Shivang Sinha

GenAI / ML Engineer

Koblenz, Germany | +4915560169619 | sinhashivang35@gmail.com
linkedin.com/in/shivang-sinha-92755012b | www.shivangsinha.website

Valid German residence permit. Available for full-time roles from March 2026.

Summary

GenAI / AI Solutions Engineer with 5+ years of experience delivering LLM- and RAG-powered systems in enterprise and industrial environments (Carl Zeiss, Barclays, Hexagon).Designs and deploys agent-based copilots and developer assistants that streamline engineering workflows and cut time spent on documentation and support tasks.Currently completing an M.Sc. in Natural Language Processing (graduating Feb 2026) with a thesis benchmarking BERT, rule-based and LLM-based models for skill extraction on 1,000+ German AI/ML job postings; targeting GenAI / ML Engineer roles in the EU.

Technical Skills

LLMs & NLP: Large Language Models (LLMs), Generative AI, Retrieval-Augmented Generation (RAG), Prompt Engineering, BERT fine-tuning, Text classification, Sentiment analysis, HuggingFace, spaCy, LangChain, LangGraph, Phoenix.

Programming & Backend: Python (FastAPI, Flask, Django), .NET (C#), REST APIs, Microservices, React, TypeScript.

Cloud & Data: Microsoft Azure (Azure OpenAI), Docker, CI/CD, SQL, MySQL, MongoDB, Vector databases (FAISS, Chroma).

Work Experience

Carl Zeiss Microscopy — Munich, Germany
Werkstudent GenAI — Nov 2024 – Present

- Contribute to an internal Microscopy Copilot built on Azure OpenAI and LangGraph, implementing agent-based workflows that automate recurring engineering tasks across R&D teams.
- Engineered an AI-driven developer assistant that converts natural-language requests into executable Python scripts, surfaces relevant internal documentation, and generates C# helper functions to accelerate feature development.
- Optimized document retrieval pipelines by refining indexing and query strategies, improving context relevance and reducing time engineers spend searching internal documentation.
- Built RAG pipelines over documentation and source-code repositories to enable structured, context-aware code generation and troubleshooting guidance.
- Orchestrated integrations of MCP-based services that streamline cross-team workflows and support the rollout of AI solutions from proof-of-concept to production usage by multiple software engineering teams.

Barclays — Pune, India
Senior Software Developer — Mar 2022 – Mar 2024

- Developed and enhanced enterprise-scale full-stack applications supporting investment banking operations for internal business stakeholders.
- Designed and tuned RESTful APIs and backend services for high-volume transactional workflows with emphasis on performance, reliability, and secure data handling.
- Helped implement an internal chatbot for operations teams that centralized common queries and reduced manual ticket handling effort.
- Collaborated within Agile Scrum teams (6–8 engineers), participating in code reviews, sprint planning, and production releases in a regulated environment.

Hexagon CCI Hyderabad, India
Software Developer Sep 2020 – Mar 2022

- Built and maintained customer-facing industrial web applications used by global manufacturing and metrology clients.
- Implemented backend services and REST APIs that enabled new product capabilities and integrations across distributed systems.
- Improved application stability and responsiveness in production by addressing performance bottlenecks and coordinating fixes with cross-functional teams.

Education

University of Trier Germany
M.Sc. Natural Language Processing (Data Science) Apr 2024 – Present

- Master's Thesis: *A Comparative Study of NLP Models for Longitudinal Analysis of AI/ML Skill Requirements in the German Job Market* (Expected Feb 2026).
- Collected and annotated 1,000+ German job postings and benchmarked BERT-based, rule-based and LLM-based skill extraction models using Precision, Recall and F1-score for longitudinal trend analysis.

Chandigarh University India
B.E. Computer Science (Cloud Computing specialization by IBM) 2016 – 2020

- Completed projects and coursework in cloud computing, distributed systems and software engineering.

Projects

NLP Skill Extraction Benchmarking (Master's Thesis)

- Built a labeled dataset from 1,000+ German AI/ML job postings to evaluate skill extraction approaches for longitudinal analysis of market requirements.
- Evaluated BERT-based, rule-based and LLM-based NLP pipelines using Precision, Recall and F1-score, focusing on accuracy and robustness of extracted AI/ML skills.

Agent-Based Research Assistant

- Developed an LLM-powered research assistant (FastAPI + React) that ingests sell-side reports and returns cross-asset investment summaries with transparent reasoning steps.
- Integrated semantic search, tool-calling and streaming chat to deliver accurate, explainable answers and traceable recommendation flows for end-users.

Languages & Additional

- **Languages:** English (Fluent, C1), German (Intermediate, B1/B2, actively learning).
- **Work Authorization:** Valid residence permit in Germany; open to EU relocation.
- **Availability:** Full-time from March 2026.