

CHENSHUHAO (CODY) QIN

Los Angeles, CA | (213) 922-0925 | chenshuh@usc.edu | codyqin.github.io/CodyQin | github.com/CodyQin | linkedin.com/in/chenshuhao-qin

EDUCATION

University of Southern California (USC) <i>Master of Science in Computer Science</i> • Coursework: Machine Learning, Deep Learning, Natural Language Processing, Data Structures & Algorithms, Operating Systems, Computer Architecture, Computer Networks, Predictive Modeling	Los Angeles, CA <i>Jan. 2026 – Dec. 2027 (Expected)</i>
Duke University; Duke Kunshan University <i>Bachelor of Science in Computer Science; B.S. in Computation and Design</i> • Honors: Apple WWDC23 Swift Student Challenge Winner (1 of 13 in China), Provincial Innovation Grant Recipient (\$3k), DKU Student Ambassador	Durham, NC / Kunshan, China <i>Aug. 2021 – May. 2025</i>

SKILLS

Languages: Python, C, Java, SQL, TypeScript/JavaScript, Swift
Big Data & Lakehouse: Apache Spark, Flink, Iceberg, Kafka, Ray, batch & streaming pipelines, object storage (OSS/S3)
Infrastructure & Cloud: Kubernetes (incl. Operator development), Docker, Linux, Git, CI/CD, MySQL, PostgreSQL, Redis, REST APIs (FastAPI, Flask), React, AWS
AI & Machine Learning: PyTorch, TensorFlow, CUDA GPU programming, LoRA fine-tuning, vLLM, Hugging Face, Pandas, NumPy
LLM & Agents: agent harness & context engineering, retrieval-augmented generation (RAG), LangChain, LlamaIndex, AgentScope, Milvus vector search

WORK EXPERIENCE

Alibaba Cloud (Top-4 Global Cloud Provider) <i>AI Engineer Intern (Agent Infrastructure)</i> • Collaborated with Alibaba's Tongyi Lab to productize Data-Juicer Agents — an open-source agent harness — into natural-language multimodal data pipelines (cleaning, filtering, deduplication) for the broader Qwen ecosystem, deployed on ADL Foundry , the team's Apsara data-lake analytics platform. • Built internal-platform integrations coupling Spark , Flink , and Apache Iceberg lakehouse engines with Data-Juicer Agents, driving batch and Kafka -backed streaming pipelines that turn raw multimodal data into analysis-ready datasets on OSS object storage. • Built an agent-to- Ray distributed execution pipeline over private-cloud object storage (OSS) and Kubernetes , with hands-on Kubernetes Operator development for job lifecycle automation — 2,793 records , 16 operators , ~5 minutes — landed upstream through a merged PR. • Analyzed agent execution traces to localize context bloat; designed a three-layer context-control architecture (tool-output compaction, memory compression, hard token budgets) cutting single-tool context payloads by 82% and stabilizing multi-turn agents on 30k-token models. • Shipped a full-stack agent workspace (React / FastAPI / MySQL) with persistent sessions, execution tracing, conversation rollback, and multi-instance isolation (4 container-image releases); contributed 5 merged PRs (unit-tested, CI green) to the 7k+ Data-Juicer ecosystem, plus 1 proposal adopted by maintainers.	Hangzhou, China <i>Jun. 2026 – Aug. 2026</i>
Blueberry AI <i>AI Engineer Intern</i> Built a Milvus + MySQL feature-extraction and auto-tagging pipeline batch-indexing 200K+ images for landmark / IP / SKU retrieval; benchmarked DINOv3 , Qwen-VL , and BGE-VL under a unified recall / precision / latency framework; delivered a YOLO + vector-search retrieval POC validated for commercial use.	Burlingame, CA (Remote) <i>Nov. 2025 – Jan. 2026</i>
Tsinghua Sichuan Energy Internet Research Institute <i>Algorithm Engineer Intern</i> Engineered a DeepLabV3+ (PyTorch) segmentation model for microgrid fault diagnosis reaching 92% accuracy ; owned the full data lifecycle of 500+ images from preprocessing to deployment, anchoring winning infrastructure bids.	Chengdu, China <i>Sep. 2025 – Oct. 2025</i>
Leiting Games (G-bits) <i>AI R&D Intern (AIGC & Automation)</i> Built a Dify + RPA automated monitoring pipeline (98.85% accuracy , 1,000+ violations detected) and fine-tuned SDXL / Flux LoRA models driving 117K+ installs , ¥24.6M billings (196% ROI) ; led a 5-person team to 3× creative output (2025 Best Practice).	Shenzhen, China <i>May. 2025 – Aug. 2025</i>

PROJECTS & PUBLICATIONS

ChatDKU — RAG Campus Q&A Assistant <i>Python, LlamaIndex, LangChain, Redis</i> Co-founded a university-adopted RAG Q&A agent indexing 12,000+ webpages over a privately deployed LLM; implemented hybrid search (BM25 + vectors) , HyDE , multi-level Redis caching, and SSO web deployment; officially deployed at DKU and documented as a DKU Signature Work.	Jul. 2024 – May. 2025
Publication: <i>Privacy-Preserving Room Occupancy Estimation Using Federated Analytics of BLE Packets</i> — ACM SenSys 2024 (Demo Track, second author), pp. 889–890; 95% real-world accuracy via federated learning over BLE.	