Giorgi Giorgobiani

Senior AI Platform Engineer

New York, NY (relocating) · giorgigiorgobiani54@gmail.com · linkedin.com/in/giorgi-giorgobiani-282883153 · github.com/Zuzuna54

SUMMARY

Senior AI platform engineer with 9 years building production systems, specialized in LLM and agent infrastructure. Architect of multi-tenant conversation-intelligence and multi-agent orchestration platforms processing thousands of hours of conversation data in production. Depth in agent runtimes, retrieval-augmented generation (RAG) over pgvector, streaming voice pipelines, multi-tenant data isolation, and infrastructure-as-code on AWS. Ships systems that stay up, stay cheap, and stay observable.

SKILLS

AI & LLM Systems: Multi-agent orchestration, agentic workflows, agent runtimes, RAG pipelines, vector search (pgvector), embeddings, semantic clustering, Model Context Protocol (MCP) servers, streaming STT/TTS pipelines, prompt versioning, LLM observability, cost and latency optimization, OpenAI / Anthropic / Gemini / Mistral APIs, open-weight TTS/STT

Cloud & Infrastructure: AWS (ECS, SQS, Lambda, S3, RDS, AppSync, CloudWatch, ECR, EC2, VPC, CloudFront), Pulumi (TypeScript), CDKTF, Docker, Kubernetes, GitHub Actions, Serverless, GCP

Data: PostgreSQL 17 (Row Level Security, pgvector), Redis, FalkorDB, Neo4j, DynamoDB, MongoDB, Elasticsearch, Kafka, Prisma, TypeORM

Languages: TypeScript, JavaScript, Python, C/C++ (embedded firmware), Java, SQL, Cypher

Backend & API: Node.js, Express, FastAPI, GraphQL, AppSync, Flask, Vert.x

Frontend: React, Next.js, React Native, Redux, Tailwind, D3.js

EXPERIENCE

Sunny Labs Inc — Founder and Principal Engineer

August 2025 – Present

- Architected the LUMI AI companion platform end to end: custom ESP32-S3 hardware, a multi-layer memory system, and a multi-tenant cloud backend licensed as a white-label “brain” platform to third-party brands.
- Built a multi-layer agent memory architecture across Redis, PostgreSQL, and FalkorDB spanning short-term context, episodic facts, emotional state, and a long-term relationship graph; reprocesses the full history of prior sessions into each new one, with no retention ceiling, to maintain consistent long-horizon recall and personalization.
- Engineered a personality-evolution subsystem using background agents that analyze conversation metadata and update character configuration over time.
- Built a low-latency streaming voice pipeline (streaming STT to LLM to custom TTS), reaching sub-500ms for real-time dialogue.
- Designed multi-tenant isolation with PostgreSQL Row Level Security guaranteeing per-device memory separation across licensed brands; took hardware from concept to first functional module in under 4 weeks.

Builders Studio — Senior Platform Engineer

October 2025 – April 2026

- Architected a multi-tenant conversation intelligence system converting 100 hours/week of calls (~120 calls) from meeting recorders into structured business intelligence for 3+ venture studio teams.
- Built a 9-worker, SQS-driven processing pipeline (transcription, speaker ID, summary, learning extraction, collage, reflection) across 7 queues with 7 dead-letter queues, bounded retries, and a shared concurrent message-pool library as the reliability backbone.
- Designed a 4-stage signal promotion ladder (candidate, emerging, validated, decision-grade, never demoting) lifting per-call learnings into cross-call signals and 31 prerequisite-chained artifact types across 6 categories, evaluated in 2 database queries.
- Built retrieval over 3,072-dimension OpenAI embeddings in pgvector spanning transcript segments, conversation summaries, and reflection insights, powering cross-portfolio semantic search and pattern detection.
- Shipped a Gemini 2.5 Pro chat assistant with a 24-tool function-calling registry (search, RAG, list, detail, temporal) using thinking-token streaming and parallel tool dispatch over SSE, plus an MCP server exposing the intelligence layer to external AI agents.
- Guaranteed complete data isolation between portfolio companies via PostgreSQL 17 Row Level Security in a single-deployment multi-tenant architecture, with all AWS infrastructure managed as code in Pulumi (TypeScript) at full dev/prod parity.
- Cut compute cost ~94% by running ECS on EC2 rather than Fargate across 3 capacity providers tuned per workload shape (compute-optimized AI workers, micro ingestion, ARM64 for FFmpeg), autoscaled on an SQS backlog-per-task metric.

