Sandro Sage

AI Engineer - Agentic AI & Generative AI Systems

+49 157 55992600 | sandro-sage@gmx.de
Munich; Germany
LinkedIn: sandro-sage

Summary

AI/ML engineer with 5+ years of Python-based software development across generative AI, computer vision, Natural Language Processing (NLP) and embedded ML. Currently building production multi-agent GenAI systems (LangChain, LangGraph, RAG, FastAPI) for a large-scale BMW initiative by bringing hands-on experience with BMW's AI ecosystem and Cloud Infrastructure. M.Sc. in Artificial Intelligence (FAU, grade 1.3).

Experience

ARRK Engineering GmbH — Munich, Germany
AI Engineer — *December 2025 - present*

- Design and develop a production multi-agent Generative AI system (one orchestrator agent plus two specialized subagents) serving 10+ active users, using Python with LangChain, LangGraph, and LangFuse for orchestration, tracing, and evaluation
- Built a RAG pipeline integrating Weaviate vector search and PostgreSQL, and developed scalable FastAPI backend services containerized via Docker and Docker Compose
- Adjusted Kubernetes/Helm deployment templates and GitHub Actions CI/CD pipelines to support a multi-service architecture (FastAPI backend and Celery workers) requiring shared persistent storage, after the default template proved insufficient for a single-service setup
- Core contributor on a 5-person team delivering a large-scale BMW Generative AI initiative; step in on internal project coordination when the project lead is unavailable
- Participate in twice-weekly meetings with BMW stakeholders to review new features, tickets, and project status, and run guided demos of the agent system for end users
- Apply MLOps/AgentOps/DevOps practices, clean code principles, and automated testing to keep multi-agent systems reliable in production, and serve as a go-to resource for the AI/ML team within ARRK's ADAS department on agent design and LLM integration
- Dedicate a portion of time to exploratory research on world models and latent space manifolds, extending academic research into applied industrial contexts for driving scenarios

Skills: *Python, LangChain, LangGraph, LangFuse, FastAPI, Docker, Kubernetes, Helm, Celery, GitHub Actions, Weaviate, PostgreSQL, RAG, Multi-Agent Systems, MLOps/AgentOps/DevOps, World Models*

Fraunhofer IIS — Tennenlohe, Bavaria
Working Student - AI & Machine Learning — *October 2023 - November 2025*

- Built and optimized computer vision pipelines across two applied research projects (SyNaKI, GAIA) using TensorFlow/TFLite, model compression and quantization via the TensorFlow Model Optimization Toolkit, and custom OpenCV image-processing algorithms, deploying models on embedded hardware
- Implemented a distributed AI tool using evolutionary algorithms, and supported funding-proposal preparation for AI research initiatives within the institute
- Recognized as a dependable, proactive contributor across both research and applied engineering work within the AI team

Skills: *Python, C++, TensorFlow, TFLite, TensorFlow Model Optimization Toolkit, OpenCV, Computer Vision, Embedded Deployment, Evolutionary Algorithms*

DIT Lab for Digitalization of AI in Electrical Engineering — Deggendorf, Bavaria
Research Assistant — *March 2022 - March 2023*

- Led sensor-based data acquisition on DC-DC converters to enable predictive maintenance through behavioral monitoring
- Engineered meta-statistical features over sliding time windows and applied classical ML models (scikit-learn, GridSearch-tuned) to detect anomalies and behavioral patterns in converter sensor data
- Built and deployed CNNs, compressed and optimized via the TensorFlow Model Optimization Toolkit, on resource-constrained TinyML devices
- Completed Bachelor's thesis on TinyML model generation, optimization, and evaluation as part of this research

Skills: *Python, scikit-learn, GridSearch, TensorFlow, TensorFlow Model Optimization Toolkit, TinyML, Time Series Feature Engineering, CNNs, Sensor Data Acquisition*

T-Systems Multimedia Solutions GmbH Dresden, Saxony
Internship - IoT Edge Computing & Machine Learning (AWS) *October 2021 - February 2022*

