JONATHAN JIN

me@jonathanj.in
linkedin.com/in/jinnovation
github.com/jinnovation
huggingface.co/jinnovation

SELECT WORK EXPERIENCE

Hinge

Staff Machine Learning Engineer — New York, NY, Apr 2024 – Current

Leading the democratization and platformization of AI at Hinge. Spearheading 0-to-1 development of generative AI infrastructure. Previously defined, led, and delivered on traditional-ML model serving strategy from end-to-end.

- Identified need for centralized LLM gateway solution to federate foundation model queries across model porviders and query patterns. Leading partnership with TensorZero team on open-source gateway and optimization solution.
- Drove the implementation and delivery of Hinge's LLM evaluation and monitoring system. Successfully led and closed out build-or-buy evaluation of vendor/OSS solutions for evaluation and prompt engineering/management. Implemented platform based on OpenTelemetry, OpenInference, and Phoenix in collaboration with stakeholders across Backend, AI, and Data disciplines.
- Defined and led the implementation and delivery of Hinge AI's model serving strategy. Collaborated with applied AI/ML teams to coalesce existing approaches/solutions into cohesive platform based on MLflow, Ray, and Spark, reducing time-to-production for new models by ~50%. Defined strategic principles, focuses, and year-long roadmap in close collaboration with platform partner teams in Data and Engineering to achieve general availability in three quarters.

Humane Intelligence

Engineering Advisor — New York, NY, Feb 2024 – Aug 2024

"Fractional" tech lead and architect. Partner with the CEO to define and execute on engineering strategies. Contribute in hands-on capacity for technical proofs-of-concepts.

Spotify

Senior Machine Learning Engineer — New York, NY, Mar 2021 – Mar 2024

Senior member of the ML Platform org. Primarily focused on managed multi-tenant compute infrastructure for Kubeflow and Ray in GCP. Later shifted focus to AI/ML governance and artifact/experiment management solutions.

- Drove early-stage evaluation and implementation of centralized, multi-tenant Ray infrastructure based on KubeRay;
- Led development of Backstage-based model governance solution in preparation for the EU Digital Services Act;
- Defined general-availability roadmap for Kubeflow-based orchestration product, centered on SLOs and SRE best practices.

NVIDIA

Senior Systems Software Engineer, AI Infrastructure — New York, NY, Dec 2019 – Jan 2021

Worked on AI infrastructure for autonomous vehicle research and development.

Twitter

Machine Learning Software Engineer — New York, NY, Aug 2018 – Dec 2019

Member of Cortex, Twitter's central ML platform organization. Worked on: workflow orchestration; experiment management.

EDUCATION

University of Chicago *B.S. Computer Science, B.A Economics* — 2011 – 2015

SPEAKING

Data Council, "Trimming the Long Tail of Production Model Ownership at Hinge" — Apr 2025
Ray Summit, "Accelerating Time-to-Production for ML at Hinge" — Sep 2024
MLconf, "Empowering Traceable and Auditable ML in Production at Spotify with Hendrix" — Mar 2022
KubeCon + CloudNativeCon, "Scaling Kubeflow for Multi-tenancy at Spotify" — Oct 2021

SKILLS

- **AI, ML, and Data**: TensorZero, Phoneix, OpenInference, Ray, PyTorch, Spark, Flink, MLflow, Airflow, Kubeflow, TensorFlow
- **Infrastructure and Distributed Systems**: Kubernetes, Terraform, Helm, Prometheus, OpenTelemetry, gRPC, Docker
- **Enterprise Platforms**: Amazon Web Services (AWS), Google Cloud (GCP), Databricks, Arize
- **Programming Languages**: Python, Go, C++, Java