Eziyo Ehsani



Location: Milan, Italy, **Mail:** eziyo.ehsani@gmail.com, **Mobile:** +393487803246, GitHub, LinkedIn

Summary

AI Solution Architect with a background in machine learning engineering and data science, and strong foundations in data structures, algorithms, and production ML systems. Skilled in Python, PyTorch, SQL, FastAPI, Docker, Kubernetes, and Git-based workflows, with hands-on experience building LLM and transformer-based NLP modules, as well as on-device/edge AI deployments. Experienced in designing and deploying scalable machine learning and RAG pipelines, focusing on algorithm development, performance optimization, and privacy-preserving, locally deployed agentic AI systems.

Professional Experiences

AI Functional Analyst – Sunprime *Milan — Oct 2026 – Present*

- Design AI solution architectures and system designs based on business and technical requirements.
- Translate requirements into technical designs, integration patterns, and scalable solution architectures.
- Assess and prototype AI capabilities to support system design and solution delivery.

AI Systems Engineer Intern – Software and Sustainability S2 Group *Amsterdam — Apr 2025 – Mar 2026*

- Deployed and optimized 8 open-source LLMs (0.5B–9B) on-device using llama.cpp, PyTorch, and ML compression techniques (INT4/INT8 quantization, pruning, distillation), enabling agent-ready NLP inference under strict memory and compute constraints.
- Built automated Linux/Bash orchestration pipelines for 480 evaluation runs, collecting ML metrics (latency, energy usage, token throughput, accuracy) to validate autonomous, privacy-preserving language model deployments.
- Profiled latency–energy trade-offs across ARM-based mobile hardware and SoC architectures, defining configurations that balance real-time NLP responsiveness with sustainable resource consumption for on-device edge AI systems.

On-Device ML Intern – DIETI Lab – University of Naples Federico II *Naples, Italy — Sep 2025 – Feb 2026*

- Trained and evaluated transformer-based NLP models (e.g., BERT, DistilBERT) for text classification, comparing full-precision baselines with quantized on-device variants.
- Applied Explainable AI techniques (SHAP, LIME) to clarify model behavior and support decisions on accuracy–efficiency trade-offs for deployment on resource-constrained hardware.
- Built supervised learning pipelines in Python (scikit-learn, PyTorch) and conducted exploratory data analysis to identify configurations that balance latency, energy use, and predictive performance.

Data Scientist – HascoNet *Tehran, Iran — Jul 2020 – Aug 2023*

- Automated KPI dashboards in Power BI, cutting manual reporting time by 30%.
- Built predictive churn and HR analysis models on 100k+ records to inform retention and staffing decisions.
- Designed ETL pipelines feeding analytics and machine learning workloads.

Data Science Intern – HascoNet *Tehran, Iran — Jan 2020 – Jun 2020*

- Built Power BI data models and DAX measures to enable dynamic filtering and performance monitoring.
- Assisted senior analysts with reporting, data validation, and documentation for analytics projects.

Education

Master's in Data Science (Grade: 108/110) – University of Naples Federico II *Sep 2023 – Mar 2026*

- **Thesis:** Green Mobile Intelligence: Quantifying Energy-Accuracy Trade-offs in On-Device LLMs.
- **Coursework:** AI Systems Engineering, Data Mining & Machine Learning, Hardware & Software for Big Data

Bachelor's in Computer Engineering – Azad University *Sep 2017 – Jul 2021*

- **Thesis:** Design of an IoT-Based Energy Consumption Monitoring and Optimization System
- **Coursework:** System Design & Analysis, Compiler Design, Microprocessor, Computer Architecture

Selected Projects

- **On-Device LLM Optimization:** Deployed 8 open-source LLMs (0.5B–9B) on a Snapdragon 8 Elite device using llama.cpp with Q4_K_M & IQ4_XS quantization; automated 480 isolated inference runs with BatteryManager energy profiling to quantify feasibility of on-device agent deployments under real hardware constraints.
- **RAG Clinic Assistant:** Designed a tool-using RAG agent for a medical clinic (Python, LlamaIndex, ChromaDB, local LLMs) that retrieves grounded context and orchestrates multi-step query resolution; deployed via FastAPI backend with a custom chat interface.
- **Device Power Profiler:** Extended an Android/Kotlin energy profiling tool interfacing with Android's BatteryManager API via ADB to support reliable telemetry for on-device agent evaluation across heterogeneous hardware.
- **Local Voice Cloning & TTS System:** Built a fully offline multilingual voice cloning system (OpenVoice, MeloTTS, Gradio) enabling autonomous, on-device voice synthesis in 6 languages with no cloud dependency.

Skills

- **Core Methods:** Supervised/Unsupervised Learning, Sequence Modeling, Reinforcement Learning, Bayesian Optimization, Cross-Validation, Robustness Evaluation
- **Machine Learning & NLP:** XGBoost, PyTorch, TensorFlow, BERT/Transformers, Word Embeddings, RAG, SpaCy, Evaluation (BERTScore, LLM-as-a-judge)
- **Agentic AI & Orchestration:** LangChain/LangGraph, CrewAI/AutoGen, Multi-Agent Systems, Tool Use & Function Calling, RAG Pipelines, Prompt Engineering, n8n, Model Context Protocol (MCP)
- **On-Device & Edge AI:** On-device LLM Deployment, llama.cpp, Quantization (INT4/INT8, GGUF, AWQ), Knowledge Distillation, Pruning, ONNX Runtime, SoC Architectures, Latency & Energy Profiling
- **Programming:** Python, C++, SQL, Bash/Shell Scripting, ADB, GDB, Linux perf
- **Engineering & DevOps:** FastAPI, Docker, Git/GitHub, CMake, CI/CD, Vector Databases (ChromaDB), Spark, Hadoop, ETL, Redis

Publications

- **Eziyo Ehsani**, Luca Giamattei, Ivano Malavolta, Roberto Pietrantuono (2026). Sustainability Is Not Linear: Quantifying Performance, Energy, and Privacy Trade-offs in On-Device Intelligence. *Under Review at Empirical Software Engineering – EMSE*. arXiv preprint arXiv:2603.26603

Certifications

- **Building with the Claude API** – Anthropic *2026*
- **Model Context Protocol: Advanced Topics** – Anthropic *2026*
- **Claude Code 101** – Anthropic *2026*
- **Data Analyst** – DataCamp *2025*

Languages

- **English:** Full professional proficiency
- **Italian:** Intermediate proficiency
- **Persian:** Native

Autorizzo il trattamento dei miei dati personali ai sensi del D.Lgs. 196/2003 e del Regolamento UE 2016/679 (GDPR).