- Independently designed and deployed end-to-end cloud data pipelines using the AWS Cloud Development Kit (CDK), including database management and visualization dashboards
- Supported predictive maintenance projects covering automated material handling systems and ultrapure water valve condition monitoring, applying SCRUM and CI/CD practices in an agile delivery environment

Skills: *AWS CDK, Cloud Architecture, Data Pipelines, SCRUM, CI/CD*

DIT Project Lab for Hardware-Related Digitization Deggendorf, Bavaria
Research Assistant - Embedded Systems *October 2020 - July 2021*

- Developed a GUI application for sensor communication via UART, SPI, and I2C protocols using the PICkit Serial Analyzer, bridging hardware and software integration across embedded systems projects

Skills: *C, Embedded Systems, UART, SPI, I2C, GUI Development*

Education

Friedrich-Alexander-University (FAU) Erlangen, Bavaria
M.Sc. Artificial Intelligence *April 2023 - March 2026*

- German grade: 1.3 | Focus: Advanced Deep Learning, Pattern Recognition, Statistical Learning, Diffusion and Attention Models
- Master Thesis: K-space Latent Diffusion for Accelerated MRI (IdeaLab) - Grade: 1.3

Deggendorf Institute of Technology (DIT) Deggendorf, Bavaria
B.Eng. Applied Computer Science *October 2019 - March 2023*

- German grade: 1.4 | Focus: Embedded Systems
- Bachelor Thesis: Generation, Optimization, and Evaluation of ML/DL Models for Tiny Machine Learning (TinyML) - Grade: 1.3

University of Southern Denmark (SDU) Odense, Denmark
Summer School - Deep Learning *August 2022 – August 2022*

- Completed a hands-on project classifying healthy vs. pneumonia-affected lungs using convolutional neural networks

Research Projects

K-space Latent Diffusion for Accelerated MRI - Master's Thesis, IdeaLab (FAU)
GitHub: github.com/sandrosage/kspace-diffusion

- Investigated applying latent diffusion models directly in k-space for accelerated MRI reconstruction - a direction not previously explored systematically - through the proposed kLD-MRI framework
- Designed and trained k-space autoencoders (K-AEs) to learn compact latent representations of fully-sampled k-space, demonstrating that high-frequency k-space structure can be preserved in the latent domain
- Built a latent diffusion model and a consistency-guidance sampling method to reconstruct undersampled k-space; identified specific limitations in diffusion stability and guidance-manifold interaction that inform directions for future frequency-domain reconstruction research
- Open-sourced the full implementation for reproducible k-space latent dataset generation and follow-on research

SMMetrDE - Segmentation-guided Monocular Metric Depth Estimation
GitHub: github.com/sandrosage/Monocular_Depth_Estimation

- Designed a framework combining semantic segmentation with monocular depth estimation, integrating and evaluating multiple state-of-the-art depth models (MonoDepth2, MiDaS, DPT, ZoeDepth) for object-wise metric prediction
- Conducted model performance evaluations (RMSE, AbsRel, inference time) across KITTI dataset subsets, gaining practical experience with transformer-based vision models (ViT, BEiT, Swin)

Technical Skills

AI/ML Frameworks: LangChain, LangGraph, LangFuse, PyTorch, PyTorch Lightning, TensorFlow, Keras, Hugging Face, scikit-learn, OpenCV, World Models, Latent Space Modeling

Agent, LLM & NLP: Multi-agent systems, AgentOps/MLOps, RAG pipelines, prompt & context engineering, Weaviate (vector DB)

Software Engineering: Python, C, C++, SQL, clean code & object-oriented programming (OOP), Git, GitHub, CI/CD

Infrastructure & DevOps: FastAPI, Docker, Docker Compose, Kubernetes, Helm, platform development, Celery, GitHub Actions, AWS (CDK), PostgreSQL, Jira

Languages: German (Native), English (C1), French (B1)