In Keun Kim

ik2619@columbia.edu • (332) 265-6478 • linkedin.com/in/nearkim • github.com/nearKim

PROFESSIONAL EXPERIENCE

Viva Republica (Toss) Seoul, Korea

Machine Learning Engineer, ML Platform Team Jul 2024 – Feb 2025

- Built a DeepFM inference cluster for personalized advertisement using Spring WebFlux(Kotlin), ONNX, Aerospike, handling A/B routing at 6k+ RPS with p95 latency under 70ms.
- Upgraded inference cluster by applying MLflow model packaging, zero-downtime deploys, and automated data integrity tests, cutting release time from 6 hours to 30 minutes and enabling hourly model updates.
- Created an LLM server platform integrating in-house models with Azure OpenAI (ChatGPT) and Milvus vector database, unifying payloads and applying intelligent request batching to boost throughput, deployed in UI/UX evaluation tools and virtual customer centers.

Server Engineer, ML Platform Team Sep 2022 – Jun 2024

- Volunteered to take ownership of a risky, high-impact refactor avoided by others. Utilizing DDD and JVM profiling, removed over 1,500 duplicate lines of code, cut pod memory usage by 25%, decreasing it from 8GB to 6GB. Also reduced endpoint delivery time from one day to under three hours.
- Partnered with non-technical PMs to convert rapidly changing fraud policies into a YAML-based, no-code framework generating Spring Beans for Kafka consumers, facilitating PMs to launch new real-time ML pipelines without developer support.
- Engineered real-time fraud scoring and blocking services powered by Kotlin, Kafka, CatBoost on Triton, and HBase/MySQL, maintaining false positives under 30% and sustained p95 200ms and p999 600ms on 1TB+ of features.
- Deployed JuiceFS on Kubernetes, allowing POSIX access from Airflow Kubernetes Operator, storing 20M+ facial images, and eliminating the NAS cluster.
- Launched Tosst, an image generation service using FastAPI, Stable Diffusion, S3, and MySQL, empowering designers to create images from natural-language prompts and saving about ₩100M (~$72.5K) per month.

Python Developer, FDS Team Jan 2021 – Aug 2022

- Prototyped a fraud detection inference server via Python, LightGBM, and Kafka, processing 150+ RPS in production and seeding an enterprise-level Fraud Detection System.
- Collaborated with the Customer Relations team to redesign an end-to-end fraud case support pipeline employing MySQL OLTP, Hive, Impala, and Slack audit logs, replacing all spreadsheet workflows and driving legal case errors to zero.

Python Developer, VivaManager Team & Workflow Silo Oct 2019 – Dec 2020

- Delivered fraud-case CRM proof-of-concept in three days in Django, introducing the first compensation and demographic dashboards later scaled by FDS.
- Enhanced system reliability by streamlining Django ORM data flow and integrating Datadog and Sentry, cutting incidents from daily system failures to annual events.

XINICS Seoul, Korea

Frontend Developer Jul 2018 – Oct 2019

- Augmented Canvas LMS modules using React (boards, file BBS, auto-grader, live quiz), boosting e-learning engagement for college students.
- Implemented a parameterized Bash shell script automating product rollout with DB and NAS provisioning, allowing seamless deployment across 30 universities in Korea with minimal configuration and zero installation errors.

RESEARCH EXPERIENCE

Columbia University New York, NY

Graduate Researcher (advisor: Baishakhi Ray) Sep 2025 – Current

- Conducting research on Cyber Security and Large Language Models (LLMs) under Projects in Computer Science (COMS 6901E).

SKILLS

Programming Languages: Kotlin, Python, Java, Rust, Scala, TypeScript, JavaScript, C
Software Engineering: Domain-Driven Design (DDD), Hexagonal Architecture, SOLID Principles
Frameworks & Libraries: Spring Boot, Spring Webflux, FastAPI, Django, Django Rest Framework
DevOps & CI/CD: Kubernetes, Istio, Docker/Docker Swarm, Jenkins
Databases & Data Systems: Apache Kafka, Apache Airflow, Aerospike, Milvus, Hadoop Ecosystem (HBase, Spark, Hive, Sqoop, Impala), MySQL, PostgreSQL, MongoDB
Machine Learning: DeepFM, CatBoost, XGBoost, LightGBM, ONNX, MLflow, NVIDIA Triton Inference Server, TensorFlow, PyTorch, TensorFlow Extended

EDUCATION

Columbia University New York, NY

M.S. in Computer Science (Machine Learning Track) Expected Dec 2026

Seoul National University Seoul, Korea

B.S. in Industrial Engineering, Minor in Computer Science Aug 2024