

Anthony Chaikovsky

Los Angeles, CA | anthonychaikovsky@gmail.com | (310) 334-9802 | [linkedin.com/in/aj-chaikovsky](https://www.linkedin.com/in/aj-chaikovsky) | github.com/AJChaikovsky

EDUCATION

University of Southern California

B.S. in Computer Science and Business Administration

Los Angeles, CA

May 2027

Relevant Coursework: Principles of Software Development, Introduction to Algorithms and the Theory of Computing, Data Structures and Object-Oriented Design, Discrete Methods in Computer Science, Database Systems

PROJECTS

Surf Alerts: Surf Forecast Alert Platform | github.com/AJChaikovsky/surf-alerts

2026

React, TypeScript, FastAPI, PostgreSQL, SQLAlchemy, Alembic, APScheduler, Leaflet

- Engineered a full-stack surf forecasting platform using React, FastAPI, and PostgreSQL, with configurable alert rules, scheduled forecast evaluation, and daily notification deduplication.
- Prevented duplicate emails by reserving each alert's surf day under a PostgreSQL unique constraint (INSERT ... ON CONFLICT DO NOTHING) before sending, so overlapping or repeated worker runs can't notify twice.
- Isolated user data by scoping every alert, match, and private-spot query to the signed-in owner and returning 404 for other users' records, alongside CSRF tokens, login lockouts, and server-side session revocation.
- Tested failure modes rather than just happy paths: cross-user access, CSRF tokens from other sessions, threshold boundaries, missing wind data, and forecast API failures that skip one spot without stopping the run.

mini-vLLM: LLM Inference Engine from Scratch | github.com/AJChaikovsky/mini-vllm

2026

Python, PyTorch, CUDA, FastAPI

- Implemented transformer inference for Qwen2.5-0.5B-Instruct from scratch, including RMSNorm, rotary positional embeddings, grouped-query attention, and SwiGLU; matched Hugging Face token-for-token on five prompts.
- Built a PagedAttention key-value cache and continuous-batching scheduler with sequence admission, retirement, preemption, lazy block reservation, and chunked prefill.
- Exposed an OpenAI-compatible streaming API with sampling controls and client-disconnect cancellation that releases cache blocks within one scheduler step.
- Reached 112.04 tokens/s at concurrency 32 on an RTX 2080 Super — 6.1x the sequential baseline — while holding median time to first token to 0.29s versus 10.30s for static batching.

LA Hacks 2026: Choices Counselor | github.com/AJChaikovsky/LA-Hacks-project

Oct 2026

Python, Fetch.ai uAgents, ASI:One, Pydantic, LA Metro GTFS

- Worked across the entire project on a 3-person team: a Spanish/English multi-agent counselor that helps LAUSD families choose magnet programs, estimate admission points, and avoid costly application mistakes before the deadline.
- Implemented deterministic Python decision engines for program eligibility, admissions scoring, and application validation, separating rule-based calculations from LLM-generated responses to improve factual reliability.
- Built a 49-program dataset where every value carries a source citation and unverified values stay null instead of being guessed; hardened edge cases (out-of-district families, missing grades, over-limit lists) with property-based tests.
- Added specialist agents for LA Metro bus routes (GTFS) and CDE school data, a paid batch-report agent on the Fetch.ai Payment Protocol, and the web chat client.

Beta Leads: AI Lead Generation for Commercial Real Estate

Oct 2026

Python, Next.js, React, TypeScript, Supabase (PostgreSQL), Claude API, GitHub Actions, Resend

- Built an AI lead-generation system for a commercial real estate brokerage, now in active use by the brokerage, that finds LA restaurants, retailers, and service businesses likely to need a storefront soon and surfaces them to brokers.
- Developed an automated Python data pipeline integrating 15 municipal datasets and county property records, with daily ingestion, entity deduplication, and lead qualification for commercial real estate prospecting.
- Made the pipeline safe to rerun and tolerant of failures: saved records are skipped, the HTTP client throttles and retries busy city servers, every Claude response is validated before saving, and failed leads keep their error and retry on the next run.
- Built a Next.js dashboard on Supabase with company-domain sign-in and row-level access rules that stop brokers from editing AI scores or other brokers' notes.

USC Study Room Booking Platform: Full-Stack Web Application

2026

Java, Spring Boot, React, TypeScript, MySQL, JPA/Hibernate, Flyway, REST APIs, Git/GitHub

- Built backend features for a USC study room booking platform supporting reservations, waitlisting, reviews, and course-based study partner matching, modeled in MySQL with Flyway migrations.
- Implemented an automatic FIFO waitlist with Spring Boot, JPA, and service-layer logic that promotes the next student when a reservation is cancelled, exposed through REST endpoints.

ACTIVITIES & LEADERSHIP

Sigma Nu, Epsilon Omicron Chapter

Aug 2024 - May 2025

Executive Council Member (Social Chairman)

- Planned and coordinated chapter events, managing logistics and communication across the membership; collaborated with the executive board to advance organizational goals and strengthen chapter engagement.

SKILLS

Programming Languages: Python, Java, JavaScript, TypeScript, C++, SQL, HTML

Frameworks & Tools: React, Next.js, FastAPI, SQLAlchemy, Alembic, PostgreSQL, Supabase, Spring Boot, MySQL, JPA/Hibernate, Flyway, Android SDK, Fetch.ai uAgents, Claude API, GitHub Actions, pytest, JUnit, Git, PyTorch, CUDA

Concepts: REST APIs, data pipelines, LLM integration, multi-agent systems, authentication and web security, relational database design, automated testing