JINGRAN ZHOU

San Francisco Bay Area · jobs@jingran-zhou.com · (510) 717-0910 · jingran-zhou.com

Summary

Senior Machine Learning Engineer at Apple with **5+ years** building production ML systems across Siri intelligence, on-device optimization, and infrastructure. Track record of shipping privacy-preserving models, agent runtimes, evaluation frameworks, and resource-constrained systems to millions of users. Previously built game engine tools at Tencent (**2M+ DAU**), backend systems at J.P. Morgan, and ML research tools at HKU. Specialize in LLM agents, RAG pipelines, on-device ML, time-series forecasting, and quality assurance for AI systems.

Skills

Focus Areas	LLM agents & tool use; Evaluation & safety; Retrieval/memory; On-device power/scheduling; Time-series forecasting; Privacy-preserving ML
Modeling & Methods	RAG; LoRA/PEFT; SFT; DPO/RLAIF; User simulation; Quantization & distillation; Adversarial testing
Systems & Infrastructure	Dataset/metric registry; CI/A/B testing; Distributed tracing; Docker; Kubernetes; FSDP/ZeRO; Flash-Attention; KV-cache
Programming Languages	Python; Java; C/C++; Swift; Objective-C; SQL; R; JavaScript; C#; Haskell; Lua; MATLAB; Shell
Frameworks & Tools	PyTorch; TensorFlow; LangChain; Hugging Face; DSPy; AsyncIO; Spring Boot; Node.js; Django; Flask; Unity; Tableau; Power BI

Work Experience

Apple — Senior Machine Learning Engineer (Siri Core Modeling) Cupertino, CA · Sep 2024–Present

- Shipped fault-tolerant NL-to-Python agent runtime covering **>2,000** intents; integrated tool-generation, clustering, and sandboxed execution for production Siri queries.
- Built an automated self-improvement engine for agent planning via meta-prompting; reduced manual diagnostics/resolution by **85%** through targeted pipelines and workflows.
- Achieved **95%** agent tool-calling accuracy in multi-step flows via zero-shot testing, prompt design, and error-driven iterations; standardized evals and failure taxonomies.
- Reduced invalid-call rate by **40%** and p95 latency by **20%** through routing/rollback policies with distributed tracing for debuggability; while operating under a fixed cost budget.
- Built pre-ship A/B gates with success/latency/cost KPIs; blocked **3** regressions and improved agent task-success rate by **15%**.
- Designed adversarial/zero-shot eval harness with perturbations and argument fuzzing; gated **5** launches and prevented escaped defects reaching production.
- Built human+LLM auto-raters with **≤5%** disagreement; cut eval time-to-signal by **60%** and bootstrapped preference data for DPO/RLAIF.
- Unified disparate evaluation platforms into a single framework; harmonized metrics and made evals reproducible across teams.
- Improved retrieval recall by **10%** and reduced hallucinations by **30%** via **RAG** pipeline + **LoRA**/SFT fine-tuning with user-simulation evals.

Apple — Machine Learning Engineer II (AI/ML) Cupertino, CA · Sep 2021–Sep 2024

- Developed and optimized Large Language Models (LLMs) for Apple Intelligence, implementing Retrieval-Augmented Generation (RAG) to enhance output accuracy and relevance in natural language processing tasks.
- Engineered Low-Rank Adaptation (LoRA) adapters for efficient model fine-tuning, achieving significant performance improvements while maintaining a small parameter footprint.
- Designed comprehensive evaluation pipelines and user simulation frameworks to assess model quality, reduce hallucinations, and improve output diversity across various language tasks.
- Contributed to the development of editing tools that supported human authors in generating high-quality, context-aware content for AI applications.
- Proposed and shipped a device-state-driven modeling approach for on-device intelligence; improved accuracy vs noisy signals with stronger privacy posture.
- Increased OS update success rate by **25%** through time-series capacity planning and forecasting models to predict OS resource utilization; reduced peak contention failures during high-traffic windows.
- Created **three patented power-management systems** that enabled Always-On Display on iPhone 14 Pro.

Apple — Machine Learning Engineer (CoreOS) Cupertino, CA · Aug 2020–Sep 2021

- Developed and shipped **privacy-preserving, on-device ML models** for Optimized Battery Charging, reducing battery aging by cutting high-state-of-charge dwell time by **20%** across millions of iPhones & AirPods.
- Shipped core **OS scheduling enhancements** to manage high-throughput background tasks for **Apple ProRes** video encoding, balancing performance against strict **power and thermal budgets** to enable the feature launch on constrained devices.
- Designed and patented a predictive, **sensor-based control system** to manage the **Always-On display** for the iPhone 14 Pro, meeting stringent **energy efficiency targets** through real-time, on-device ML.