Cere Network — Senior Software Engineer

April 2025 – October 2025

- Built the NLP vertical on Cere's decentralized data platform: a 10-agent pipeline where a ConciergeAgent meta-orchestrator fans batched events out to 7 Gemini-backed AI agents (sentiment, embedding, toxicity, spam, emoji, topic, relationship) and 3 coordination agents in parallel, using stateless fan-out with a single-writer merge into a social graph of users, messages, topics, and inferred relationships.
- Implemented content-hash idempotency making the pipeline fully replayable — Postgres can be wiped and rebuilt from source events with identical output.
- Built semantic retrieval over 384-dimension embeddings in pgvector with dynamic topic clustering at a 70% similarity threshold, backed by a 27-table schema with 49 indexes.
- Exposed 5 query operations as MCP (Model Context Protocol) tools through the platform gateway, making the social graph directly queryable by external LLMs.
- Declared pipelines as infrastructure using CDKTF (Terraform CDK, TypeScript) over custom Stream / Raft / Agent resource types, so a new pipeline ships as a stack file and deploys without recompiling any backend service.

YellowPad — Founding Engineer

April 2024 – February 2025

- Founding engineer at a legal-AI startup; built the full platform from zero (AWS, Next.js, Node.js, TypeScript), serving 2,000 users.
- Built the document-intelligence pipeline behind contract due diligence: unstructured.io parsing and chunking, OpenAI embeddings, and LLM-driven extraction, cutting the manual review hours attorneys spent per deal.
- Orchestrated that pipeline as coordinated AWS Lambdas with SQS queueing, offloading heavier LLM processing to EC2 where Lambda limits bound throughput.
- Designed AppSync GraphQL endpoints for real-time updates and offline sync, cutting API response times 40% while preserving cross-platform consistency.
- Orchestrated containerized microservices on EC2 with Docker and Kubernetes; established CI/CD via GitHub Actions; configured VPC, CloudFront, RDS, and S3.
- Led the company through SOC 2 Type 2 certification, implementing security controls across the platform.

Remotasks (Scale AI) — NLP and Prompt Engineering, Contract

September 2023 – April 2024

- Developed prompt strategies and evaluation loops for LLM output quality using spaCy and Pandas, iterating on model behavior against project-defined objectives.

Genentech — Software Engineer

August 2021 – August 2023

- Built a model federation dashboard that let researchers schedule, sequence, and automate model-training jobs, removing most of the manual coordination from the training workflow.
- Visualized the full training-node topology and inter-job relationships as an interactive browser graph over Neo4j, using AntV G6/G2 and Graphin.
- Built full-stack applications (React/Redux, Java/Vert.x, Node.js, GraphQL) deployed on AWS with Docker and Kubernetes for department-wide internal use, backed by PostgreSQL and Neo4j.

EARLIER EXPERIENCE

StoreTasker — Software Engineer (July 2019 – July 2021) — Built a Ruby proxy layer for frontend request capture and third-party integration; led React/Node.js adoption.

Biz2Credit — Software Engineer (July 2017 – July 2019) — Contributed to monolith-to-microservices migration (Node.js/Express, Ruby on Rails); built React/Redux/TypeScript frontends; deployed serverless services on AWS Lambda.

EDUCATION

App Academy, New York, NY — Full-Stack Web Development, 2017

SUNY Plattsburgh, New York — B.S. Business Administration, 2015