Artem Vozniuk

Senior Software Engineer – AI Infrastructure & R&D

■ artemv.tech | ■ LinkedIn: linkedin.com/in/artem-vozniuk-6036b290

■ artvoz911@gmail.com | ■ DemoHub Live: artemv.tech/face-fusion

Summary

Senior Software Engineer with 10+ years of experience building high-performance systems and AI infrastructure. Skilled in building end-to-end solutions for Generative AI. Experienced in both Python and C++, with a strong foundation in R&D, applied math, low-latency systems, and modern cloud tooling.

Experience

Allegory — AI Startup

Senior Software Engineer – AI Infrastructure (Sep 2024 – Present)

- Designed and implemented the full backend and cloud infrastructure for a generative AI platform built around diffusion models.
- Developed an orchestration engine allowing declarative workflow configuration in YAML, integrating LLMs, image/video generation APIs, and model pipelines (SDXL, Flux, Hedra, Sonic).
- Built scalable FastAPI-based services with AWS, Kubernetes, Terraform, PostgreSQL, and Redis.
- Implemented observability stack (Prometheus, Grafana, Sentry) and CI/CD pipelines via GitHub Actions.
- Enabled rapid experimentation by allowing non-engineers to define and deploy new workflows in minutes, improving iteration speed during R&D.

Inworld AI

Senior Software Engineer – SDKs & Edge Inference (Feb 2022 – Sep 2024)

- Developed cross-platform C++ SDKs and on-device inference systems powering low-latency, real-time AI workloads.
- Created NDK core library for gRPC-based communication with Inworld’s AI infrastructure platform.
- Ported client-side inference for VAD (Silero) and Whisper (speech-to-text) to optimize responsiveness.
- Built Unreal Engine SDK from scratch, reducing integration time to minutes for partner studios.
- Set up CI/CD pipelines via GitHub Actions across Windows, macOS, Linux, iOS, and Android.
- Collaborated with partners including NVIDIA, Ubisoft, and Disney on demos and integrations.

Game Development (Saber Interactive / Playrix / The Multiplayer Group)

Senior Software Engineer – Gameplay / Engine / Physics (Feb 2014 – Jan 2022)

- Contributed to AAA and mobile titles including Quake Champions, World War Z, Gardenscapes and more.
- Worked on gameplay, and engine systems with a focus on performance and cross-platform optimization (PC, consoles, mobile).

Personal Projects

DemoHub – github.com/art-vozniuk/demo-hub

- • Developed a production-ready MVP template for building AI-powered products with microservices architecture, async job processing, and scalable infrastructure.
- • Integrated PyTorch-based inference workers and optimized ONNX Runtime GPU execution for improved throughput.
- • Deployed observability and CI/CD for production-like workloads.

Core Skills

Languages: Python, C++, Go

AI/ML: PyTorch, Diffusion Models, ONNX Runtime

Backend/Infra: FastAPI, Docker, Kubernetes, Terraform, AWS, PostgreSQL, Redis

Observability: Prometheus, Grafana, Sentry | CI/CD: GitHub Actions

Other: Workflow orchestration, R&D, experimentation, inference optimization, GPU pipelines

Education

Saint Petersburg State University — Master’s degree, Mathematics and Computer Science

Certifications

Neural Networks and Deep Learning – DeepLearning.AI (Feb 2024)

Credential: coursera.org/account/accomplishments/verify/MN6V7AMFLFXU