a deterministic scanner, a tool-calling analyst grounded on the store, or a grounded news engine — every answer returning **inline citations**.

grounded over 3 live data feeds 10 grounding tools ~3-7s answers · \$0 API cost
15-min refresh · systemd timers

Next.js · React 19 · TypeScriptPython · PostgreSQL / JSONBOllama · Qwen3-32Btool-calling · inline citationsnginx · systemd

Real data. Real stakes. Fail-safe.

AUTONOMOUS TRADING PLATFORM

● operating

Research-to-Execution Trading Stack

A full quant-desk stack — market-data backbone, alpha, execution, and institutional-grade validation — built and run by one engineer.

A centralized **async market-data backbone** unifying 10+ live venues into a time-series tick store, with self-healing "half-open feed" watchdogs that catch frozen-but-alive feeds. A **calibrated alpha ensemble** fed by a four-model regime engine, order-flow, and an LLM news pipeline. A **multi-strategy execution engine** (3,400+ lines) running strategies off independent sockets with hard, no-override risk gates and wallet-anchored P&L reconciliation. And an **institutional validation harness** (Deflated Sharpe, PBO/CSCV, purged walk-forward CV) that runs daily in production — and honestly refuses to certify strategies that don't survive it.

~96 supervised services · 0 failed · 49-96-day uptimes · 200M+ ticks stored · ~465ms median signal-to-fill

Python asyncio

QuestDB · PostgreSQL

LightGBM · CatBoost · HMM

Platt / isotonic calibration

Qwen-32B news pipeline

systemd · nginx

QUANT INFRASTRUCTURE

deep domain

Algorithmic Trading Systems (10+ yrs)

Neural-net Expert Advisors, cross-market arbitrage engines, and tick-level data pipelines — built end to end.

70+ **Expert Advisors** and 15+ custom indicators, including from-scratch **deep-neural-network** and **LSTM** implementations bridged to external ML; **five cross-market statistical-arbitrage engines** (dual-listed ADR/HK with session sync + FX, ratio and cointegration stat-arb with parameter optimization); and a tick-level data pipeline (multi-source ingestion, GMT alignment, LEAN-format conversion) feeding rigorous, lookahead-free backtests.

70+ EAs · 15+ indicators · 5 arbitrage engines · 90+ analyses for one strategy

MT5 / MT4 · MetaTrader

Python · pandas · numpy

QuantConnect LEAN

LSTM / DNN · socket IPC

Rigor, and reach across modalities.

KNOWLEDGE GRAPH + LLM

client engagement

AI Backbone for Strategy Validation

A self-hosted AI system that reasons over a trading knowledge graph and auto-reviews strategies for leakage, overfitting, and look-ahead bias.

A **six-service, one-command Docker stack**: a Neo4j knowledge graph modeling the domain (strategies → signals → trades), a pgvector semantic-search service, and an **LLM proxy** (local Mistral 7B, versioned YAML prompts, Pydantic-validated structured verdicts) that scores each strategy's hypothesis and risk and emits typed red flags — with full observability (Prometheus / Grafana / MinIO audit) and security hardening (PII redaction, prompt-injection blocking).

8 entity + 8 relationship types 384-dim semantic search every LLM call audited

Neo4j · APOC

pgvector · Sentence Transformers

FastAPI · Ollama / Mistral 7B

Prometheus · Grafana · MinIO

MULTIMODAL & XR

client engagement

XR Studio — AI Communication Coach

Real-time multimodal AI coaching across browser, mobile, and VR — one training system with live feedback.

A **real-time pipeline** that reads facial expression, gaze, voice stress, pacing, and gestures from a single webcam and streams them over WebSocket against **WebGL avatars** with LLM-driven conversation and adaptive difficulty; a **parallel VR delivery** on standalone headsets (solving external-texture rendering onto the composition-layer WebView) closing a bidirectional voice loop (STT → analysis → **20–30ms** local TTS); and an enterprise analytics layer with a natural-language briefing agent and scenario builder.

3 delivery modes: browser / mobile / VR 7+ real-time signals, one webcam sub-**30ms** local TTS

React / Next.js · WebGL

real-time computer vision

Whisper · Piper TTS

Unity · C# · PICO XR SDK

WebSocket streaming

Native to cloud, desktop to headset.

FULL-STACK PLATFORM

client engagement

Signal Automation Platform

A native desktop + SaaS platform that turns free-text trading signals into live broker orders, with an LLM fallback parser.

A **C++17 / MFC Windows client** integrating Telegram's TDLib (user-account auth with 2FA, channel monitoring, dedup) and a signal parser (symbol normalization, multiple take-profits, stop detection); a **broker layer behind a pluggable adapter interface** (real-time price streaming, session/token management, order placement); a full **PHP/MySQL SaaS** around it (auth, Stripe billing, licensing, feature flags, audit); and an **LLM normalizer** (text + image/OCR) as a fallback for signals regex can't parse.

50+ file native desktop app pluggable broker adapters full SaaS + licensing LLM text + OCR fallback

C++17 · MFC · TDLib

PHP · MySQL · Stripe

OpenAI · WebSocket

— ALSO BUILT & OPERATED

AI · signals

AI Financial-News Signal Engine — LLM-in-the-loop pipeline scoring the news firehose into signals, delivered to trading platforms 24/7.

Trading

Prediction-Market Trading Suite — 23k-line async system running five concurrent strategies over sub-second market data.

AI · agents

Local-First Agent Workforce — six role-specialized agents running entirely on a single consumer GPU.

XR

VR AI Coaching Trainer — standalone PICO headset app with an on-device bidirectional voice coach.

Let's build something that runs itself.

If it can run without a human, I can build it — and keep it running.

andrei@daemonworks.dev

daemonworks.dev

github.com/NrSlionHeart

Engagements typically run **\$3k-\$20k / month**, sized to the system — fixed-scope builds or operate-and-improve retainers. One person replacing a team, and staying to run it.