JINGRAN ZHOU

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

Summary

Founding engineer of Apple's next-generation Siri agent platform, leading LLM post-training (LoRA/SFT/DPO/RLAIF), synthetic data generation, and agentic AI development from early prototype to production across 1B+ devices. Architected the full training–evaluation flywheel: synthetic data pipelines, alignment via reward modeling, adversarial red-teaming, and multi-agent reasoning — driving planner accuracy from ~60% to 92%+. 3 granted US patents, UC Berkeley M.Eng., HKU First Class Honors. Seeking to bring production-scale post-training and agentic AI expertise to frontier model development.

Skills

Focus Areas	LLM post-training & alignment; Agentic AI & tool-use orchestration; Synthetic data generation; Multi-agent reasoning; Evaluation & red-teaming; RAG & retrieval systems
Modeling & Methods	LoRA/PEFT; SFT; DPO/RLAIF; RLHF; Reward modeling; Synthetic data pipelines; RAG; Quantization & distillation; Adversarial red-teaming
Systems & Infra	LLM training & serving; FSDP/ZeRO; Flash-Attention; KV-cache; Dataset/metric registry; CI/A/B; Docker; Kubernetes
Languages	Python; C/C++; SQL
ML Stack	PyTorch; Hugging Face; vLLM; LangChain; DSPy

Work Experience

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

- Architected and shipped Apple's LLM-based agentic AI platform converting natural language to executable tool calls across **2,000+** intents, with fault-tolerant sandboxed execution serving **1B+** Apple devices.
- Led end-to-end synthetic data generation platform combining LLM-based generation, templated workflows, and human-in-the-loop quality control — driving agentic planner accuracy from ~**60%** to **92%+**.
- Owned LLM post-training pipeline: LoRA/SFT fine-tuning and DPO/RLAIF alignment on Apple's internal foundation models for server and on-device (edge) deployment; reduced hallucinations by **30%** through iterative training–evaluation loops.
- Pioneered human+LLM auto-rating system achieving **≤5%** inter-annotator disagreement, accelerating evaluation cycles by **60%** and bootstrapping preference data for RLAIF reward modeling.
- Architected automated reasoning and self-improvement engine for agentic planning via meta-prompting and self-play, reducing manual error diagnosis by **85%** and improving agent task-success rate by **15%**.
- Led adversarial red-teaming program with perturbation testing and argument fuzzing; gated **5** production launches and established safety standards preventing escaped defects from reaching users.
- Optimized RAG retrieval pipeline, improving recall by **10%**; reduced invalid API call rate by **40%** and p95 latency by **20%** via intelligent routing and rollback policies under fixed cost budget.
- Drove cross-functional strategy across model training, synthetic data, evaluation, and deployment as technical lead for post-training and alignment roadmap, accelerating Siri's evolution into an LLM-powered agentic system.

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

- Developed and optimized Large Language Models for Apple Intelligence, implementing RAG and LoRA fine-tuning for server and on-device deployment, enhancing accuracy across natural language understanding and planning tasks.
- Built synthetic data generation workflows for pre-training and post-training teams, combining LLM-based generation with human-authored templates to produce high-quality training corpora at scale.
- Designed evaluation pipelines and user simulation frameworks to assess model quality, measure hallucination rates, and validate output diversity across language tasks.
- Created three patented power-management and display-control systems, driving innovation from ideation through patent filing and production deployment on iPhone 14 Pro.

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

- Shipped privacy-preserving, on-device ML models for Optimized Battery Charging and patented predictive sensor-based Always-On Display control system on iPhone 14 Pro, reducing battery aging by **20%** across millions of devices.
- Delivered core OS scheduling enhancements for Apple ProRes video encoding, balancing high-throughput background tasks against strict power and thermal budgets on constrained devices.

Patents (3 Granted US Patents)

US 12,141,012 B2:	Energy saving for battery powered devices (2024)
US 12,436,593 B2:	Sensor-based display power control (2025)
US 12,436,594 B2:	Predictive display power control (2025)

Education

University of California, Berkeley — M.Eng. in EECS (GPA: 3.96), 2019–2020
The University of Hong Kong — B.Eng. in Computer Science (First Class Honors, GPA: 3.85), 2015–2019
Princeton University — Exchange, Computer Science (GPA: 4.0), Spring 2017