

ARA YEROYAN

Senior AI Engineer | LLMs, GenAI, Voice AI | Agentic AI & RAG
Yerevan, Armenia • ar23yeroyan@gmail.com • linkedin.com/in/ara-yeroyan

PROFILE

Senior AI Engineer with 5+ years of experience building end-to-end ML and GenAI systems across **voice, retrieval, and applied NLP**. My work sits at the intersection of research and production - including **published work with the NVIDIA NeMo team on Audio/NLP** - where I design, evaluate, and deploy AI systems that must perform reliably in real-world environments. Across roles in AI/ML and NLP, I have built large-scale pipelines spanning data processing, training, inference optimization, deployment, monitoring, and reporting. My recent focus includes **LLM-powered products, voice agents, RAG pipelines, multi-agent orchestrations**, and evaluation/observability for GenAI systems. Beyond implementation, I shape system design, define KPIs, improve reliability, and connect technical work to business outcomes. I have also contributed to the broader AI community through **peer-reviewed publications** (SIGIR, INTERSPEECH), **Google Summer of Code** mentorship, open-source software, datasets, AI models, and technical education.

KEY EXPERTISE

LLMs, GenAI & Agentic AI	RAG & Information Retrieval	Voice AI & Applied NLP
System Design & Architecture	ML Pipelines & MLOps	LLM Eval & Observability

PROFESSIONAL EXPERIENCE

AI Engineer, Research - ServiceTitan

Sep 2024 - Present | Yerevan, Armenia

- Built LLM evaluation and monitoring pipelines for voice agents, tracking quality, safety, drift, and regressions.
- Developed LLM-based reporting systems for high-volume operational data with complex ETL pipelines.
- Optimized LLM inference achieving **3.2x performance improvement** using state-of-the-art deep learning and engineering practices.
- Defined and operationalized **customer-facing KPIs** tied to revenue-driving factors.

ML Architect / Consultant - Akryl Digital (Client: GIZ)

Aug 2025 - Apr 2026 | Hamburg, Germany (remote)

- Designed and deployed **production-grade multi-agent chatbot systems** for GIZ, integrating RAG pipelines with **sub-2-second** end-to-end latency.
- Built **visual RAG systems** transforming PDFs into high-resolution images with multi-vector retrieval and token-level filtering.
- Hosted & optimized medium-scale LLM applications in multi-tenant environments under constrained hardware.
- Primary **technical liaison for GIZ**, bridging complex AI implementation with non-technical stakeholders.

Senior ML Engineer / Consultant - Plat.ai

Sep 2022 - Mar 2025 | Yerevan, Armenia

- Built AutoML workflows integrated into the **Snowflake Marketplace** as a native application.
- Integrated explainability and interpretability mechanisms into financial ML systems.
- Developed domain-specific features improving **F1 score by 17%** for banking/insurance clients.
- End-to-end ML systems: data extraction, ETL, feature engineering, modeling, deployment, cloud integration.

Senior ML Engineer - EPAM Systems (Enterprise client: retail/fashion)

Jun 2024 - Sep 2024 | Yerevan, Armenia

- Built a successful MVP for a **real-time multimodal recommendation system ("Fashion Assistant")** for a major enterprise client in the **fashion sector**.
- System retrieved products based on user preferences using multimodal AI techniques.

Senior NLP Engineer - Stealth Startup (YC-backed)

Jan 2024 - Jun 2024 | Palo Alto, USA (remote)

- Created a **real-time voice assistant** that interacts with users over live calls while using external tools and third-party APIs to manage appointment booking and cancellation.
- Implemented offline sales assistant workflows using advanced prompt engineering techniques such as tree-of-thought, self-reflection agents, and **agent-style reasoning patterns**.

