# Bingxi(Paul) Jiang

(949) 344-0799 | pauljiang137@gmail.com | Availability: Jun. 15 - Sept. 18 | https://www.linkedin.com/in/bingxi-jiang/

## EDUCATION

**University of California, Irvine (UCI)** *(Expected) 06/2027*

*Bachelor's Degree, Double Major: Computer Science; Business Information Management - GPA: 3.725/4.0*

## TECHNICAL SKILLS

**Coding Language:** Python, C++, SQL, JavaScript, Java
**Skills:** PyTorch, Scikit-learn, LoRA, TensorBoard, MLflow, LLaMa-Factory, Prompt Engineering, Tool Orchestration
**Coursework:** Artificial Intelligence, Machine Learning, Database management, Human Computer Interaction, Information retrieval

## EXPERIENCE

***Fresh Road Inc, Software Engineer Intern | San Jose, CA*** *09/2025 - 12/2025*

- Built a production **backend agent framework** in **Node.js** integrating **OpenAI** for conversational flows and **Gemini** for expert reasoning; implemented **tool schemas**, prompt templates, and traceable execution for multi-step agent behaviors.
- Reduced cost and improved responsiveness via **caching**, **prompt compression**, and selective **model routing** (fast model for easy turns, stronger model for hard turns); cut $/conversation by **32%** and p95 latency by **18%**.
- Redesigned a B2B AI-agent auto-callback pipeline; added call ownership attribution, transcript-level **event tracking**, and end-to-end **observability**; improved operational efficiency by **50%**.
- Implemented evaluation and reliability tooling, including **prompt regression suites**, **golden conversations**, and structured **telemetry** (latency, tool error rate, token usage, cost); enabled **A/B testing**, monitoring, and fast rollback when quality degrades.

***Ping An Technology, Computer Vision Algorithm Engineer | Guangdong, China*** *06/2025 - 09/2025*

- Fine-tuned Qwen2.5-VL for vehicle-damage detection on 2.27M samples; standardized data preprocessing, LoRA/QLoRA fine-tuning, and evaluation using LLaMA-Factory + Swift under limited GPU budget; delivered 69% precision / 55% recall.
- Built an automated damage-reporting system with prompt + tool orchestration; integrated multi-view ingestion, structured output schema, and validator-based post-processing to reduce manual review by 20% and improve report consistency by 27%.
- Optimized inference performance with batching, mixed precision(FP16); reduced p95 latency by 38%, increased throughput by 50%, and enforced safe rollout via canary, monitoring, and rollback-ready model registry with MLFlow.
- Designed a low-quality image enhancement module using CLIP-semantic residual guidance with adaptive histogram equalization; improved robustness on low-light/motion blur, lifting recall by +7pts at fixed precision.

***Relational Cognition Lab, Research Assistant | Irvine, CA*** *01/2026 - Present*

- Conducted representation probing on frontier LLMs (Gemini, GPT-4, Llama-3) to evaluate internal consistency of moral/virtue concepts across layers and context lengths; translated findings into deployment-relevant evaluation protocols.
- Built a noise-estimation and significance framework by interspersing stochastic Wikipedia text in prompts; established statistical baselines for virtue-to-morality vector projection reliability, improving reproducibility across runs and model versions.
- Developed experiment infrastructure for dataset curation, batched inference, and metrics dashboards; hardened pipelines against prompt drift and leakage to support consistent comparisons.

***DeepEM Lab, Research Intern | Irvine, CA*** *01/2025 - 06/2025*

- Built scikit-learn baselines (PCA + SVM) to predict lithium-ion battery performance from material ratios; delivered feature pipeline, cross-validation evaluation, and interpretable decision boundaries to guide lab iteration.

## PROJECTS

***Syllabus-to-Schedule: Intelligent Academic Parser & ETL Pipeline*** *12/2025 - Present*

- Built an LLM-native ETL pipeline using **Gemini 2.0 Flash** to extract structured schedules from PDFs, borderless tables, and images; improved extraction quality versus OCR-centric parsing through **VLM layout** + **semantic reasoning**.
- Engineered a dual-stage filtering and conflict resolution engine in **GPT-4o** using a **Trust Hierarchy**; reconciled system assignments and document-extracted dates with **100% logical consistency** in final outputs.
- Built **CI/CD** workflows to validate schemas, run deterministic post-processing, and execute **offline evals** on a golden set; blocked deployments when **consistency**/**latency**/**cost regressions** exceeded thresholds

***Artified Self: Multimodal AI Personal Narrative Engine*** *01/2026 - Present*

- Architected an end-to-end multimodal ML pipeline that ingests high-frequency telemetry, performs temporal segmentation, generates embeddings, and orchestrates long-context LLM reasoning to convert raw user activity into structured life-log events.
- Built a production data + inference stack integrating Google Calendar/Tasks APIs, embedding services, and LLM generation with monitoring of latency, token usage, and failure retries.

***Sora Persona Engine: AI-Powered Prompt Customization System for Sora2*** *10/2025 - 12/2025*

- Designed a data collection + analytics pipeline that scrapes large-scale prompt data via Chrome extension, stores structured histories in SQL, and performs batch LLM analysis to extract high-performing prompt patterns.
- Built an internal prompt-optimization service that exposes distilled prompt strategies through APIs and scheduled batch jobs; added monitoring, versioning, and automated evaluation to continuously improve prompt generation quality.