James Fang

Email | Linkedin | Github | Personal Page

Experience

Architect Chips Jul 2025 - Now

Founding ML & Fullstack Engineer *Palo Alto, CA*

- Startup focuses on building AI agents for chip design, such as spec-to-RTL and physical design
- Architected and designed an end-to-end multimodal retrieval augmented generation (RAG) pipeline for information retrieval (IR) on semiconductor design and spec documents
- Supported text, image, tables, LaTeX equations for multimodal RAG to provide context for AI chip design agents, overcoming challenges with IR on PDFs
- Scaled, optimized and profiled Python multiprocessing pipelines by 170x to increase indexing speed of PDF to markdown + embedding system from 0.1 to 17 pages a second, while ensuring high fidelity
- Utilized Docling for PDF file processing, to serve embeddings, OpenSearch for hybrid keyword and embedding search & index on document chunks, SeaweedFS for file storage on input PDF, and processed images / markdown
- Benchmarked and evaluated large-scale open-source models (e.g., Qwen-2.5 VL 32B, Qwen-3 Embedding 8B) for inference and multi-modality tradeoffs; optimized throughput performance via parameter tuning in vLLM (constraints on model quality, hardware / model size, batch size, context length, KV-cache utilization at 90–95%)
- Orchestrated distributed CPU/GPU server microservices with Python FastAPI and vLLM endpoints for retrieval, embedding generation, and document chunk-by-chunk processing, ensuring fault tolerance and scalability
- Managed and provisioned infrastructure to directly serve the distributed system above, wrapping the setup into Docker container for on-prem deployment with Google

Rx Jot Oct 2023 – Feb 2025

Co-Founder & Fullstack Software Engineer

- Architected and implemented a system to use large language models (LLMs) to provide medical documentation and appeal letters in oncology prior authorization
- Processed more than $4 million in oncology prior authorizations
- Built tools that reduced the time to process for insurance to process oncology authorization requests by half and reduced denial rates by two thirds
- Launched and maintained the Rx Jot frontend application for users and built the entire backend API to handle requests to generate documentation
- Worked with governmental regulations such as HIPAA and implemented stringent data security practices
- Configured the database to handle the volume of data made by healthcare providers for constant real-time access
- Worked on data science and analytics to extract insights to optimize the prior authorization documentation pipeline to further reduce insurance denials
- Added integration with resources such as NCCN guidelines, PubMed research, FDA indications, ICD & CPT/HCPCS coding
- Engineered LLM prompts to model behavior and eliminated hallucination points of failure
- Validated with users and customers to reduce 95% of user friction and input times by 80%

Technical Skills

Languages: Python, JS/TypeScript, Java, C, C++, HTML/CSS (Bootstrap), SQL, Linux Shell
Libraries & Frameworks: React, NextJS, Git/Github, OpenAI/GPT, OpenCV, Pytorch, AWS, Kubernetes, Docker, MongoDB, SQL, Flask
Skills: Generative AI, Prompt Engineering, Fullstack Development, Deep Learning, Machine Learning, Computer Vision, Natural Language Processing

Education

UIUC Computer Science

Master's Degree *Aug 2024 – May 2025*

- Related Coursework: Advanced Algorithms, Computer Security, Machine Learning for Signals, Social Spaces, Advanced Data Structures — **GPA: 3.7**

Bachelor's Degree *Aug 2021 – May 2023*

- Related Coursework: Algorithms & Models of Computation, Data Structures, Database Systems, Machine Learning, Natural Language Processing, Intelligent Agents, Distributed Information Systems — **GPA: 4.0**

Consulting Work

The Residency Jul 2025 - Aug 2025

Fullstack Engineering, Guest Booking Platform

- Engineered an end-to-end guest booking system with integrated guest stay applications, confirmations, payments, reviews for The Residency, a selective program where talented founders and builders live and learn together
- Designed a multi-house application workflow with linked application IDs and a state-managed booking engine
- Integrated secure Stripe Checkout with server-side price calculation and webhook automation to confirm bookings and update related records
- Implemented automatic date conflict detection to prevent double-bookings; validates overlaps against proposed dates and specific room/spot assignments
- Built staff/admin dashboards for real-time booking review, date amendments, status changes, spot management, and bulk actions; created role-based access control (RBAC) for Admins and Community Architects with house-scoped permissions and audited mutations
- Added post-stay NPS ratings and analytics, enabling data-driven quality tracking across homes and cohorts

Roadmap – ADHD AI Task Initiation Jul 2025

Fullstack Engineering / Systems Design

- Rebuilt and productionized the AI task initiation platform, refactoring core systems for performance, reliability, and scalability, making the product launch-ready for real-world ADHD task management use
- Developed a sophisticated task orchestration system with Firestore, supporting hierarchical task breakdown, AI-powered task selection, and real-time state management across multiple user sessions
- Architected a real-time, task-specific AI task support chat system with streaming responses (Gemini 2.5 Flash)
- Designed an AI-powered memory classification system to auto-categorize user interactions into profession, systems, drivers, struggles, and solutions for deep personalization
- Built an intelligent multi-channel notification engine (AWS SNS for text messages, Resend for emails) with AI-driven nudge strategy tailored to ADHD behavioral patterns
- Implemented Google Calendar & Recall AI integration with automated meeting bot deployment, event sync, transcript processing, and task generation
- Created a secure multi-tenant API architecture (Firebase Authentication, token validation, internal secrets) supporting 15+ endpoints for AI processing, notifications, and calendar integrations

Industrial AI Jun 2025 - Jul 2025

Fullstack / Software Engineering

- Rewrote Industrial AI's core product, a manufacturing ticketing system to reduce factory floor downtime by streamlining communication between operators and the maintenance team
- Integrated Zod schema validation into Google Gemini 2.5 outputs, enforcing type consistency and boosting reliability across the ticketing system
- Built a streamlined, low-friction interface tailored for non-technical factory staff, with plug-and-play onboarding that lets new organizations start using the system productively in under 15 minutes
- Implemented WebSockets for real-time updates and circumventing race conditions across multiple users
- Designed, built and deployed a secure backend on Google Compute Engine with Nginx + Certbot
- Integrated organizational role-based access control (RBAC) security, JWT backend authentication, with WorkOS for organizational sign-on

Mosaic AI Labs (YC W25) Apr 2025 – May 2025

Fullstack / Software Engineering

- Improved AI output quality by designing clearer prompts for Google Gemini 2.5, helping users generate more accurate and reliable results
- Added new features such as an agent pause/cancel button, including implementing additional backend state management to support agent control
- Designed and implemented color wheel UI to simplify color selection, replacing manual RGB input and improving usability
- Integrated Mosaic with downstream video editing tools like Kling by crafting effective prompts and coordinating integration logic
- Reported and documented a bug in a downstream service, improving system robustness and communication between services
- Contributed across the stack, propagated feature changes through both frontend and backend components in a complex codebase