Valerii Verezhynskyi

Principal Architect, AI Systems | B2B

✉ work@valerii.cc   📄 US Entity (Wyoming LLC) | US Billing (ACH/Wire)

COMMERCIAL ENGAGEMENTS

Refact.ai ⬈

Founding Team Member, Data & Backend Engineer   Nov 2021 – Feb 2024 / Remote

Joined as the first engineering hire (Employee #1) to build the initial product alongside ex-OpenAI engineer, Oleg Klimov.

- Built data ingestion scripts to scrape **80 Million** Code Repositories & co-authored in a creation of a Dataset used for training Code Completion model Refact-1.6B-fim (SOTA 2022, 32% HumanEval pass@1)
- Engineered enterprise-grade LLM Inference Backend supporting OSS and on-prem
- Developed a Telemetry & Statistics for Code Completion Model, User Activity

Founding Team Member, AI Developer   Feb 2024 – Jan 2025 / Remote

- Engineered LLM Client in Rust
- Developed RAG for Code Completion & Chat using AST and Vector Search code retrieval
- Implemented Agent Capabilities for LLM: repository exploration, planning, autonomous changes application (SWE bench-compatible)

Coxit, *AI Developer, Consulting Engagements* ⬈   Feb 2025 – Jun 2025 / Remote

Retained as a Lead AI Developer to bootstrap Prototypes of LLM-powered applications.

- Developed Named Entity Recognition Model for Personal Information Identification, WASM-compilable
- Engineered MCP Server with RAG for feeding LLM with Technical Documentation in Construction Domain AND Evaluation using LLM-as-a-Judge; Medium ⬈

BeeSensible, *Interim Head of AI* (0-to-1 Build) ⬈   Jul 2025 – Dec 2025 / Remote

Architected and shipped complete production-ready AI/ML Stack in <6 months

- Deployed the complete air-gapped ML Stack: training, storage, deployment, synthetic data generation
- Developed a streaming-first latency-optimized BERT/LLM Inference backend
- Created a reusable pipeline for Synthetic Dataset Generation for fine-tuning LoRAs/BERTs
- Delivered production-ready AI/ML Backend, ML Ops, Synthetic Data Pipeline, Finetuned BERTs & LoRAs (70 labels, production-grade F1)

Nautiloid Protocol LLC, *Founder & Principal Architect*   Jan 2026 – Present

Case studies:

- Local Voice-to-Voice Cascade Stack ⬈ | STT+VAD+LLM+TTS | gRPC, websockets, HTTP
  Real-time, Chat Completions, Audio, Transcriptions APIs implementation with Parakeet V3, gpt-oss-20b, kokoro TTS that produce **sub 400ms TTFA** on consumer-grade GPU.

- Zero-dependency MCP Server (Python) ⬈ | Streamable HTTP, JSON-RPC
  Custom MCP Server implementation with explicit global state and without opaque decorators unlike reference implementations.

STACK

Languages: Python, Rust
Tools & Technologies: uv, fastapi, pydantic, protobuf, websockets, vllm, llama.cpp, onnx, aim, mlflow, otel+clickhouse+grafana
Environment: macOS, Proxmox, bare-metal/HPC deployment