

Zhengxuan YUAN

Applied AI Engineer

Munich, Germany
✉ kevinxyuan@gmail.com
🌐 kevinxyuan.com
in zhengxuan-yuan
🔗 Kevoyuan



Profile

AI Engineer with three years of experience building GenAI and machine-learning solutions for enterprise and industrial environments. Experienced in translating ambiguous requirements into RAG and agentic workflows, AI automation, data pipelines, evaluation logic, and production-oriented cloud services. Strong Python and machine-learning background with hands-on experience in retrieval, structured outputs, validation, fallback handling, cost optimization, FastAPI, Docker, AWS, SQL, and stakeholder-facing AI delivery.

Work Experience

08/2023–Present

Applied AI Engineer, *Applus+ IDIADA*, Munich, Germany

- Delivered AI, GenAI, and machine-learning solutions for external clients and internal engineering teams, translating ambiguous technical requirements into testable AI workflows, prototypes, validation criteria, and practical implementation paths.
- Co-engineered AutodidactAgent, a Case-Based Reasoning agent for enterprise knowledge reuse across technical documents, structured data, and raw artifacts, combining case extraction, retrieval, reusable reasoning skills, evaluation, and fallback logic.
- Improved answer accuracy by 10% and reduced LLM inference cost by 50% through retrieval and context optimization, objective evaluation, acceptance thresholds, caching, fallback handling, and cost-aware orchestration.
- Developed Python-based AI services and containerized inference workflows using FastAPI, Pydantic, Docker, AWS Lambda, S3, SageMaker, and ECR, covering data preprocessing, structured inputs and outputs, inference, validation, logging, and failure handling.
- Built reusable data-processing and model-validation pipelines for multichannel time-series, simulation, sensor, and computer-vision data, supporting automated quality checks, anomaly detection, digital-twin validation, and engineering decision-making.
- Developed and evaluated machine-learning and deep-learning solutions using PyTorch, TensorFlow/Keras, scikit-learn, and XGBoost, including time-series, computer-vision, and graph-based models.
- Built reusable components for AI and ML workflows, including retrieval, evaluation, data preprocessing, validation, and structured inference pipelines, to support repeatable experimentation and solution delivery.

Technical Skills

GenAI & Agentic AI

RAG, Case-Based Reasoning, agentic workflows, LLM orchestration, embeddings, retrieval, prompt engineering, context engineering, structured outputs

AI Evaluation & Reliability

Objective evaluation, acceptance thresholds, validation gates, fallback/retry logic, failure analysis, caching, token and inference-cost optimization

Machine Learning & Data

Python, SQL, PyTorch, TensorFlow/Keras, scikit-learn, XGBoost, Pandas, NumPy, time-series analysis, computer vision, GNNs

AI Engineering & Data Pipelines

FastAPI, Pydantic, REST APIs, ETL/data pipelines, Parquet, data validation, reusable processing and inference components

Cloud & ML Engineering

Docker, AWS Lambda, S3, SageMaker, ECR, GCP Vertex AI, BigQuery, Git, CI/CD, containerized inference, logging, failure handling, deployment workflows

Education

04/2020–03/2023

M.Sc. Mechatronics and Robotics, *Technical University of Munich*, Munich, Germany
Master's thesis on Deep Learning–FEM integration for material characterization.

10/2014–01/2020

B.Sc. Mechanical Engineering, *Leibniz University Hannover*, Hanover, Germany
Bachelor's thesis on FEM analysis of stress distribution in dry-lubricated roller bearings.

Selected Publication

2025

Yuan, Z. et al., **Automated Machine Learning for Monitoring Physical Twins in Passive Vehicle Safety**, DATA 2025.

Languages

German: C1 English: C1 Mandarin/Cantonese: Native