

ANDREI-CRISTIAN RUSU

Full-Stack & AI Engineer

Constanța, Romania | Remote | English (C1) | andrei.rs
rusuandreicristian369@gmail.com | github.com/rusuandreicristian369 | linkedin.com/in/andrei-cristian-rusu

Summary

Full-stack engineer (TypeScript, React, Node.js, Python) with five years' experience owning products end to end: as a technical lead, as the sole engineer for a long-term client, and as co-founder and CTO of Metricor, whose ERP runs in production at a large ready-mix concrete producer. MSc in Distributed Systems, with research on fine-tuning LLMs.

Experience

Co-Founder & CTO · Metricor Technologies, Bucharest Jan 2026 – Present

- Lead engineering and architecture, in a two-engineer team, for Metricor Concrete ERP — a vertical ERP built with a Bucharest producer (200,000+ m³ of concrete a year) and running in production there.
- Shipped order-to-cash in one system: quotes, e-signed contracts, automated credit checks, hour-by-hour dispatch, invoices generated from delivery tickets, a client portal and a driver app. React, Node.js, MongoDB, Hetzner.
- Building Metricor BI, an agentic analytics platform (LangGraph with custom orchestration): agents query databases, run Python and search across data sources, backing every answer with traceable, step-by-step evidence.
- Kept it model-agnostic: developed it on local open-weight LLMs, demoed it to the ERP client on Claude Opus 4.8, and track answer quality with an evaluation suite on synthetic databases.

AI Platform & Infrastructure Engineer · Confidential AI company (NDA), freelance Jul 2026 – Oct 2026

- Designed, built and ran a self-hosted AI platform on a multi-node cluster of 16 NVIDIA GPUs (about 464 GB VRAM, including an RTX PRO 6000 Blackwell).
- Served open-weight LLMs (vLLM, SGLang, llama.cpp) with tensor parallelism and FP8/NVFP4/INT4 quantization.
- Built per-user AI agents, one Docker container each, with tool calling, memory, RAG, speech and OCR.
- Secured it (VPN-only access, SSO, encrypted secrets); ran backups and monitoring on Docker, k3s and Incus.

Freelance Full-Stack Engineer · Long-term client, Austria 2023 – 2025

- Sole engineer for one founder: took three products from idea to launch, owning every technical decision. React, Next.js, NestJS, MongoDB, Kubernetes on DigitalOcean, GitLab CI, Grafana.
- Built **VRon**, a no-code platform for immersive 3D and VR showrooms, in the browser and on VR headsets; showcased at Web Summit 2025. Solved real-time networking (WebRTC) and frame-rate optimization on headsets.
- Built **Make3D**, a 3D-printing marketplace where partner printers print and ship uploaded models.
- Built **Makeronline.shop**, an e-commerce platform integrated with VRon for 3D product presentation.

Backend Developer · Bee Coded 2023

- Built from scratch a VTEX product-import pipeline for Excel files of up to about 1M rows (Excel's limit): Node.js streams process them without loading whole files into memory, on AWS Lambda, SQS and DynamoDB.

Full-Stack Developer, promoted to Technical Lead · Cult of Coders 2021 – 2023

- Led cross-functional teams of about seven building client MVPs, from VR to healthcare, from architecture to deployment; promoted to Technical Lead after two months.
- Ran technical interviews, code reviews, automated testing and sprint planning; owned DevOps on AWS, across manual and CI/CD deployments.

Education

MSc, Distributed Systems · University of Bucharest 2022 – 2024

- Thesis: text-to-SQL and text-to-Entity Framework. Built a dataset by translating 87% of Spider's SQL into Entity Framework, then fine-tuned Llama 3 8B, raising accuracy from 31% to 51%. [GitHub](#)

BSc, Computer Science · University of Bucharest 2019 – 2022

- Thesis: a computer-vision 3D Sudoku solver — reads a scanned puzzle with OpenCV, solves it by backtracking and draws the solution back onto the image. [GitHub](#)

Skills & Awards

Stack	TypeScript, Python, React, Next.js, Node.js, NestJS, Three.js, MongoDB, PostgreSQL; testing with Jest, Pytest, Playwright, Cypress
Cloud & DevOps	AWS (EC2, S3, Lambda, DynamoDB, Route 53), Docker, Kubernetes, GitLab CI, Grafana, Linux, Hetzner, DigitalOcean
AI & LLMs	LangGraph, Claude API, agents & tool calling, RAG, text-to-SQL, LLM evaluation, fine-tuning, PyTorch, Hugging Face, vLLM, SGLang, llama.cpp, quantization
Awards	SmartHack Innovation Prize (2020, 2023), Cornerstone Technology 2nd Prize (2023), EEML Summer School Best Poster (2022)