Lecturer - Introduction to Data Science and Machine Learning - Armenian Code Academy (ACA)

Jul 2024 - Nov 2024 | Yerevan, Armenia

- Delivered a comprehensive course consisting of:
 - Higher Mathematics for ML: Linear Algebra, Calculus, Probability and Statistics
 - Python for Data Science
 - Introduction to Machine Learning and Deep Learning

Machine Learning Specialist - Podcastle

Jul 2021 - Aug 2022 | Yerevan, Armenia

- Worked on audio synthesis and enhancement including voice cloning, dynamic range compression, and volume normalization leveraging deep learning and DSP techniques.
- Researched, reproduced, and implemented state-of-the-art methods, delivering proofs of concept, research insights, and optimizations for distributed training pipelines.
- Led data collection initiatives and designed pipelines for scheduling training runs and model serving, keeping models continuously updated with new annotated data.

Data Scientist - Funday Agency

Apr 2021 - Sep 2021 | Yerevan, Armenia

- Automated customer engagement monitoring across social media, applying statistical analysis and anomaly detection to derive actionable business insights.
- Designed scrapers and subsequent text processing and data cleaning pipelines for large-scale unstructured data.
- Developed an internal platform featuring interactive dashboards and automated reporting mechanisms, enhancing decision-making processes across the company.

Machine Learning Engineer - FAST Foundation / CYNORA

Jan 2021 - Mar 2021 | Yerevan, Armenia

- Led a regression analysis project on molecular datasets, applying chemoinformatics techniques for data cleaning and exploratory analysis.
- Designed and trained deep learning models using Graph Convolutional Neural Networks (GCNNs) with PyTorch Geometric to predict molecular properties.

Machine Learning Researcher - Hero House

Sep 2020 - Feb 2021 | Yerevan, Armenia

- Image enhancement and image-to-image translation using deep learning techniques.
- Developed conditional generative adversarial networks (GANs) that outperformed existing physics-based and clinical computer vision algorithms.

EDUCATION

BSc in Data Science - American University of Armenia • GPA: 3.85/4.0 (Top 1% of cohort)	Aug 2020 - Jun 2024
LLM Summer School - YerevaNN	Jul 2025
Startup Incubator (MedGPT) - Entrepreneurship and Product Innovation Center (EPIC)	Sep 2023 - Feb 2024
Advanced Machine Learning (231 hrs) - Armenian Code Academy	Dec 2020 - Jun 2021
DevOps - Armenian Code Academy	Dec 2021 - May 2022
AI BootCamp - FAST Foundation	Sep 2020 - Feb 2021
1991 Military Unit - AI/ML Program - FAST Foundation	Feb 2020 - Jul 2020
Nationally recognised intensive program covering: Mathematics for Machine Learning, Programming for Machine Learning, Classical Machine Learning, Deep Learning, and an industry-tied Capstone.	

PUBLICATIONS

ORCID: 0009-0009-3605-9702 | Google Scholar: d8lf8-gAAAAJ

- Visual RAG Toolkit: Scaling Multi-Vector Visual Retrieval with Training-Free Pooling and Multi-Stage Search - Accepted at SIGIR 2026 (arxiv, Feb 2026)
- From Scarcity to Sufficiency: Speech Recognition Pipeline for Zero-resource Language - Published Conference Paper, INTERSPEECH 2025 (in collaboration with the NVIDIA NeMo team)
- Methods to Increase the Amount of Data for Speech Recognition for Low Resource Languages - arxiv, Jan 2025
- Enabling ASR for Low-Resource Languages: A Comprehensive Dataset Creation Approach - arxiv, Jun 2024

HONORS & AWARDS

- Presidential Award for Best Bachelor Student in the field of IT (Oct 2024) - presented by the heads of the Republic of Armenia and Synopsis.
- Selected for the Silicon Valley Exchange Track (Jun 2023) - Palo Alto / Sunnyvale, California, USA.
- Mentor @ Google Summer of Code 2026, DBpedia.

LANGUAGES

Armenian (Native) • English (C1; GMAT 720) • Russian (C1)