Tencent — Game Software Engineer (Interactive Entertainment Group) Shenzhen, China · Jun–Aug 2019

- Developed Unity Editor tool in C# for spawning objects that accelerated game designers' workflow by **3×** and designed collision avoidance algorithm to prevent spawned object overlapping.
- Taught myself Lua to implement Skill Panel and After-Battle Score Settlement, integrating UI, FX, animation, and logic, impacting **2M+** daily active users when mobile game launched.
- Implemented Unity Inspector to display and translate character attributes at runtime in tree view with expansion/collapse control, enabling game designers to interact with parameters in real-time.

The University of Hong Kong — Research Assistant (Data Engineering Group) Hong Kong SAR · Dec 2018–May 2019

- Built **9 interpretable ML models** trained on features extracted from **54 unstructured court cases** to help colleagues from Faculty of Law better estimate and understand criminal sentencing decisions. Supervised by Prof. Benjamin C.M. Kao.
- Developed Python web application based on decision tree model with Flask and Bootstrap to predict, explain, and visualize sentencing decisions, which **HKU Law & Technology Center adopted** for public law education.

J.P. Morgan — Software Engineer (Corporate & Investment Bank) Hong Kong SAR · Jun–Aug 2018

- Taught myself Spring Boot & Node.js and collaborated with London team on full-stack development of Java-based web monitor which German clients used daily to track millions of financial instruments.
- Optimized Oracle SQL queries to global instrument database from **120s** to **3s** on average; handled software testing on **6+** modules used daily by international clients.
- Led **6-person** team designing end-to-end LSTM-based system with firm-wide scalability for predictive monitoring of business processes; delivered proof-of-concept to trading desk.

CLP Power Hong Kong — Data Science Intern (Center of Excellence) Hong Kong SAR · Jun–Aug 2017

- Developed ARIMAX model in R for time-series forecasting of electricity consumption with anomaly detection throughout **14 Districts** of Hong Kong, enabling company to take precautions against cable faults.
- Created **3 live dashboards** with Power BI used by **>2,400 field electrical engineers** daily to monitor distribution board readings and predict equipment failures.
- Cleaned, classified, and analyzed **3 years** of internal electricity data to identify consumption patterns and fault indicators across Hong Kong's power grid.

Selected Projects

20 Million Particle Simulation on Supercomputer Spring 2020

C/C++, OpenMP, MPI, CUDA | Berkeley Parallel Computing

- Developed 3 parallel collision simulations of **20M particles**; reduced time complexity from quadratic to linear; measured strong & weak scaling on Cori (NERSC) and Bridges (PSC) supercomputers.

Optimized Matrix Multiplication on Supercomputer Spring 2020

C, Assembly, SIMD | Berkeley Parallel Computing

- Optimized matrix multiplication on Cori supercomputer using SIMD, blocking, memory alignment, and loop unrolling to achieve **18 GFLOPS**.

Distributed Hash Table for Genome Assembly Spring 2020

C++, UPC++ | Berkeley Parallel Computing

- Implemented distributed hash table with UPC++ to parallelize de novo genome assembly across multiple nodes.

Fake News Stance Detection Spring 2020

Python, TensorFlow | Berkeley Machine Learning

- Implemented 5-layer neural network using TF-IDF, Universal Sentence Encoder, and cosine similarity; achieved **82%** accuracy on highly-imbalanced dataset.

Optimized Large Integer Addition Spring 2017

Assembly | Princeton Programming Systems

- Assembly-optimized module for adding very large integers that outperforms GCC by **400%**.

Patents & Honors

US Patent 12,141,012: Energy saving for battery powered devices (2024)
US20240077992A1: Sensor-based display power control (2024)
US20240077930A1: Predictive display power control (2024)

Dean's Honors List (HKU) · CLP "Powering a Sustainable Generation" Scholarship · Zhiyuan Scholarship

Honors & Awards

Dean's Honors List — The University of Hong Kong 2016, 2018, 2019
Interdisciplinary Contest in Modeling (ICM) Honorable Mention — COMAP 2017

Education

University of California, Berkeley — M.Eng. in EECS (GPA: 3.96) 2019–2020
Concentration: Data Science and Systems
Capstone: Harnessing Natural Language Processing to Automate Questionnaire Completion in the Finance Industry (advised by Prof. Kurt Keutzer)

The University of Hong Kong — B.Eng. in Computer Science (First Class Honors) 2015–2019
GPA: 3.85 · Dean's Honors List 2015-2019
Final Year Project: Style Transfer on Non-Parallel Text by Iterative Matching and Translation (supervised by Prof. Benjamin C.M. Kao)
Relevant Coursework: Software Engineering, Operating Systems, Database, Networks, Functional Programming, Compilers, Programming Languages, Algorithms, Discrete Math, Linear Algebra, AI & ML

Princeton University — Exchange, Computer Science (GPA: 4.0) Spring 2017
Advisor: Prof. Brian Kernighan
Relevant Coursework: Algorithms & Data Structures, Programming Systems, Multivariable Calculus