

[Alan Zeng](#)

510-930-8801

zienzeng0510@gmail.com[Alan Zeng](#)

San Jose, CA

EDUCATION

University of California, Berkeley | Berkeley, CA, USA

- **Degree:** B.A. in Computer Science and Applied Mathematics, Minor in Data Science
- **Period:** Aug. 2021 – Aug. 2022, Aug. 2023 – Aug. 2025
- **Relevant Coursework:** Machine Learning, Artificial Intelligence, Optimization Algorithms, Internet Architecture and Protocols, Cyber Security

SKILLS

- **Languages:** Java, Python, TypeScript, SQL
- **Frameworks:** Spring Boot, Flask, FastAPI
- **Databases:** PostgreSQL (pgvector), MySQL, Redis
- **Tools:** Alembic, JUnit, Postman, Stripe API
- **Cloud & DevOps:** AWS (EC2, App Runner, RDS), GCP, Docker, Git
- **AI/ML:** OpenAI API, Google Gemini, scikit-learn, NumPy, Pandas, RAG, Vector Search

WORK EXPERIENCE

Backend Engineer (Contractor) | Pias Tech | Palo Alto, CA

November 2025 - Present

Project 1: TikTok Shop Creator Discovery Platform (Magic Search)

- **AI-Powered Search Engine:** Architected and independently developed an **end-to-end semantic search system** for TikTok Shop, enabling brands to discover ideal creators from a database of **64,000+ creators** using natural language queries.
- **Vector Embedding Pipeline:** Designed a multi-stage profiling pipeline that analyzes creator attributes (niche, demographics, content style, commercial potential) using **Google Gemini**, then generates **OpenAI embeddings** (1536-dim) stored in **PostgreSQL with pgvector** for efficient cosine similarity search.
- **RAG-Enhanced Retrieval:** Implemented a **Retrieval Augmented Generation (RAG)** layer with query rewriting, where user queries are first transformed to match the profile text format, then candidates are reranked by GPT-4o-mini with contextual explanations for each recommendation.
- **Business Impact:** Solved a critical pain point for brands by replacing manual creator discovery with an intelligent matching system, reducing search time from hours to seconds.

Project 2: USDT-Based Stock Trading Platform

- **Full-Stack Architecture:** Built a **Flask** backend with **TypeScript** frontend and **MySQL** database, implementing a thread-safe operation locking mechanism using Python's **threading.Lock** with timeout support, ensuring all buy/sell orders are serialized to prevent race conditions and double-spending in a multi-user environment.
- **Infrastructure & DevOps:** Architected a production environment using **Docker multi-container deployment** with frontend/backend service isolation, enabling rolling updates without full system downtime and protecting internal API endpoints from external exposure. Maintained a simplified single-server development environment on **GCP VM** for rapid iteration.
- **Production Deployment:** Preparing to launch the platform for 100 initial users managing up to \$100,000 USD in trading volume, with comprehensive transaction auditing and balance monitoring.

AI Backend Engineer | Griting | Palo Alto, CA

September 2025 - November 2025

- **Backend & Payment Integration:** Built backend services on **AWS (App Runner, RDS)** with **FastAPI** for 200+ users, integrating Google **OAuth2** authentication and **Stripe** payment processing.
- **Recursive AI Matching System:** Engineered a tournament-style **RAG** pipeline where **GPT-4** iteratively filters mentor candidates (100 → 5 per round) until converging to a final group, then selects top N matches based on user-provided biography.

Backend Development Intern | JD.com | Beijing, China

January.2024 - May.2024

- **Core System Development:** Built a package management backend using **Spring Boot** and **JPA**, implementing **RESTful APIs** with role-based data visibility for 2,000+ buyers and delivery personnel.
- **Performance Optimization:** Implemented **Redis caching** with Cache-Aside pattern and city-based data partitioning, reducing database read queries by 50%+.

PROJECT EXPERIENCE

AI Agents for Pacman

August.2024 - December.2024

- **Technologies:** Python, NumPy
- **Reinforcement Learning & Search:** Implemented **Q-Learning** with reward shaping and **Minimax** with **Alpha-Beta Pruning**, training agents to navigate **MDPs** and optimize decision-making strategies.
- **Probabilistic Inference:** Built Particle Filters using **Bayesian inference** for real-time ghost tracking under uncertainty.