Milorad Trifunovic

Senior AI Product Engineer (LLMs, RAG, AI Agents)

mtrifunovic435@gmail.com https://www.linkedin.com/in/milorad-trifunovi%C4%87-5a16b63b5/ Belgrade, Serbia

SUMMARY

Senior Applied AI Engineer with a software engineering foundation built across game development, full-stack product engineering, and applied AI. My background started with backend and product-focused development, then gradually moved into machine learning, NLP, workflow automation, and modern generative AI systems. I've worked hands-on across the full lifecycle of AI products, including building retrieval-based and agent-driven solutions, developing data and inference pipelines in Python, and shipping APIs, orchestration layers, and user-facing tools with React and TypeScript. I'm strongest in roles where AI has to work in production, integrate cleanly with real systems, and deliver practical value rather than remain as a prototype.

SKILLS

AI Engineering & Machine Learning:
Python, PyTorch, TensorFlow, scikit-learn, NumPy, LLMs, Retrieval-Augmented Generation (RAG), Generative AI, AI Agents, NLP, Prompt Engineering, Model Evaluation, Fine-Tuning, LangGraph

Backend, APIs & Data Systems:
Node.js, Express.js, Java, Spring Boot, Django, FastAPI, Flask, REST APIs, Microservices, Model Serving, Data Pipelines, ETL, Workflow Automation (n8n, Zapier), Webhooks, PostgreSQL, MongoDB, Pinecone, Weaviate, FAISS, BullMQ, Celery

Frontend & Product Engineering:
React, Next.js, TypeScript, JavaScript, HTML5, CSS3, Tailwind CSS, REST API Integration, SSR/CSR

Cloud & Infrastructure: AWS (ECS, S3, CloudWatch, Lambda), Docker, GitHub Actions, Vercel, Redis, API Gateway, SQS

EXPERIENCE

Senior AI Product Engineer

Remote

Netguru

01/2024 - Present

Digital product consultancy delivering software engineering, design, and AI solutions for startups and enterprises.

- Architected and delivered **LLM-powered workflow systems** that transformed unstructured inputs into structured, actionable outputs, reducing manual processing effort by **30–40%** across high-volume operational flows.
- Built Python-based **RAG pipelines** integrating curated knowledge sources, retrieval, and prompt-grounding logic, improving factual accuracy and reducing unsupported or hallucinated responses by **~30%**.
- Developed **agent-style orchestration workflows** for multistep reasoning, task routing, and exception handling, increasing successful end-to-end workflow completion rates by **25%** without human intervention.
- Designed and exposed **model-backed APIs** enabling real-time AI capabilities across internal and customer-facing products, achieving **sub-300ms median latency** and supporting high-throughput usage scenarios.
- Implemented full-stack AI product features using **React and TypeScript**, including workflow dashboards and review interfaces, reducing manual review time by **25%** and improving user adoption across internal teams.
- Integrated frontend, backend, and AI inference layers into **cohesive production systems**, enabling seamless AI-driven user experiences and reducing system friction across workflows.
- Introduced evaluation pipelines for retrieval quality, grounding accuracy, and response reliability, reducing model iteration and validation cycles by **30–35%**.
- Deployed and scaled **containerized AI services** on AWS (ECS, S3, CI/CD pipelines), improving platform availability to **99.9%** and reducing deployment time by **40–50%**.

AI Engineer

Remote

Rossum

07/2022 - 12/2023

Cloud-native AI document automation company building end-to-end workflow automation for transactional business processes.

