Namit Bhalendra Kharade

Berlin, Germany

namit.b.kharade@gmail.com | +49 176 37249426 | namitkharade.github.io | linkedin.com/in/namit-k

Professional Summary

AI & Software Engineer (M.Sc.) with 3+ years of experience specializing in backend infrastructure for Generative AI. Proven track record of taking LLM applications from prototype to production, with specific expertise in GraphRAG systems, inference optimization, and automated evaluation pipelines. Strong background in core software engineering and ETL processes ensuring solutions are scalable, testable, and cost-efficient.

Experience

AIME GmbH Berlin, Germany
AI Engineer October 2024 - Present

- Deployed scalable **AI inference endpoints** on AIME's **GPU Cloud**, configuring the **AIME API Server** to manage high-concurrency model requests across distributed workers.
- Built a **GraphRAG system** using **LangChain** and **Knowledge Graphs** to support complex multi-hop queries, and developed an evaluation pipeline with **Ragas** and **G-Eval** to ensure high response quality.
- Developed a rigorous evaluation pipeline using **Ragas** and **LLM-as-a-Judge(G-Eval)**, benchmarking **GraphRAG vs. Standard RAG vs. In-Context LLMs** to scientifically determine the optimal retrieval strategy for enterprise datasets.
- Containerized Generative AI models (Stable Diffusion, Whisper) into **AIME MLC (Machine Learning Containers)**, optimizing them for low-latency real-time inference in production.
- Documented **performance metrics** and research findings to build a comprehensive internal **knowledge base**, streamlining R&D workflows and enhancing team-wide knowledge sharing.

ai-coustics GmbH Berlin, Germany
Data Engineer February 2024 – August 2024

- Orchestrated scalable **ETL pipelines** using **Apache Airflow** on Google Cloud Platform (GCP), automating the ingestion of training data into **BigQuery**.
- Developed **Python automation scripts** to handle raw audio sorting and archiving, reducing manual data processing time by **50%**.
- Integrated audio analysis tools (**Yamnet** and **DNSMOS**) directly into the pipeline to automatically tag content and score audio quality, eliminating the need for manual labeling and filtering.
- Implemented automated quality checks to catch corrupted files and silence early, ensuring that only high-quality data made it into the training sets.

aiio GmbH Magdeburg, Germany
Software Engineer March 2023 - January 2024

- Diagnosed and resolved bugs in a **React and Django** based web application, enhancing functionality and improving the overall user experience.
- Integrated **Large Language Models (LLMs)** such as OpenAI's GPT-3.5 and GPT-4 models, applying **AI** to improve user productivity and ease of use.
- Enforced high code quality standards by developing comprehensive **unit and system tests**, ensuring reliable deployment.
- Supported the development, configuration, and documentation of the Project Management application, contributing to increased project efficiency and collaboration.

Accenture Solutions Pune, India
Associate Software Engineer August 2021 - August 2022

- Collaborated with **cross-functional teams** to create and train **machine learning programs** that made customer service better for British Telecom.
- Analysed and prepared data using **Python and SQL** for data analysis and preparation which we then used to train our **AI models**.

- Implemented **CI/CD automation** via **Jenkins**, reducing deployment cycle times from **2 days** to **4 hours** and ensuring build stability.
- Actively participated in **code reviews**, finding ways to make it better, and following the best practices to maintain code.

AI Adventures Pune, India
ML Engineer Intern December 2020 - May 2021

- Developed a conversational AI agents using **DialogFlow** (pre-LLM) to provide real-time COVID-19 updates, handling natural language user queries.
- Performed **data augmentation** and cleaning to expand training datasets, improving the robustness of downstream AI models.

Education

Otto-von-Guericke University Magdeburg, Germany
Master of Science in Data and Knowledge Engineering October 2022 - Dec 2025

- **Master's Thesis:** Deployment and Evaluation of a Scalable GraphRAG System for MHQA using Diverse Open-Source LLMs on a CaaS Infrastructure.
- **Focus Areas:** Deep Learning, Cloud Computing, Natural Language Processing, Big Data Storage.

Savitribai Phule Pune University Pune, India
Bachelor of Engineering in Computer Engineering May 2017 - July 2021

- **Key Coursework:** Data Structures & Algorithms, Database Management Systems, Software Engineering.
- **Final Grade:** First Class with Distinction

Skills

Generative AI & LLM Engineering:
LLM Inference, Fine-tuning (LoRA/QLoRA), vLLM, SGLang, RAG & GraphRAG, LangChain, LlamaIndex, Knowledge Graphs (Neo4j, RDF), Prompt Engineering, Evaluation Frameworks (Ragas, G-Eval), Transformers.

Machine Learning & Data Science:
Deep Learning, Neural Networks, PyTorch, TensorFlow, Keras, Recommenders, Computer Vision (OpenCV), NLP, Speech-to-Text (Whisper), Audio Analysis (Yamnet, DNSMOS), Explainable AI (XAI), Scikit-learn, Pandas, NumPy.

MLOps, Cloud & Infrastructure:
Google Cloud Platform (GCP), AWS (EC2, S3), Docker, Kubernetes, AIME GPU Cloud, CI/CD (Jenkins, Git/GitHub Actions), API Development (FastAPI, Sanic), Linux/Unix Environments.

Data Engineering & Databases:
Apache Airflow (ETL), Web Crawling, BigQuery, Vector Databases (LanceDB, PostgreSQL (pgvector)), SQL (PostgreSQL), NoSQL (MongoDB), Data Visualization (Matplotlib, Seaborn, Plotly, PowerBI).

Languages & Web:
Python, SQL, JavaScript, TypeScript, Bash/Shell, HTML/CSS, React, Vue.js, Django. **Tools:** Selenium, Playwright, Pytest, Jira, Git, Agile/Scrum.

Languages (Spoken):
English (C1 - Full Professional Proficiency), German (A2 - Elementary).

Publications

Deep-learning based Helmet Violation Detection System [IEEE Xplore]
2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)

- Engineered a real-time object detection pipeline using **YOLOv4** to identify helmet violations, optimizing inference speed and accuracy to outperform standard CNN baselines on surveillance feeds.

COVID-19 Disease Prediction & Real-Time Mask Detection [IEEE Xplore]
2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)

- Implemented a hybrid Deep Learning system combining Computer Vision for real-time mask detection and predictive modeling for disease spread, optimized for low-latency edge deployment.