MILOŠ BOŽIĆ

Lead ML/AI Engineer | PhD in Machine Learning

linkedin.com/in/milosbozic
milosbozic.com
Belgrade, Serbia

SUMMARY

Lead ML/AI engineer with a PhD in Machine Learning and 10+ years taking models from research to production across energy, retail, automotive marketplaces, AI hardware and robotics. I have built and led AI teams of up to 15+ scientists and engineers. Most recently I founded Pyxero.ai, a bootstrapped AI startup funded through ML/AI consulting (LLM fine-tuning and quantization, RAG pipelines, agentic systems, computer vision), while building Agent Gate, a governance layer for enterprise AI agents. I am looking for a full-time Staff/Lead ML or AI Engineer role where I can own hard technical problems end to end and grow a team for the long term.

WORK EXPERIENCE

January 2024 - Present

Founder & Lead AI Engineer, Pyxero.ai, Belgrade.

Bootstrapped AI startup, self-funded through ML/AI consulting and engineering services for clients, while building our own AI governance product.

Client work - LLM & GenAI engineering:

- Fine-tuned LLMs for domain-specific tasks and optimized and quantized models for efficient production inference.
- Designed and shipped RAG pipelines and agentic systems for client use cases.
- Trained and deployed computer vision models for client projects.

Product - Agent Gate:

- Designed and built Agent Gate, an AI governance tool: a policy-enforcement gateway between AI agents and enterprise tools, with an MCP server for Claude/Cursor/ChatGPT clients, 30+ maintained connectors and credential vaulting with short-lived scoped tokens.
- Implemented DLP guardrails that inspect tool-call inputs/outputs to block, mask or redact sensitive data.
- Built prompt-injection and AI-security research into the product (talk: Belgrade AI Week 2025).
- Built observability: immutable audit trail streamed to SIEM, per-agent/tool/tenant latency and error metrics.

May 2023 - September 2024

AI Solutions Architect, Alascom, Italy (Belgrade-based).

Designed and led end-to-end AI solutions for industrial automation and robotics clients, from concept to deployment.

Architecture & leadership:

- Led the design of end-to-end AI systems: architecture blueprints, technical specifications and integration plans that ensured scalability, security and performance in production.
- Took 3+ AI projects from concept to deployment, defining scope, objectives and deliverables with product teams and client stakeholders.

- Guided the development team technically and evaluated and selected AI tools and frameworks for the company's stack.

Engineering:

- Built vision-guided robotic pick-and-place for objects of varying shapes and structures.
- Developed camera-robot calibration methods.
- Built visual inspection and anomaly detection systems for manufacturing quality control.
- Developed LLM-based text and voice interfaces for natural-language human-robot interaction.
- Built a synthetic data pipeline that combines realistic 3D rendering with generative-AI post-processing, reducing the need for real labeled data.

December 2021 - May 2023

AI Lead, Anari AI, Belgrade.

Deep-tech startup building AI hardware and a compiler stack for efficient deep learning inference on FPGA.

Engineering:

- Led the architecture and hands-on development of transformer-based 3D point cloud segmentation models for autonomous driving and mapping use cases.
- Optimized deep learning models for hardware acceleration using quantization, pruning and operator fusion, achieving 10× faster inference with less than 3% accuracy loss on FPGA targets.
- Contributed to the ML compiler that maps neural networks onto custom FPGA architectures, enabling deployment of models on the hardware.
- Designed training and evaluation data pipelines for large-scale 3D point cloud datasets.

Leadership:

- Built and led a team of 5 AI scientists and engineers, covering hiring, mentoring, goal-setting, and code and research reviews.
- Owned the AI technical roadmap: selected, scoped and prioritized AI projects to balance research ambition against product and hardware constraints.
- Partnered with product and hardware teams to turn research results into product capabilities and customer demos, and shaped the company's AI research agenda.

August 2019 - December 2021

Chief Data Scientist, NIS a.d. Novi Sad, Belgrade.

Led the Digital Data Lab, the company's central ML hub, delivering data products across retail, geology, production, refining and seismic operations at one of Southeast Europe's largest energy companies.

Leadership & strategy:

- Defined data and ML strategy across 5 business domains, turning business challenges into a prioritized portfolio of data products.
- Built and ran the Digital Data Lab as the company's in-house center for ML delivery, leading a team of 10+ data scientists and engineers through 20+ projects from concept to production.
- Worked directly with business stakeholders to translate requirements into ML solutions and find new opportunities, growing the portfolio to more than 20 projects.
- Directed partnerships with external vendors and academic institutions to speed up R&D, and designed and set up the lab's complete on-premise compute infrastructure.

Delivered ML portfolio:

- Retail: customer churn prediction, segmentation and product recommendation.
- Geology: lithology facies classification and grain size distribution prediction from well data.
- Production & refining: predictive maintenance for electrical submersible pumps (ESPs) and multi-stage centrifugal pumps.
- Seismic & remote sensing: GAN-based seismic data inpainting and computer vision for aerial drone terrain imagery.