- Built **Python** services for AI-driven document and transaction **workflows**, reducing manual processing time by **45%** across high-volume business operations.
- Developed **AI agent–style orchestration logic** to coordinate document ingestion, validation, and exception handling workflows, enabling adaptive decision-making across automated processing pipelines.
- Applied **PyTorch** and **scikit-learn** to improve document classification, field extraction, anomaly detection, and exception routing within automated workflow pipelines.
- Designed and implemented **workflow orchestration logic** for validation, approval, and post-processing flows, improving end-to-end workflow completion speed by **30%** and reducing manual intervention in document processing.
- Delivered model-backed capabilities through **REST APIs** and internal service endpoints, allowing workflow engines and external systems to consume predictions in near real time.
- Deployed production-grade ML inference services using **Docker and AWS (ECS, S3)**, enabling scalable model serving and reducing deployment time by **50%** through automated CI/CD workflows.
- Designed workflow automation using **n8n** to orchestrate AI-powered document processing pipelines, integrating model outputs with downstream systems through APIs and webhooks.

EXPERIENCE

Machine Learning Engineer
Remote

DeepL
03/2021 - 06/2022

Language AI company building machine-learning-powered translation, writing, and communication products.

- Engineered **Python**-based NLP pipelines for language processing tasks, improving the quality and consistency of text understanding workflows used in production translation services.
- Trained and fine-tuned neural models with **PyTorch** and **TensorFlow** for machine translation quality estimation, language detection, and text classification, improving offline model performance by **15%**.
- Applied **scikit-learn** to prototype baseline models, feature extraction workflows, and offline evaluation utilities that accelerated experimentation across NLP projects.
- Orchestrated data preparation and feature generation jobs for large multilingual corpora, increasing dataset reliability and reducing preprocessing failures by **25%.**
- Operationalized inference services with containerized deployment workflows, helping move ML functionality from experimentation into stable internal and customer-facing systems.
- Refined model evaluation and error analysis processes for multilingual outputs, raising confidence in production releases and shortening iteration cycles for new model versions.

Software Engineer
Belgrade, Serbia

SAP
01/2020 - 12/2020

Global enterprise software company building data, analytics, and intelligent business platforms.

- Developed a full-stack enterprise application using **React, TypeScript, and Node.js**, delivering scalable features for data-driven business workflows.
- Built reusable UI components and dashboards to support **analytics and intelligent automation scenarios**, improving usability and feature delivery speed.
- Implemented backend services and **RESTful APIs** to process and expose structured business data across distributed systems.
- Integrated frontend applications with backend and data services to enable **near real-time data visualization and operational insights.**
- Contributed to features aligned with **automation and machine learning–assisted workflows**, enhancing system capabilities for intelligent decision support.
- Applied **TypeScript, OOP principles, and modular architecture** to improve code maintainability and scalability across shared components.
- Collaborated with cross-functional teams in an agile environment to deliver end-to-end functionality across UI, API, and data layers.
- Enhanced application performance by optimizing API response handling and frontend rendering efficiency, reducing page load times by **23%.**

Software Engineer I
Belgrade, Serbia

Nordeus
08/2018 - 12/2019

A leading game development company building large-scale, real-time mobile gaming platforms.

- Developed gameplay systems and backend services using **Java (Spring Boot)** with strong **object-oriented programming (OOP)** principles, improving modularity and maintainability of game logic.
- Designed and implemented scalable services for **player data, game events, and session management**, supporting high-concurrency gameplay environments.
- Built internal tools and dashboards using **JavaScript (React)** for game configuration, monitoring, and analytics, reducing manual setup and configuration effort by **30%** and improving visibility into live game operations.
- Modeled game mechanics and event systems using **OOP design patterns**, enabling flexible feature extensions and shortening new feature rollout cycles by **20%.**
- Integrated **RESTful APIs and messaging systems** to support real-time game interactions and event-driven architecture, improving system responsiveness and ensuring stable handling of high-concurrency player activity.
- Optimized performance of backend services by improving data access patterns and reducing response latency.
- Investigated and resolved production issues, delivering timely hotfixes that reduced recurring live-service incidents by **30%** and improved system stability during peak usage periods.

EDUCATION

B.Sc. in Computer Science
10/2014 - 08/2018

University of Belgrade