Yosuke Kuroki

+818066733555 | yosukekuroki000@gmail.com | linkedin.com/in/yosuke-kuroki

Role: SR Machine Learning Engineer | NLP, Computer Vision, OCR | Cloud, MLOps

Summary

AI Engineer specializing in machine learning, LLMs, and product development, with a proven track record of building scalable, production-grade AI systems and bringing innovative products to market quickly. My expertise spans zero-to-one prototypes, viral open-source projects, and large-scale deployments—from AI-powered chatbots handling millions of tokens to SaaS, enterprise, and fintech solutions.

Technical Skills

FILEDS: Language Processing (NLP), ML Operations, Image Processing, Text Generation, Image Segmentation, Image Denoising, Object Detection , Sentiment Analysis, Named Entity Recognition (NER), Text Detection, Optical Character Recognition (OCR), RAG, LLMs, Computer Vision, Data Generation, Model Creation, Code Analysis, CUDA (GPU)
LIBS: PyTorch, OpenCV, NumPy, Caffe2, Scikit-learn, Skimage, Scipy, Mxnet, Pandas, PySpark, ELK, TensorFlow
FRAMEWORKS: React.js, Next.js, Vue.js, Tailwind CSS, MUI, Actix, Rocket, Node.js, Express.js, Django, Flask
LANGUAGES: Rust, Golang, JavaScript, HTML5, CSS3, Solidity, Python, C++, TypeScript
DEVELOPER TOOLS: Jupyter, Kubernetes, AWS(GCP), LangChain, Jenkins, Querybook, Docker, CI/CD pipelines, Docker, ESLint, TSLint, Test-first Automation, Selenium
DATABASES: MySQL, MongoDB, Redis, PostgreSQL, GraphQL,
AGILE PRACTICES, TEAM TOOLS: Agile, Scrum, Kanban, Scrumban, Jira, Trello

Education

Tokyo Institute of Technology Tokyo, Japan
Bachelor of Science in Computer Science *Feb.2014 - June 2018*

Experience

SR ML/LLM, Generative AI engineer Jan.2025- Present
Ultim Group *United States*

- Designed a novel GPT-4o-based evaluation framework for long-text generation and summarization, reducing feedback cycles from days to minutes and enabling rapid iteration.
- Spearheaded R&D for ”Client Intelligence,” leveraging LLMs to extract structured insights from historical broker chatrooms and developing a text-to-SQL chatbot for natural language retrieval.
- Built automated ingestion and preprocessing pipelines using Azure AI and cloud functions, and embedded unstructured documents into a vector database for downstream search.
- Scaled an AI document processing platform by 100× (17k+ tenants/year), optimizing UI/UX, refining SME-driven prompt engineering, and doubling analyst efficiency—saving 8.5k+ hours annually.
- Led a major system refactor (10k+ LOC), slashing per-tenant costs by 80% (from $25 + to5$) and reducing latency from 45 minutes to 5.
- Led the development of a CNN-based OCR model optimized for GPU inference, reducing latency by 41% with TensorRT and Docker.

.NET, Full Stack Developer(Contract Position) Nov.2024 - Present
ZOJAX GROUP *United States*

- Full-stack web development using .NET, JavaScript, and modern frameworks to build scalable, high-performance applications for enterprise clients.
- Collaborate with cross-functional teams (product, UX, QA) to design, develop, and deploy end-to-end solutions that enhance user experience and business efficiency.
- Optimize application performance through code refactoring, database tuning, and cloud integration (Azure/AWS).
- Ensure maintainability and scalability by implementing clean architecture, RESTful APIs, and CI/CD pipelines.

SR Machine Learning Engineer, MLOps Mar.2022 - Dec.2024
SPOTCUBE.INC *Riyad, Saudi Arabia*

- Created hyperspectral tree species classifier achieving 30% higher accuracy than RGB baseline
- Implemented LiDAR-RGB fusion pipeline, reducing object detection false negatives by 30%
- Redesigned training infrastructure, cutting model development cycles from 7 days to 20 hours and reducing AWS costs by 80%
- Co-developed graph neural network architecture that reduced 3D point cloud processing time by 50%
- Streamlined annotation workflows, increasing labeling team efficiency by 50-70%

Machine Learning Engineer Mar.2020 - Feb.2022

CTI-Construction Testing & Inspection, Inc. *Gunma, Japan*

- Developed a medicine ranking system leveraging Python, Databricks, and SparkXGBRanker, implementing EDA, feature engineering, and hyperparameter tuning to enhance model accuracy.
- Managed cloud infrastructure on AWS EKS and Azure Kubernetes Service, provisioning resources via Terraform and Helm.
- Designed and deployed scalable data pipelines on AWS S3/EC2 & Google Cloud Storage, optimizing preprocessing and model training for medical imaging datasets.
- Enhanced model performance using data augmentation, transfer learning, and cross-validation techniques, improving robustness for real-world clinical applications.
- Containerized ML applications using Docker, orchestrating secure, HIPAAcompliant deployments with Kubernetes to ensure real-time inference in clinical settings.
- Deployed ranking models as UISE JVM Chassis applications, implementing realtime tracking, alerting, and periodic retraining, strengthening risk and customer engagement strategies.
- Refactored and migrated legacy ML pipelines to Java & gRPC/OIPx protocol, modernizing a low-latency orchestrator for high-frequency trade execution and data consistency.
- Established CI/CD pipelines on AWS & GCP cloud infrastructure, enabling seamless model updates and real-time risk reporting for financial analytics.
- Designed & executed ETL pipelines using Pandas, PySpark, Jenkins, Airflow, Databricks, and Querybook, optimizing data ingestion & transformation workflows across cloud platforms.
- Developed automated ETL workflows on AWS Lambda, EC2, and SageMaker, efficiently handling large-scale financial datasets.
- Tuned Hadoop & Spark configurations, reducing processing time for critical big data operations, improving scalability & cost efficiency.

AI Chatbot, Full Stack developer May 2018 - Feb.2020

Metadata *United States*

- As a dedicated AI chatbot, Frontend Developer with over two years of freelance experience on Metadata, I specialized in creating dynamic and responsive web applications that deliver exceptional user experiences.
- My work involved collaborating with clients to understand their needs and translating them into functional, visually appealing websites and applications.