

Levon Barseghyan

barseghyan.levon@gmail.com | (818) 640-4390 | linkedin.com/in/levonbars

SUMMARY

Machine Learning Engineer with 6+ years building large-scale production systems, ML platforms, and iOS/Android mobile apps, specializing in LLM agents, multi-agent systems, and model evaluation. Lead the evaluation, observability, and experimentation framework for Amazon's agentic AI platform—farms of collaborating LLM agents in production—and co-author of research on multi-agent text-to-SQL. Deep experience across LLM-as-judge evaluation, RAG, knowledge distillation, distributed tracing, and ML infrastructure on AWS, plus deep mobile (iOS/Android) and geospatial engineering expertise.

EXPERIENCE

Software Development Engineer, Machine Learning

January 2026 – Present

Amazon

Los Angeles, CA

- Built and lead the evaluation and observability framework for an internal agentic AI platform that runs farms of collaborating LLM agents for geospatial inference at Amazon delivery scale—the org's first agent-evaluation capability, now adopted by 4 teams across the science org and supporting 5 projects.
- Built end-to-end distributed tracing across the full agent-farm lifecycle—orchestrator routing, agent-to-agent delegation, and worker agents—capturing every LLM call, tool invocation, and reasoning step for debugging and quality measurement.
- Designed an experimentation system comparing agent configurations (model, prompt, tools) on shared inputs, scoring against ground truth with deterministic and LLM-as-judge scorers plus regression detection.
- Re-architected the trace backend from a Postgres-backed Langfuse prototype to a custom DynamoDB- and S3-based store, cutting infrastructure cost ~60% while scaling to far larger trace volumes.
- Established the curated execution-trace dataset powering the platform's knowledge-distillation pipeline, enabling distillation of frontier models into smaller student models to cut inference cost.
- Championed AI adoption on the team by automating on-call workflows with agentic workstreams that pick up assigned tickets and drive them to resolution, improving SLA attainment 30% over the prior baseline while reducing operational load.

Software Development Engineer

March 2022 – January 2026

Amazon

Los Angeles, CA

- Architected an automatic map-caching mechanism for iOS and Android that enabled seamless offline navigation while cutting on-device storage 40%, and built a real-time location-validation algorithm (Kotlin/Swift) that improved delivery-point accuracy in production.
- Engineered a scalable mapping solution that helped drivers identify delivery points faster (part of a team initiative delivering ~\$93MM annual savings) and partnered across orgs on a parking-guidance solution contributing to \$100MM+ in annual savings.
- Resolved 95% of map-rendering issues within 3 months using A/B testing and data analysis, and cut incident response time 25% via improved logging and metrics.

Senior Software Engineer (Contract)

October 2022 – January 2024

FreightRight

Los Angeles, CA

- Led 4 engineers building a full-stack freight-tracking web app serving 5,000+ concurrent users, with Spring Boot auth and a fully automated AWS CI/CD pipeline (Elastic Beanstalk, CloudFront).

Software Developer

April 2020 – March 2022

PlanetBids

Los Angeles, CA

- Migrated a legacy Flash application to Ember (Django backend) within 10 months to meet the Flash Player end-of-life deadline, reaching 100% client-criteria compliance for a smooth cutover.

RESEARCH & PROJECTS

MaGS: Multi-Agent Geospatial SQL

January 2025 – May 2025

Georgia Institute of Technology

Research

- Co-authored a research paper on a hierarchical multi-agent LLM framework for geospatial text-to-SQL, enabling non-technical users to query and visualize spatial data; manuscript prepared in IEEE Access format.
- Designed a supervisor-led architecture with specialized agents (selector, decomposer, refiner, visualizer) using RAG over PostGIS documentation to generate, validate, and execute spatial SQL.
- Raised end-to-end query accuracy to 83% versus 67% for a LangChain SQL-agent baseline (largest gains on complex multi-table spatial joins: 60% vs. 40%); evaluated across GPT and DeepSeek models and added automatic map visualization (Folium) for non-experts.

EDUCATION

M.S. in Computer Science • Georgia Institute of Technology

GPA 4.0

Coursework: Machine Learning, Deep Learning, Computer Vision, Reinforcement Learning, AI Agents, LLMs, NLP.

B.A. in Computer Science • University of California, Berkeley

GPA 3.8

SKILLS

LLM & Agentic AI: LLM agents, multi-agent systems & orchestration (agent farms), RAG, prompt engineering, agent evaluation, knowledge distillation, LangChain, LangGraph, AWS Bedrock.

ML & Evaluation: LLM-as-judge & deterministic scorers, experimentation & A/B evaluation, regression detection, distributed tracing (OpenTelemetry), PyTorch, TensorFlow, NumPy, Pandas.

Data & Infrastructure: vector stores, Postgres/PostGIS, AWS (EKS, Elastic Beanstalk, CloudFront), Kubernetes, Docker, Jenkins, CI/CD.

Languages: Python, Java, JavaScript, TypeScript, C, Kotlin, Swift, SQL, HTML/CSS.