VATCHE THOROSSIAN

Senior AI/ML 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/ML Engineer with 7+ years designing and shipping production-grade LLM platforms, RAG pipelines, multi-agent systems, and MLOps infrastructure at scale. Built and owned end-to-end AI systems serving 24,000+ SKUs and 1M+ records across large-scale e-commerce — from feature engineering and model training through real-time inference and deployment. Led AI architecture decisions across 3 engineering teams, defined technical roadmaps, and delivered systems still running in production. Seeking a Senior AI/ML Engineer role where deep LLM and systems expertise translates directly into measurable product impact.

PROFESSIONAL EXPERIENCE

Khanoumi March 2025 – Present

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

AI & ML Engineer Lead

- Architected and shipped a multi-agent reinforcement learning pricing engine across 24,000+ SKUs from scratch — historical sales, real-time inventory signals, and demand elasticity ingested via Python/Pandas/PostgreSQL pipelines, served as a low-latency gRPC microservice; eliminated manual pricing review for the entire catalog.
- Reduced PDP content production time by 60% for 10,000+ product listings by building an end-to-end AI content system: RAG pipeline for structured attribute extraction, LLM prompt engineering for copy generation, and image-synthesis orchestration for brand-consistent product visuals.
- Eliminated the ad-hoc SQL request queue entirely for 10+ non-technical stakeholders by building a natural-language-to-SQL BI platform (LangChain, custom prompt engineering) enabling direct self-serve querying of revenue, margin, churn, and conversion KPIs.
- Owned full MLOps stack: Apache Airflow DAGs, MLflow experiment tracking and artifact versioning, GitLab CI/CD, Docker/Kubernetes, and Terraform IaC on AWS (Lambda, S3) — reduced model release cycle from multi-day to same-day deploys.

ISRAN June 2024 – March 2025

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

Data Scientist & LLM Engineer

- Cut internal knowledge retrieval turnaround by 35% across a 200+ person organization by building enterprise RAG pipelines over private datasets — LangChain, Ollama-based LLMs, CrewAI multi-agent orchestration, and Qdrant vector database with HyDE-style embeddings replacing manual document search entirely.
- Led architecture decisions across 3 engineering teams by designing domain-driven microservices for data-intensive AI workloads, enabling independent deployment of 5+ AI services and removing cross-team integration bottlenecks.
- Implemented full MLOps lifecycle with MLflow and Apache Airflow — training, deployment, monitoring, rollback — reducing deployment errors and cutting release time from multi-day cycles 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% by building scikit-learn ML demand models replacing rule-based heuristics across operational processes.
- Increased customer engagement by 15% and marketing effectiveness by 25% by analyzing 500K+ transaction records (SQL, Python), surfacing behavioral segments that shaped campaign targeting strategy.
- Aligned 3 cross-functional teams on a single prioritized backlog through KPI frameworks and PRDs, 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

- Drove a 20% increase in actionable business insights by building Python/Pandas/NumPy analytics pipelines over 1M+ data points, cutting weekly reporting effort from 8 hours to under 2 hours.
- Increased product catalog coverage by 3x and reduced data acquisition costs by 35% by building automated scraping pipelines (Scrapy, BeautifulSoup, Selenium) and RESTful APIs (FastAPI, Flask, Django).
- Boosted system throughput by 40% and cut infrastructure costs by 25% via Docker/Kubernetes scaling, Nginx load balancing, and server-side caching — sustained a 5x concurrent traffic spike without downtime.
- Reduced average query latency by 67% (900ms to 300ms) across 10+ seller-facing features by optimizing PostgreSQL query plans and MongoDB schemas for high-volume performance data.

Omid Iranian Holding August 2018 – August 2020

Technology and education holding group

Junior Software Engineer & UI/UX Designer

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

LANGUAGES & TECHNOLOGIES

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

Vector & Data: Qdrant, 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, gRPC, REST, Microservices, Domain-Driven Design, Clean Architecture, OOP, Design Patterns

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