VATCHE THOROSSIAN

Senior AI Engineer | LLM & RAG Systems Architect | MLOps & Distributed Inference

+374 41044241 | vatche.thorossian@gmail.com | linkedin.com/in/vatche-thorossian | github.com/vatche-t | Yerevan, Armenia

PROFESSIONAL SUMMARY

Senior AI Engineer with 8+ years across software, data and AI engineering; prior work spans production RAG and MLOps.
Current Cisco Silicon One work covers agentic LLMs, RAG, fine-tuning and AI infrastructure for PD, DFT and MBIST.
Production responsibilities include deployment, monitoring, troubleshooting and reliability through AIOps.

PROFESSIONAL EXPERIENCE

Cisco Silicon One

July 2026 - Present

Agentic LLMs, RAG, model adaptation, AI infrastructure and production AIOps for silicon engineering

Senior AI Engineer

- Engineer agentic LLM workflows across PD, DFT and MBIST in Cisco Silicon One chip engineering.
- Build RAG systems for domain-specific retrieval and LLM-assisted engineering workflows.
- Fine-tune and adapt LLMs for specialized silicon engineering tasks.
- Develop and maintain AI infrastructure for model serving, RAG and agentic workloads.
- Handle production AI operations through AIOps, including deployment, monitoring, troubleshooting and reliability.

Digitain

April 2026 - June 2026

Large-scale iGaming technology company - sportsbook, casino, and platform solutions

Senior AI Engineer / AI Automation Engineer

Yerevan, Armenia - On-site

- Built agentic AI automation for product, marketing and QA teams across sportsbook, casino and platform workflows.
- Developed marketing and competitor intelligence systems covering campaigns, game positioning, odds movements, market availability and operator game rankings.
- Built Playwright-based agentic QA to analyze behavior, surface issues and support structured coverage of complex user flows.
- Automated knowledge capture through AI interviews with specialists and managers, structuring domain knowledge for assistants and internal training.

Khanoumi

March 2025 - April 2026

Large-scale beauty & health e-commerce platform - a leading vertical commerce company

AI & ML Engineer Lead

- Architected a multi-agent reinforcement learning pricing engine for 24,000+ SKUs using sales, inventory and elasticity data; served low-latency gRPC inference and replaced catalog-wide manual pricing review.
- Cut PDP content production time by 60% for 10,000+ listings using RAG attribute extraction, LLM copy generation and image-synthesis orchestration.
- Eliminated the ad-hoc SQL queue for 10+ stakeholders with a text-to-SQL BI platform using LangChain and custom prompts for revenue, margin, churn and conversion KPIs.
- Owned Airflow, MLflow experiment tracking and artifact versioning, GitLab CI/CD, Docker/Kubernetes and Terraform on AWS (Lambda, S3); reduced model releases from multi-day to same-day.

ISRAN

June 2024 - March 2025

AI research and solutions organization - data science & LLM engineering division

Data Scientist & LLM Engineer

- Cut private knowledge retrieval turnaround by 35% for a 200+ person organization using LangChain, Ollama, CrewAI and Qdrant with HyDE-based retrieval, replacing manual document search.
- Led architecture decisions across 3 engineering teams; domain-driven microservices enabled independent deployment of 5+ AI services and removed integration bottlenecks.
- Implemented MLflow and Airflow workflows for training, deployment, monitoring and rollback, reducing deployment errors and cutting release time from multi-day to same-day.

Pizzaro - Golrang Industrial Group January 2024 - September 2024

Subsidiary of Golrang Industrial Group, one of the largest FMCG and retail holding companies in the region

Data Scientist & Product Manager

- Improved inventory workflow efficiency by 30% and cut overstock incidents by 20% with scikit-learn demand models replacing rule-based heuristics.
- Increased customer engagement by 15% and marketing effectiveness by 25% through SQL and Python analysis of 500K+ transactions and behavioral segmentation.
- Aligned 3 cross-functional teams through KPI frameworks, PRDs and a prioritized backlog, accelerating roadmap delivery by 2 sprint cycles.

Selleryar March 2021 - December 2023

Early-stage startup - AI and data tools for seller performance optimization on e-commerce platforms

Full-Stack Python Developer & Data Scientist

- Increased actionable business insights by 20% with Python, Pandas and NumPy pipelines over 1M+ data points; cut weekly reporting from 8 hours to under 2 hours.
- Expanded catalog coverage 3x and cut data acquisition costs by 35% using Scrapy, BeautifulSoup, Selenium and REST APIs built with FastAPI, Flask and Django.
- Increased throughput by 40% and cut infrastructure costs by 25% through Docker/Kubernetes, Nginx and caching; sustained a 5x traffic spike without downtime.
- Cut query latency by 67% (900ms to 300ms) across 10+ seller-facing features through PostgreSQL query-plan and MongoDB schema optimization.

Omid Iranian Holding August 2018 - August 2020

Technology and education holding group

Junior Software Engineer & UI/UX Designer

- Built a centralized full-stack student data dashboard, cutting administrative processing time by 50% for 200+ users.
- Mentored 30+ junior developers; 80%+ completed independent projects within 6 months through structured programming courses.

LANGUAGES & TECHNOLOGIES

Silicon engineering AI: Agentic AI, LLM workflows, RAG systems, LLM fine-tuning, AI infrastructure, AIOps, production AI operations, Physical Design (PD), Design for Test (DFT), Memory Built-In Self-Test (MBIST)

Core Languages: Python (expert), Go, SQL, Bash

AI / ML & LLMs: LangChain, CrewAI, Ollama, RAG pipelines, multi-agent systems, PyTorch, Keras, scikit-learn, Transformers, NLP, reinforcement learning, MLflow, Qdrant, HyDE

Vector & Data: PostgreSQL, MongoDB, Oracle SQL, Microsoft SQL, Pandas, NumPy

Infrastructure & MLOps: Docker, Kubernetes, Apache Airflow, Terraform, AWS (Lambda, S3), GitLab CI/CD, Linux, Nginx

APIs & Architecture: FastAPI, Flask, Django, gRPC, REST, microservices, Domain-Driven Design, Clean Architecture, OOP, Design Patterns

Automation: Playwright, Scrapy, BeautifulSoup, Selenium

EDUCATION

Azad University, Central Branch - M.Sc. Computer Science (Bioinformatics)

Azad University, Central Branch - B.Sc. Computer Science

CERTIFICATIONS & TRAINING

Data Scientist & LLM Engineering - ISRAN · Product Management Bootcamp - Quera · Deep Learning - Udemy · Design Patterns - Udemy · ML, NLP, Python - Coursera / Udemy