

Bingxi (Paul) Jiang

(949) 344-0799 | pauljiang817@gmail.com | www.linkedin.com/in/bingxi-jiang | github.com/Bingxi-Jiang

EDUCATION

University of California, Irvine (UCI) | Irvine, CA

Bachelor's Degree, Double Major: Computer Science; Business Information Management | GPA: 3.822/4.0 | Expected December 2027

TECHNICAL SKILLS

Languages: Python, Java, C++, JavaScript/TypeScript, SQL

Frontend: React, Next.js, Vue.js, HTML, CSS

Backend: Node.js, Spring Boot, FastAPI, WebSocket, REST APIs, Express, Groovy

Databases & Messaging: Redis, MySQL, PostgreSQL, MongoDB, Kafka, Elasticsearch

Cloud & DevOps: GCP, AWS, Docker, Kubernetes, Github Actions, CI/CD, Nginx

AI/ML: PyTorch, Scikit-Learn, Pandas, NumPy, LangChain, LoRA, OpenAI, Gemini, Claude

PROFESSIONAL EXPERIENCE

Dassault Systèmes (Software corporation) | San Diego, CA

Software Engineer Intern | June 2026 – Present

- Engineered a closed-loop automation pipeline that converts legacy Groovy/Geb workflows into maintainable Playwright/TypeScript.
- Architected a modular Playwright/TypeScript framework with reusable API, session, nested-frame, AG Grid, and reporting layers.
- Cut a 23-step enterprise workflow from 40–50+ min to 3–4 min while expanding validation checkpoints from 6 to 23.
- Enabled idempotent reruns with isolated BrowserContexts, storageState, and persisted workflow IDs across user roles.
- Replaced fixed waits with response, DOM, and lifecycle-driven synchronization, retaining trace/video artifacts for diagnosis.

GTechFin Inc (B2 Fintech startup) | New York, NY

Software Engineer Intern | April 2026 – June 2026

- Architected a Python/Type workflow engine and MCP service with rerunnable DAG stages, parallel branches, and persisted handoffs.
- Modeled end-to-end pipeline state in MySQL with normalized tables, foreign keys, transactions, indexes, and idempotent retries.
- Benchmarked MongoDB and Elasticsearch, then built a self-hosted cosine index for 500–5K videos with incremental updates.
- Cut video production from ~3 hours to ~20 minutes across 7 benchmarks against professional and freelance workflows.

Fresh Road Inc (B2B SaaS startup) | San Jose, CA

Software Engineer Intern | September 2025 – December 2025

- Built a real-time Node.js/Express voice backend streaming Twilio audio to Deepgram, LLMs, and CRM services over WebSockets.
- Orchestrated six BullMQ/Redis queues with retries, backoff, delayed jobs, deduplication, and failed-job recovery.
- Cut LLM cost 32% and p95 latency 18% via caching, compressed contexts, task-based routing, and batched/parallel inference.
- Modeled multi-tenant call state in MongoDB with compound indexes, transactions, audit fields, and idempotency controls.
- Traced HTTP, WebSocket, queue, LLM, and CRM paths with correlation IDs, dashboards, Sentry, alerts, and typed failures.

Ping An Technology (Financial corporation) | Guangdong, China

Algorithm Engineer Intern | June 2025 – September 2025

- Architected a six-stage Python pipeline for multi-view ingestion, quality checks, localization, classification, scoring, and reports.
- Built an async inference API with job IDs, status polling, multipart uploads, and structured damage responses.
- Optimized INT8 serving with dynamic batching, cutting p95 latency 38% and raising throughput 50% across fixed benchmarks.
- Built a 2.27M-image data pipeline with deduplication, schema checks, leakage-safe splits, augmentation, and class balancing.
- Packaged prototypes in pinned Docker images and Kubernetes services adopted downstream in systems handling 10K+ daily requests.

PROJECTS

FocusTrail: AI Task Breakdown & Real-Time Monitor | <https://github.com/PST-Protocol/FocusTrail>

- Built Gemini-powered task orchestration system with React and Vite to break complex tasks improving execution efficiency by 43%.
- Built Node.js and Express backend with Server-Sent Events streaming real-time focus telemetry from a macOS osascript agent
- Deployed full-stack productivity platform on GCP with REST APIs for breakdown, monitor sessions, stats aggregation and history.

Noted: Real-Time Multimodal Transcription & Memory System | <https://github.com/Bingxi-Jiang/Noted>

- Architected real-time transcription pipeline streaming 16kHz PCM audio via WebSocket to Deepgram Nova-2, achieving <300ms latency.
- Engineered multimodal memory system combining transcripts, speaker diarization, and Gemini Vision analysis for context preservation.
- Implemented cross-session RAG with embedding similarity search over transcript chunks, enabling context-aware retrieval across folders.

Course Helper: Intelligent Academic Parser & ETL Pipeline | www.course-helper.com

- Built schema-driven ETL pipeline converting heterogeneous syllabi into typed JSON reducing manual extraction by 93% per syllabus.
- Deployed processing service with PostgreSQL backend and Redis caching handling 1K+ syllabus uploads daily with automated validation.
- Containerized microservices with Docker and Kubernetes orchestration deploying on AWS EC2 with auto-scaling handling traffic spikes.

AWARDS & LEADERSHIP

- 1st Place, Streamlit AI Hack; built and demoed an AI application using Snowflake Cortex under time-constrained product requirements.
- Event Organizer, Chinese Union in Computer Science; organized technical events and supported student developer community program.