Michael Lubich

San Francisco, CA | +1 (415) 275-0094 | michaelle.lubich@gmail.com | LinkedIn | Github | mishalubich.com
U.S. Citizen (W-2) | Authorized to work in the U.S. without sponsorship | **Job preferences:** Onsite, Hybrid, or Remote

Professional Summary

Platform and AI software engineer with **ex-Apple** and **ex-Walmart** background deploying production-grade LLM and AI systems. **Tripled backend throughput** on Java and Spring Boot microservices at Walmart and kept **macOS storage releases stable across 100M+ devices** at Apple. Replaced a vendor agent with one that has never written a wrong record across **10,000 support cases/month**, cutting cost per case from $2.00 to ~$0.02 (**saving $240K/yr**). Runs on-prem open-weight models on **vLLM** behind carrier-scale platforms answering questions in **under 200ms** across billions of records. Technical partner to a $10M+/yr events business and publishes open-source agent tooling.

Skills

Languages	Python, TypeScript, JavaScript, Java, Go, Rust, C/C++, SQL, Bash, HTML/CSS
Backend & APIs	Java/Spring Boot, Spring MVC, Hibernate/JPA, FastAPI, Flask, Django, Node.js/Express; REST, GraphQL, gRPC, WebSockets; microservices, serverless (AWS Lambda), event-driven (Kafka, RabbitMQ); distributed systems, system design, concurrency, caching, data structures and algorithms, OOP, SOLID, design patterns, DDD
Frontend & Full-Stack	React, Next.js (App Router, SSR/ISR), TypeScript, Tailwind CSS, Redux/Zustand, Angular, responsive/accessible UI, Supabase, Stripe payments and webhooks
AI / ML Engineering	vLLM, PyTorch, scikit-learn, LangChain, LangGraph, agentic workflows, agent loops and harnesses (Claude Code, Hermes, OpenClaw), context engineering, LLM fine-tuning (LoRA/QLoRA), MLOps and ML pipelines, RAG, embeddings, Vector DBs (pgvector, FAISS), Model Context Protocol (MCP), Pydantic, MLflow, LangSmith
Data & Streaming	Kafka, ClickHouse, ElasticSearch, Spark, Airflow, Postgres, MySQL, Oracle, Redis, MongoDB, DynamoDB, Snowflake, BigQuery; ETL pipelines, SQL query tuning, schema design
Cloud & DevOps	AWS (EC2, ECS, Lambda, S3, RDS, Bedrock, SageMaker), GCP, Azure; Kubernetes, Docker, Terraform, Ansible, Helm; CI/CD (GitHub Actions, Jenkins), Git, Linux, Agile/Scrum, code review
Testing & Reliability	pytest, JUnit, Playwright, Vitest/Jest; TDD, unit/integration/E2E testing, code-coverage gates, load testing; Prometheus, Grafana, OpenTelemetry, Datadog, SLOs, incident response
Security	OWASP Top 10, RBAC, IAM, OAuth 2.0, SSO/SAML, secrets management, least-privilege tool access, prompt-injection defense, tool-abuse guardrails

Education

- **University of California, Berkeley** | *Bachelor of Arts, Computer Science; GPA 3.7* — Berkeley, CA | Aug 2019 – Aug 2023

