Vincent Li

Technical Leader & Distributed Systems Architect | Reliable GenAI & ML Infrastructure

San Jose, CA | jyli.dev@gmail.com | vincentli.dev | LinkedIn | GitHub

SUMMARY

Technical leader and distributed systems architect with **10+ years building large-scale infrastructure**, including 8+ years across ML and AI systems. Led platform strategy spanning LLM serving, capacity planning, reliability, observability, and GenAI safety. Known for turning ambiguous cross-organization problems into measurable product contracts, durable architecture, and systems teams can operate at scale.

CORE EXPERTISE

Infrastructure: Distributed systems, high availability, LLM serving, capacity forecasting, accelerator allocation, cost optimization
AI Platforms: Inference serving, training and monitoring infrastructure, AI safety, traffic governance, observability, SLOs
Leadership: Technical vision, 0-to-1 delivery, cross-org roadmaps, executive alignment, mentorship, incident management

EXPERIENCE

Google | Tech Lead - GenAI Infrastructure & Capacity

Apr 2021 - Present
Sunnyvale, CA

Workspace GenAI Foundation | Gmail Spam & Abuse Infrastructure

- Lead engineering strategy for Gmail anti-abuse infrastructure protecting **billions of users** from spam, phishing, and malware.
- Modernize distributed serving for global email classification, with emphasis on low latency, high throughput, safe rollout, and operational resilience.
- Integrate LLM-based detection into the abuse pipeline while extending infrastructure across training, monitoring, serving, and safety.

Vertex AI | Provisioned Throughput & GenAI Serving Infrastructure

- Founded and led Provisioned Throughput, Vertex AI's fixed-cost subscription for guaranteed LLM inference capacity; scaled to **hundreds of millions in ARR in year one**.
- Pioneered quota-tiered LLM inference serving and shipped an executive-priority MVP in **30 days**, establishing a reusable model for differentiated service levels.
- Implemented end-to-end telemetry across metrics, logs, traces, and events, reducing **MTTR by 40%** and enabling platform reliability and cost dashboards.
- Led cross-org demand forecasting and GPU/TPU allocation, reducing supply-demand mismatches by **30%** and eliminating throttling for top enterprise customers.
- Codified engineering principles, UI-backend contracts, automation strategy, and multi-quarter roadmaps to improve delivery across Gemini infrastructure.

Network Infrastructure | Capacity Delivery

- Led a cross-functional automation program for ML hardware network capacity, adapting delivery systems to new product introductions and operational constraints.
- Designed a switch-port reservation system that eliminated manual planning errors and maintained zero downtime for new ML hardware launches.

EARLIER EXPERIENCE

State Street | Tech Lead & Architect (AVP) - Machine Learning

Jun 2018 - Apr 2021
Boston, MA

- Built intelligent email governance using NER and anomaly detection, achieving **99.99%+ accuracy**, preventing sensitive-data leaks, and eliminating 30+ FTE of annual workload.
- Developed time-series LSTM models for securities finance, improving borrow strategies and generating **$750K savings in Q1**.

Panda Electronics | Software Engineer

Jun 2012 - Aug 2014
Nanjing, China

- Developed embedded software and backend services for consumer electronics delivered at manufacturing scale across millions of units.

SELECTED INDEPENDENT WORK

Aether - Safe GenAI Platform - End-to-end reference platform connecting content safety, traffic governance, inference, and observability across four services.

Atlas - LLM Traffic and Quota Gateway - Redis-backed quotas, token-aware rate limiting, priority traffic shaping, streaming reservations, and usage forecasting; 46 automated tests validated locally.

Sentinel - GenAI Safety - Tiered safety analysis with PII redaction, prompt-injection protection, toxicity filtering, and explicit safety compute budgets.

Hyperion - ML Inference - GPU-aware serving, dynamic batching, caching, autoscaling, metrics, tracing, and alerting as a production-oriented reference implementation.

MonitorX - ML Observability - Instrumentation, drift and resource metrics, multi-channel alerting, FastAPI, InfluxDB, and Streamlit; 97 non-API tests pass locally.

vLLM - Contributor - Contributions to inference-serving and observability work in the open-source vLLM ecosystem.

Awesome LLM Infra - Curated guide to pre-training, post-training, inference, optimization, monitoring, diagrams, and structured learning paths.

EDUCATION & RECOGNITION

MS, Computer Science - Rochester Institute of Technology, 2018 | **BE, Industrial Design** - Xidian University, 2012

Google: 5 spot bonuses and 20+ peer bonuses | Microsoft/RIT Coding Competition: 3rd place | Chinese Mathematical Olympiad: 3rd prize

SELECTED WRITING

From Batch Size to TPU Topology: A Capacity Equation for ML Serving

GenAI Capacity Is a Product, Not Just an Accelerator Pool

Where I Put Safety Checks in an LLM Serving Path

PROFESSIONAL DEVELOPMENT

Deep Learning Specialization - DeepLearning.AI | Natural Language Processing - Microsoft | Enterprise Design Thinking Practitioner - IBM

ADDITIONAL

Open-source infrastructure maintainer and technical writer. Competitive strategy gaming: World of Tanks NA Top 10; Brawl Stars NA Rank 1 Trio.