Yosuke Kuroki

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

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

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 and implemented an end-to-end automation pipeline for payment processing using Selenium and RPA tools to reduce manual intervention and ensure reliable deployment in CI/CD workflows.
- Developed intelligent document parsing scripts integrated with GPT-4o-mini to extract structured information, and leveraged Python for backend processing and cloud-native deployment
- Built a monitoring dashboard that tracks automation workflows and triggers alerts via cloud logging and monitoring services to improve visibility and system uptime.
- Build a cloud-native RPA solution that combines Selenium and Kubernetes to orchestrate scalable and fault-tolerant automation processes.
- Designed a secure automation system using containerized workflows and integrated CI/CD pipelines to streamline form submission and data validation tasks.
- Built automated ingestion and preprocessing pipelines using Azure AI and cloud functions, and embedded unstructured documents into a vector database for downstream search.
- Designed a scalable augmented search generation (RAG) pipeline using a cloud-native vector database and Kubernetes to implement advanced contextual search.
- Implemented cognitive automation by integrating AI models with RPA bots, improving extraction accuracy and contextual decision-making across document workflows.
- Developed AI-powered document search systems utilizing Azure Blob Storage, embedding models, and REST APIs to optimize latency and retrieval relevance.
- Implemented a Graph RAG system with Neo4j on Kubernetes, enhancing relationship mapping for AI-assisted analytics.
- Integrated BERT-based GCN KIE models within LangChain pipelines to improve document intelligence workflows, leveraging Weights, Biases for experimentation.
- Applied Named Entity Recognition (NER) and summarization techniques in Kubernetes-deployed NLP microservices for intelligent document understanding.
- Utilized Llama2 models via LangChain for enterprise content generation, integrating outputs with REST-based business systems.
- Enhanced NLP document processing efficiency using CUDA-accelerated stacks in containerized environments.
- 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*

- As a seasoned Full Stack Developer at Zojax Group, I leverage my extensive experience in both front-end and back-end technologies to deliver robust and scalable web applications.
- My role involves collaborating with cross-functional teams to design, develop, and maintain innovative solutions that enhance user experience and drive business success.

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.