April 2018 - August 2019

Division Research Manager, Everseen, Belgrade.

Promoted from Senior Research Scientist (April 2018 - January 2019)

Everseen builds computer vision AI for retail, deployed in 250+ superstores to detect checkout errors and losses in real time.

Research leadership:

- Led the research teams at the Belgrade R&D center (15+ scientists and engineers), covering hiring, mentoring and project reviews.
- Defined the center's research strategy and roadmap, choosing research directions that balanced technical ambition against cost and delivery timelines.
- Turned research results into production models for the core product, working with engineering teams in Ireland and Romania.

Computer vision research & engineering:

- Developed one-shot and few-shot learning methods (CNN-based) for object recognition and detection, letting the system recognize new retail products from fewer than 20 images per product instead of large labeled datasets.
- Applied weakly supervised learning to object detection, cutting manual annotation effort by 20% by moving from bounding boxes to image-level labels.
- Built synthetic data generation pipelines to augment training data for rare products and edge cases.

July 2015 - July 2018

Data Scientist, Vast.com, Belgrade.

Vast.com powers vehicle search and marketplace platforms. I built and led the company's deep learning computer vision work within the data science department.

Leadership:

- Built and led the deep learning computer vision team (3 engineers), owning the roadmap from problem selection to production deployment.
- Introduced deep learning to the company's data science practice, moving image analysis from manual or rule-based processes to production CNN models.

Computer vision:

- Built a car make / model / generation recognition system covering 1k vehicle classes, used to auto-tag and validate 100M listings.
- Developed models for car orientation detection and transmission type recognition from listing photos, enabling automated listing quality checks and enrichment.
- Built salvage car detection to flag damaged vehicles in listings, improving marketplace trust and data quality.
- Implemented VIN recognition (OCR) from photos, automating vehicle identification.
- Designed a visual search system that lets users find cars by image similarity.

Pricing & anomaly detection:

- Built car price prediction models and anomaly detection for flagging mispriced or fraudulent listings.

October 2012 - July 2015 — Research Associate, Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade.

- Developed mutual-information-based feature and instance selection algorithms for time-series forecasting, applied to electrical load forecasting, stock market trend prediction and heart disease prediction.
- Published in peer-reviewed journals including Neurocomputing, Entropy and Energies (200+ citations across my research), and contributed to 2 national research projects.

2010 - 2012 — Research Associate, Faculty of Electronic Engineering, University of Niš

Built a system for collecting data on electrical lighting usage in business buildings and segmented users by their lighting behavior using clustering methods. The resulting user profiles defined optimal working plans for automating electrical lighting in business buildings.

2010 - 2015 (part time) — Data Scientist, R&D Alfatec

Worked on outsourced projects, mainly for US partners:

- Water pump failure prediction and anomaly detection system.
- Cable TV user segmentation and churn prediction.

2008 - 2010 — Young Researcher, Faculty of Electronic Engineering, University of Niš

EDUCATION

2010 - 2015: PhD in Computer Science and Electrical Engineering (Machine Learning), Faculty of Electronic Engineering, University of Niš, Serbia.
Thesis title: "*Short-term load forecasting based on machine learning methods*".

2008 - 2010: Doctoral studies in Electrical Engineering (energy savings in electrical lighting based on consumer behavior profiling), Faculty of Electronic Engineering, University of Niš, Serbia. Discontinued due to my supervisor's death.

2002 - 2008: Master of Science in Electrical Engineering (Electrical Power Engineering), Faculty of Electronic Engineering, University of Niš, Serbia.

SKILLS

Technical skills and competences

- Machine learning, Deep learning, Artificial intelligence, LLMs & Generative AI, AI agents, Computer vision, Time-series forecasting, Robotics, Statistics
- LLMs & GenAI: fine-tuning, quantization and inference optimization, RAG pipelines, agentic systems, MCP, AI security (prompt-injection defense, guardrails)
- Programming languages: Python, C#

- Favorite packages and libraries:
    - Deep learning & LLMs: PyTorch, TensorFlow, Hugging Face, LangChain, LangGraph, LlamaIndex
    - Python: scikit-learn, NumPy, pandas, matplotlib, NLTK, spaCy, Transformers, OpenCV
    - MLOps & experiment tracking: Weights & Biases
- Web development: FastAPI, Flask, React, Next.js
- Databases: PostgreSQL, MongoDB
- System administration/DevOps: Linux, Git, Docker, Kubernetes
- Cloud platforms: AWS, Azure

Leadership and other skills

- Building and leading AI teams: hiring, mentoring, goal-setting, code and research reviews
- AI strategy, research roadmaps and project scoping
- Stakeholder management and client communication
- Research methodologies, technical writing and public speaking