IVAN CHARAPANAU

AI SYSTEMS ARCHITECT & ADVISOR

Agentic Infrastructure | LLM Platforms | Local Inference

+48 572 729 429

av@av.codes

Warsaw, Poland

av.codes

COMPETENCIES

System design · LLM infrastructure · Agentic orchestration · RAG architectures · Local inference (llama.cpp, vLLM, ROCm/CUDA) · OpenAI-compatible API design · Docker / container orchestration · TypeScript · Python · Distributed systems · Technical documentation

RECOGNITION

- Top 1% contributor, r/LocalLLaMa
- 3.6k+ stars across maintained open-source projects
- Top 1% on Codewars (algorithmic problem-solving)
- Technical writing: >200k article reads across Medium and personal blogs

PROFILE

AI systems architect specializing in **agentic platforms**, **LLM infrastructure**, and **local-first inference** for enterprises building beyond off-the-shelf AI products. Ten+ years of full-stack engineering across web, desktop, server, and embedded systems, now focused exclusively on the architecture layer where most enterprise AI initiatives stall: orchestration, evaluation, security, and cost.

Recognized practitioner in the open-source AI tooling ecosystem - **3.6k+ GitHub stars** across authored projects, **Top 1% contributor on r/LocalLLaMa**, and creator of widely adopted Open WebUI extensions.

SIGNATURE EXPERTISE

- **Building agentic systems end-to-end** - production agent runtimes from prototype to enterprise traffic: composable modules, structured memory and context, long-running task orchestration, evaluation, failure recovery.
- **Deep, complex integrations** - MCP, Python sandbox execution in agent runtimes, custom binary protocols on AWS Lambda, SDK adapters bridging legacy services to new APIs. Pragmatic on build-vs-buy-vs-integrate decisions.
- **Infrastructure** - local and hybrid LLM inference (llama.cpp, vLLM, Ollama, ROCm/CUDA) including on-prem and air-gapped deployments; modular service composition, OpenAI-compatible APIs, container and CI/CD economics.
- **End-to-end delivery** - twelve years across servers, browsers, desktops, and microcontrollers. Two Chapter Lead tenures with hiring and architectural scope.
- **AI security & guardrails** - prompt injection defense, output filtering, audit and observability; creator of OpenGuard, a composable LLM security proxy.

SELECTED ENGAGEMENTS

JITERA

Lead AI Architect, Agentic Workflows · Sep 2024 – Present · jitera.com

Architect and primary engineer of **Jitera Boost**, the AI agent runtime powering the platform's chat and automation surface for enterprise clients including [**Sony, Hitachi, SoftBank,** and **Deloitte**]. Built on Harbor (my OSS project), Jitera Boost grew from prototype to production system serving tens of thousands of daily agent interactions.

- **Agentic Platform**: designed and shipped the agentic platform from scratch - framework for authoring and maintaining agents and agentic systems both from scratch and based on industry-standard frameworks (Agno, LangGraph, OpenAI Agents SDK, Claude Code SDK, and more). Agent versioning and orchestration, telemetry and observability, OpenAI and Anthtopic-compatible API surfaces, integrations with off-the-shelf and bespoke Chat UIs, A2UI, Agentic Teams, built-in tools.
- **Context / Tribal Knowledge** system: semantic lookup, multi-source memory ingestion (documents, code, integrations), lineage tracking, recall-frequency ranking — making agents persistently aware of project state
- **Long-running agentic tasks**: planner/mapper architecture with progress events and client polling — escaping the request-response timeout ceiling that blocks real workflow automation
- **Deep integrations**: Python sandbox execution and MCP (Model Context Protocol) as first-class agent capabilities
- **Leadership**: led the team implementing the product, later architect role across the GenAI chapter

HARBOR

Creator & Maintainer · Aug 2024 – Present · github.com/av/harbor · 2.9k+ stars, 200k+ visits

ACross-platform CLI and desktop application for orchestrating local LLM stacks. Single-command deployment across **16+ inference backends, 12+ frontends, and 60+ supporting services** with automatic service wiring. Used as reference architecture by teams evaluating self-hosted AI infrastructure. One of sub-proejcts, Harbor Boost implements dozens of advanced LLM workflow optimization techniques.

OPENGUARD.SH

Creator & Maintainer · Dec 2025 – Present · openguard.sh

Composable LLM security proxy addressing **prompt injection, output filtering**, and **audit requirements** for enterprise agentic pipelines. Built for the buyer profile that needs to ship agents past a security organization.

PRIOR EXPERIENCE

Senior Engineer → Chapter Lead · 2014-2024

Twelve years of full-cycle delivery across enterprise **SaaS, IoT**, and **BI** platforms — including two **Chapter Lead** tenures with hiring, outage ownership, and cross-cutting architectural scope.

- **Craft.co** (Oct 2020 – Jul 2024) - Chapter Lead at a BI platform serving 1M+ company profiles, where I delivered a 3.84× page load improvement (60s+ → 11s for the largest enterprise client), 3× CI cost reduction, and shipped the production LLM prototypes (news RAG, NER, knowledge graphs) that directly informed Harbor and Harbor Boost.
- **Evrythng** (~2017 – 2020; acquired by Digimarc, 2022) - Frontend Chapter Lead at an enterprise IoT/product-digitization platform serving Fortune 500 brands, where I architected the Application Customization Framework and microfrontends-style "Satellites" architecture used by client engineering teams to extend the platform without forking it.