Lai Jiang

Santa Clara, CA 95054
(626)-495-6862 | jianglllai@outlook.com | LinkedIn

EDUCATION

Northeastern University — Silicon Valley, CA
M.S. in Computer Science (GPA: 4.0/4.0) — *Sep. 2025 – May 2028*

University of Southern California — Los Angeles, CA
M.S. in Engineering Management — *Aug. 2021 – Dec. 2022*

Chongqing University — Chongqing, China
B.S. in Construction Management — *Sep. 2016 – Jun. 2020*

TECHNICAL PROJECTS

MoodVerse — AI-Powered 3D Mood Tracking Web App | moodverse.space — 2026

- Built a full-stack **TypeScript**/**Next.js** (App Router) app end to end: a **React**/**Three.js** frontend, REST API routes with NextAuth **Google OAuth**, and a **Prisma**/**PostgreSQL** data layer, rendering 89 mood types as an interactive 3D galaxy.
- Implemented the 3D galaxy with **React Three Fiber** (log-spiral time layout, emissive glow, bloom post-processing) and profiled frame drops in the canvas mood picker, restoring smooth interaction by caching glow textures on an offscreen canvas.
- Built LLM application features: post-entry reflection prompts and an insights panel that computes mood patterns server-side (day-of-week, time-of-day, streaks, top moods) and generates a natural-language summary from them.
- Replaced the third-party LLM with self-hosted **vLLM** inference, deploying a quantized 7B model on a single 24GB GPU so mental-health data stays local and never reaches third-party APIs.

vLLM Inference Engine — Independent Systems Study — 2026

- Traced vLLM's offline inference pipeline from where a request enters, through the scheduler, down to GPU execution, tracking how data passes between separate processes.
- Studied how PagedAttention eliminates GPU memory fragmentation to serve more concurrent requests, and how the engine keeps CPU and GPU busy in parallel by overlapping scheduling work with GPU computation.
- Ran benchmarks to test vLLM's scheduler by changing it to run prefill and decode in separate batches, then comparing with the default mixed batching on **Qwen2.5-7B**. With steady traffic, mixing cut time to first token from 23s to 0.25s. With a burst of requests, separating gave 12% higher throughput.

Optimal EV Charging Strategy — Algorithm Final Project | GitHub — 2026

- Modeled EV road-trip charging as a shortest-path problem over a **(charger, state-of-charge) state graph** in **Python**, minimizing total trip time over which stops to make and how much to charge at each.
- Built four solvers (greedy, brute force, **dynamic programming**, and **Dijkstra**) and cross-checked them: DP and Dijkstra matched the exact brute-force optimum, confirming correctness.
- Cut runtime $\sim$**5760**$\times$ (14 s to 2 ms) by memoizing each subproblem so it is solved only once.

RESEARCH EXPERIENCE

BenchFlow — Remote
Research & Engineering Contributor — *Sep. 2026 – Present*

- End-to-end full-stack engineering and research on a benchmark that measures AI coding agents' design taste from real designer-agent sessions, built with **TypeScript**, **Next.js**, **discord.js**, and **Supabase**/**PostgreSQL**.

PROFESSIONAL EXPERIENCE

Haulic Endeavor — Burlingame, CA
Junior Project Manager, E-commerce and Retail Launches — *Feb. 2023 – Feb. 2025*

SenseTime — Beijing, China
Project Manager, Artificial General Intelligence Department — *Dec. 2020 – Jul. 2021*

- Supported four reinforcement-learning research teams, including DI-star (a grandmaster-level StarCraft II agent), coordinating across research, product, and client stakeholders to scope AI game requirements.

SKILLS

Languages: Python, Java, TypeScript, JavaScript, SQL, HTML/CSS
Frameworks & Tools: Next.js, React, Three.js, Node.js, Spring Boot, Prisma, PostgreSQL, Supabase, Git
Cloud & DevOps: AWS, Docker, Terraform
AI/ML: vLLM, quantized LLM inference, LLM application development, NumPy
Other: PMP Certified, Bilingual (English, Mandarin)