Bingxi (Paul) Jiang

Phone: (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
Relevant Coursework: Artificial Intelligence, Machine Learning, Information Retrieval, Probability and Statistics, Database Management

TECHNICAL SKILLS

Languages: Python, C++, SQL, JavaScript/TypeScript, Java, Bash
ML/DL: PyTorch, TensorFlow, Scikit-Learn, Hugging Face Transformers, Numpy, Pandas, XGBoost
LLM & Evaluation: LoRA, RAG, Embedding, Quantization, Mixed precision, Prompt Engineering, A/B Testing
MLOps & Serving: MLflow, TensorBoard, ONNX, FastAPI, REST APIs, WebSocket, Docker, Kubernetes, Redis, PostgreSQL
Data & Cloud: Pyspark, Airflow, Kafka, AWS (EC2, S3, SageMaker), GCP, Prometheus, Grafana

PROFESSIONAL EXPERIENCE

Dassault Systèmes (Software corporation) | San Diego, CA
AI Engineer Intern | June 2026 – Present
- Built a closed-loop LLM agent pipeline that generates Playwright/TypeScript code from legacy Groovy/Geb workflows.
- Integrated Playwright and 3DEXPERIENCE MCP servers for live UI inspection, selector synthesis, and iterative failure repair.
- Cut a 23-step enterprise workflow from 40–50+ min to 3–4 min while expanding validation checkpoints from 6 to 23.
- Evaluated agent-generated workflows by execution success, selector reliability, validation completeness, and reproducibility.
- Engineered reusable TypeScript modules for nested iFrames, dynamic grids, multi-user sessions, APIs, and state-aware interactions.

Relational Cognition Lab | Irvine, CA
Research Assistant | January 2026 – Present
- Built a PyTorch evaluation harness over 100K+ prompts for Gemma, Qwen, and Llama, standardizing prompts, metrics, and runs.
- Identified model-family-specific noise sensitivity across virtues, including large creativity-score shifts in selected Gemma variants.
- Probed layer-wise representations with linear classifiers, cosine similarity, logit lens, and activation patching to trace relational signals.
- Designed random-text injection baselines with repeated trials and variance analysis to separate model signal from prompt noise.

GTechFin Inc (B2C Fintech startup) | New York, NY
Software Engineer Intern | April 2026 – June 2026
- Built a Python/Typer MCP pipeline ingesting nine top videos weekly for trend mining, Claude scripting, and HeyGen rendering.
- Built a custom embedding vector store after benchmarking MongoDB and Elasticsearch on retrieval quality, cost, and control.
- Designed structured prompting, JSON validation, quality scoring, and API fallbacks for multi-stage content generation.
- Cut video production from ~3 hours to ~20 minutes, validated across 10 comparisons with professionals and freelancers.

Fresh Road Inc (B2B SaaS startup) | San Jose, CA
Software Engineer Intern | September 2025 – December 2025
- Built a real-time Node.js/Express LLM backend integrating Twilio, Deepgram, OpenAI/Gemini, WebSockets, and CRM services.
- Cut LLM cost 32% and p95 latency 18% through response caching, prompt compression, and model quantization.
- Queued transcription, summarization, CRM sync, and LLM jobs with BullMQ/Redis to isolate provider latency, retries, and failures.
- Instrumented 13 B2B client deployments across healthcare, finance, and SMBs with correlation IDs, dashboards, and Sentry.

Ping An Technology | Guangdong, China
Computer Vision Algorithm Engineer Intern | June 2025 – September 2025
- Designed a multi-stage vehicle-damage pipeline for multi-view ingestion, localization, classification, and report generation.
- Curated a 2.27M-image training set, then fine-tuned Qwen2.5-VL with LoRA for damage recognition, lifting recall 10% and precision 6%.
- Optimized inference with INT8 quantization, and dynamic batching, cutting p95 latency 38% and raising throughput 50%.
- Packaged Dockerfiles and supported Kubernetes deployment of Redis-cached inference serving 10K+ daily production requests.
- Raised recall by 7 points on low-quality images via CLIP-semantic residual guidance with adaptive histogram equalization.

DeepEM Lab | Irvine, CA
Research Intern | January 2025 – June 2025
- Built PCA/SVM classifiers on ~210K battery records to predict whether cells met cycle-life standards, reaching 87% test accuracy.
- Standardized preprocessing, cross-validation, and hyperparameter search with pandas, scikit-learn, and MLflow for reproducibility.

PROJECTS

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.

Longevity LLM Benchmark: Domain LLM Evaluation Suite | https://github.com/Bingxi-Jiang/LongevityLLM-Benchmark
- Designed a 7-task evaluation suite across classification, A/B test, set generation, and regression, scored with accuracy, mean set F1, MAE.
- Built ontology-grounded reasoning scorer validating gene, allele, MP-ID, and longevity-pathway logic beyond final-answer accuracy.
- Built provider-agnostic evaluation harness for GPT, Claude, Gemini, and fine-tuned Qwen with cached inference, and token/latency metrics.

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.
- Added data quality gates with schema validation, anomaly checks and deduplication designing outputs for calendar sync and reminders.
- Deployed processing service with PostgreSQL backend and Redis caching handling 1K+ syllabus uploads daily with automated validation.