Professional Experience

- **Independent Consulting (W3 Sourcing, ERIA, LUPFR)** | *Technical Partner & Software Engineer* — San Francisco, CA | Mar 2026 – Present
  - **Saved $240K a Year on Support for a Major Water Filtration Company** Support was buried in cases and Salesforce Agentforce charged $2.00 to handle each one. Built a triage agent where the model suggests and plain Python decides what actually gets saved, so it has never written a wrong record across 10,000 cases a month. Cost per case fell from Salesforce Agentforce's $2.00 to about $0.02, saving roughly $240K a year and returning 40 hours a week to the team.
  - **Replaced Manual Revenue Work at a $10M+/yr Business (ERIA)** An events company serving Anthropic, OpenAI, and Johnson & Johnson was running ticketing and revenue by hand. Lead engineer on the React, Next.js, and Supabase store that replaced it, handling payments, refunds, and duplicate charges safely, plus an admin agent for the repetitive RSVP work. Shipped behind 2,895 tests. (www.eria.co)
  - **Made 19,029 Candidates Searchable (W3 Sourcing)** No recruiter can read 19,029 profiles. Built the TypeScript funnel that finds, screens, and scores them, with a chat agent whose server checks every quoted claim against the real profile before showing it, so recruiters never repeat something a candidate never said. (www.w3sourcing.com)
  - **Handed a Non-Technical Team the Keys (LUPFR)** This team needed an engineer every time they published an event. Made one link do the whole job and handed over a portal they run themselves, with analytics and CSV export. (lupfr.com)
- **Polaris Wireless** | *Platform & AI Software Engineer* — Santa Clara, CA | Sep 2024 – Jul 2026
  - **Cut Pre-Release Verification from Days to Minutes** Releases waited on manual lab setup and thin regression coverage. Built an agent that books and frees lab machines on request plus a pipeline that generates its own test cases, taking coverage from 40 to 300+ scenarios and pre-release verification from days to minutes, and teaches the practice internally.
  - **Gave Analysts Direct Access to Carrier Data** Analysts waited days on an engineer for every question about carrier location data. Owns AigisQuery, the plain-English query platform they now use themselves, on Model Context Protocol (MCP) tool servers under least-privilege access, each person reaching only the data they are cleared to see.
  - **Cleared Agents to Run Unattended in Production** An agent that writes to live systems needs proof it is safe before anyone signs off. Set the guardrail architecture (loop detection, step budgets, prompt-injection defense, write allowlist) that cleared these agents to run unattended in production.
  - **Cut Inference Cost by Leaving Frontier APIs** The frontier-API bill grew with every query. Moved production workloads onto in-house open-weight models (DeepSeek, Qwen on multi-GPU vLLM) with deterministic RAG grounding, cutting inference cost and driving SQL syntax errors to near zero.
  - **Held Queries Under 200ms Across Billions of Records** Agent tool calls time out when the data layer is slow. Scaled Kafka, ClickHouse, and ElasticSearch over billions of real-time spatial records, holding query latency under 200ms at full production load.
- **Apple** | *Software Development Engineer in Test, Core OS* — Cupertino, CA | Jan 2023 – Jun 2024
  - **Tripled Testing Throughput** Manual QA was the bottleneck on every macOS storage release. Migrated and rewrote 20+ aging test scripts in Python, tripling how much testing ran per cycle.
  - **Protected 100M+ Macs from Storage Bugs** A filesystem bug that ships to 100M+ Macs cannot be taken back. Worked 1,100+ high-priority tickets down to root cause so APFS updates went out stable.
  - **Took the Manual Steps Out of Releases** Release steps done by hand do not scale and do not repeat. Automated validation and release with Ansible, cutting manual work 30% and tripling deployment speed.
  - **Stopped the Same Bugs Recurring** The same defects kept coming back release after release. Turned those recurring patterns into reusable, parameterized Python test suites, cutting duplicated test code 40%.
  - **Got Three Teams to Agree on Priorities** Product, engineering, and QA disagreed on what had to be fixed before shipping. Laid out what was broken and what it would cost, which lifted throughput 20%, and wrote the troubleshooting guides that halved new-hire ramp time.
- **Walmart** | *Software Engineer Intern* — Sunnyvale, CA | May 2022 – Aug 2022
  - **Tripled Backend Throughput** Traffic was outgrowing the existing services. Rebuilt them in Java, Spring Boot, and Hibernate on Kubernetes with Kafka, tripling throughput with Datadog and Grafana monitoring in place from day one.
  - **Made Large Data Lookups Fast** Designed serverless REST and gRPC APIs on AWS Lambda, RDS, and Redis so large lookups returned faster and failed less often.
- **Lawrence Berkeley National Laboratory** | *Software Engineer* — Berkeley, CA | May 2021 – Sep 2021
  - **Gave Researchers Back 200+ Hours a Year** Researchers were rerunning experiments they could not reproduce. Built a versioned pipeline in Python, Docker, and Airflow that automates the model tuning and records every run, saving 200+ hours annually.
- **Honda Innovations** | *Software Engineer Intern* — Mountain View, CA | Jan 2021 – May 2021
  - **Automated Medical Delivery Routing** Medical supply routes were planned by hand. Built a model that predicts the fastest route and shipped it to AWS behind an automated GitHub Actions, Docker, and Terraform pipeline.

Projects

- **Open-Source Agent Tooling (74 public repos)** An AI agent is only worth as much as the systems it can safely reach. Publishes 6 Model Context Protocol servers and 5 CLIs that let agents drive mail, iMessage, WhatsApp, Notes, Jenkins, Bitbucket, Vercel, and Git worktrees, each scoped to only what it needs, shipped to real users through Homebrew and PyPI with Rust-accelerated local search.
- **Multi-Agent Data and Research Pipelines** Built CrewAI pipelines where separate agents research, analyze, and audit each other's work to produce synthetic training data, backed by a Tokio pub/sub event bus in Rust.
- **Document Agent That Resists Tampering** Invoices arrive as email attachments, and anyone can hide instructions inside one. Built an intake agent on the OpenAI Agents SDK that reads fields buried in embedded images while treating the document itself as untrusted.
- **Cheaper Long-Running Agents** Long autonomous runs get expensive fast. Wrote a portable harness that sends easy work to cheap models and hard work to expensive ones, then checks the output instead of assuming it is right.

Publications

6 peer-reviewed papers in machine learning and AI for environmental science (Google Scholar)