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, LLM, 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.

Chief Technology Officer Mar.2022 - Dec.2024

SPOTCUBE.INC *Riyad, Saudi Arabia*

- Spearheaded the creation of a hyperspectral tree species classifier, achieving a 30% increase in accuracy over traditional RGB methods, enhancing the company's technological edge in environmental analytics.
- Implemented a LiDAR-RGB fusion pipeline that significantly reduced false negatives in object detection by 30%, improving the reliability of data-driven decisions.
- Redesigned the training infrastructure, reducing model development cycles from 7 days to 20 hours while cutting AWS costs by 80%, thereby maximizing resource efficiency and operational budget.
- Co-developed a graph neural network architecture that halved the processing time for 3D point clouds, streamlining workflows and enhancing project turnaround times.
- Streamlined annotation workflows, resulting in a 50-70% increase in labeling team efficiency, fostering a productive working environment and accelerating project delivery, while mentoring mid-level developers to enhance their skills and productivity.

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.

Machine Learning developer Mar 2019 - Feb.2020

Google *United States*

- Developed a CNN-based OCR Text Detection Model, reducing inference time for scanned documents and card images, optimizing processing speed and accuracy.
- Designed and fine-tuned an OCR pipeline using Tesseract, digitizing and extracting textual metadata from archival labels, handwritten notes, and printed documentation.
- Customized OCR models to recognize specialized fonts and handwriting styles in historical records, enhancing digital metadata enrichment.
- Integrated denoising, text detection, and Tesseract OCR, leading to reduction in inference time for scanned certificate documents and PDFs.
- Developed computer vision-based techniques for character & table detection in document scene images, leveraging OpenCV to enhance document processing efficiency.
- Deployed distributed document processing pipelines on AWS (EC2, S3), ensuring scalability and secure high-volume storage of legal documents
- Developed cloud-based RESTful APIs that seamlessly integrated ML models with legacy e-discovery systems, improving review turnaround time and operational efficiency.

- Developed synthetic training data for logo detection, applying data augmentation techniques using OpenCV & C++ to improve model robustness
- Deployed ranking models with Docker & Kubernetes as UISE JVM Chassis applications, implementing an orchestrator for performance tracking, alerting & retraining.
- Developed a machine learning system for automated classification of physical media (tape formats, film reels, optical discs) via image and audio feature extraction.

AI Chatbot, Full Stack developer May 2018 - Feb.2019

Metadata *United States*

- Integrated real-time performance monitoring dashboards with Datadog, enhancing model observability and troubleshooting capabilities.
- Designed automated ETL & preprocessing pipelines for legal & archival datasets, ensuring efficient data ingestion & transformation on AWS Lambda & SageMaker.
- Conducted load testing with K6 & Locust, documenting findings & optimizing document processing pipelines for scalability & speed.
- Designed automated ETL & preprocessing pipelines for legal & archival datasets, ensuring efficient data ingestion & transformation on AWS Lambda & SageMaker.
- Designed & implemented an end-to-end NLP pipeline to automate the classification, review, and summarization of legal documents, reducing manual e-discovery time.
- Collaborated with legal teams & compliance specialists to validate model fairness & transparency, ensuring ethical AI deployment in litigation support systems.

Certifications

- Deep Learning Specialization
- Certified Analytics Professional (CAP)
- AWS Certified Machine Learning
- Azure Data Scientist Associate
- Certified Associate Developer for Apache Spark

Past Projects

- **Chatbot Building SaaS Platform:** Built a scalable AI agent platform (Next.js, Flask, Firebase) enabling businesses to deploy custom chatbots with 5+ data source integrations (PDFs, APIs, DBs), handling 500+ daily customer inquiries with 92% accuracy. Architected real-time chat for 100+ concurrent users using WebSockets, and implemented semantic search (Vector DB + Knowledge Graph) to reduce response time by 40%.
- **Real-Time Facial Recognition System:** Developed a high-performance backend (Python, FastAPI) processing 1,000+ facial recognition transactions/day, cutting fraud losses by 30% for 100+ clients. Integrated ML models (PyTorch, TensorFlow) with ¡500ms latency, and built real-time dashboards (Power BI) for fraud analytics.
- **Health Assistance Application (LLM):** Architected a comprehensive Health Assistance Application leveraging the Gemini 1.5 Pro LLM model, Retrieval-Augmented Generation (RAG), and LangChain. This solution supports doctors and patients by providing detailed medical information, symptom diagnosis, treatment suggestions, and preventive healthcare advice. The integration of generative AI ensures accurate and personalized health recommendations, significantly enhancing patient care and medical consultations.
- **Chatbot Building SaaS Platform:** Built a scalable AI agent platform (Next.js, Flask, Firebase) enabling businesses to deploy custom chatbots with 5+ data source integrations (PDFs, APIs, DBs), handling 500+ daily customer inquiries with 92% accuracy. Architected real-time chat for 100+ concurrent users using WebSockets, and implemented semantic search (Vector DB + Knowledge Graph) to reduce response time by 40%.
- **Customer Satisfaction Prediction System (MLOps):** Created a predictive analytics system using ZenML and MLflow within an MLOps framework to achieve 93% accuracy in forecasting customer satisfaction. Streamlined pipeline management and deployed models efficiently, providing actionable insights that improved business decision-making and customer retention strategies.
- **AutoML-Studio:** Built a low-code/no-code machine learning platform enabling rapid development and deployment of predictive models. Incorporated MLflow for tracking, XAI (SHAP) for explainability, and data drift detection to ensure model reliability. Deployed with FastAPI and Streamlit, the platform supports classification, regression, time-series forecasting and clustering, empowering non-technical users to build robust AI solutions.

- **Conversational AI Chatbot for Customer Support (GPT-4o, LangChain, Bedrock):** Designed and deployed an enterprise-grade chatbot powered by GPT-4o, integrated with LangChain and AWS Bedrock. Implemented advanced NLP techniques such as intent recognition and dynamic response generation, reducing customer query resolution times by 40% and improving customer satisfaction.
- **AI-Powered Blog Generator (LLM):** Developed an AI-powered content generator leveraging Llama 3.1 8B, AWS Bedrock, and RAG for real-time fine-tuning on dynamic datasets. Integrated with Streamlit for a seamless user interface, the system automated blog creation, editing, and publishing, increasing efficiency in content marketing and engagement for enterprises.
- **Harvestify-AI-Powered-Plant-Health-Assistant:** Created and implemented an AI-driven application leveraging Azure AI Studio and Azure App Services to diagnose plant leaf diseases through real-time image analysis. The system integrates deep learning models to recommend crops and fertilizer solutions tailored to soil and weather conditions, enhancing agricultural decision-making. Deployed on Microsoft Azure, the solution reduced disease detection time by 70%, improved crop yield predictions by 30%, and empowered farmers with actionable insights to optimize productivity and sustainability.
- **Intelligent Document Summarization Tool (LLM):** Developed an Intelligent Document Summarization Tool using transformer-based LLM models like Pegasus to generate concise summaries from lengthy documents while preserving key information. Deployed with FastAPI and Streamlit APP for real-time interaction, supporting formats like PDF and DOCX, and leveraging semantic analysis for relevance extraction. Reduced manual document review time by 60%, boosting productivity in legal and academic workflows.