



AKIRA CHANG

Senior AI Engineer

AI engineer with expertise in building **production-grade AI systems** from **prototype to deployment**. Optimized image generation to **outperform standard frameworks**. Skilled at bridging the gap between cutting-edge **ML research** and **real-world product requirements**, with a strong focus on **cross-functional collaboration**. Combines deep technical knowledge in **ML, GenAI** and **data** with practical engineering discipline to deliver **scalable, cost-efficient infrastructure**.

Experience

2024 - 2026

Kog

Senior AI Product Engineer

- **Achieved a 25% reduction in inference latency** for a production-grade **image generation pipeline** by implementing advanced caching, torch compile, and float8 quantization, while integrating complex workflows including ControlNet, IP-Adapter, Inpainting, and LoRA.
- Created **AI-powered game creation platform** featuring a **LangGraph**-based coding agent with **multi-agent architecture** and **RAG**-powered context retrieval, enabling autonomous game generation and **iterative refinement**.
- Engineered production **ML orchestration servers** with **Python, FastAPI** and **Pydantic**, partnering closely with the frontend team to integrate **NVIDIA A100 GPU-backed ML endpoints** into the product with stable, well-specified request/response contracts and **Azure cloud storage** integration.
- **Reduced infrastructure costs** by implementing shared model caching across ML servers, eliminating redundant loads and **optimizing GPU memory** utilization.
- **Accelerated ML development cycles** by 2x with Gradio-based QA tooling, enabling rapid visual regression testing and real-time output comparison.
- **Scaled synthetic data generation** to 10,000+ samples/day using TypeScript/Puppeteer automation, eliminating manual orchestration across ML servers.
- Enabled semantic image search by **building vector pipelines**, replacing keyword-based retrieval with visual similarity matching, resulting in **improved image search precision** rate across the platform.
- Achieved **8x payload reduction** for **VLM** segmentation masks via custom binary compression, cutting response latency and enabling real-time inference over **Azure Service Bus**.
- Iterated on the product by translating UX needs into ML-backed features with strong focus on **latency, scalability, and maintainability**.

2022 - 2023

Arianee

Data Scientist

- Developed NFT **recommendation engine** using hybrid filtering algorithms to analyze user transaction history and NFT metadata, resulting in **measurable increase in platform sales volume and user engagement**.
- **Reduced fraud risk** by building a **Dagster-orchestrated anomaly detection system** that **automatically flagged** suspicious blockchain transactions, with Neo4j visualizations to trace and track fraudulent activity.
- Designed and **deployed scalable ETL pipelines** on Google Cloud Platform using Dagster to process blockchain data, delivering **actionable user segmentation and persona insights** that informed marketing strategy and product development.
- Architected **real-time analytics dashboard** in Looker Studio monitoring 1-2M+ daily blockchain transactions, integrating BigQuery and PostgreSQL data sources to **enable data-driven decision-making** across operations teams.

Contact

Phone

+33 766453120

Email

akira.shengchang@gmail.com

Address

20 Avenue de Versailles, Paris 75016

Website

<https://akira-chang.vercel.app/>

Github

<https://github.com/AkiraChang>

Linkedin

<www.linkedin.com/in/akirashengchang/>

Education

Master's Degree in Big Data Management and Analytics

CentraleSupélec - Paris Saclay

Bachelor's Degree in Computer Science

Tsinghua University - Beijing

Skills

AI/ML

Transformers, Diffusions, LoRA, LLM, RAG, LangChain, LangGraph, PyTorch, HuggingFace

Backend

Python, FastAPI, Pydantic, Azure, Docker, GCP, Typescript

Data

SQL, BigQuery, Dagster, PostgreSQL, Supabase

Languages

English

Native / Bilingual Proficiency

Chinese

Native / Bilingual Proficiency

Japanese

Professional Working Proficiency