GURU DEEP SINGH

Von-Hünefeld Strasse 24, Neu-Ulm, Bavaria 89231, Germany | +491704661651 | gurudeep1998@gmail.com |
linkedin.com/in/guru-deep-singh | www.gurudeepsingh.com
Nationality: Indian (**Niederlassungserlaubnis**)

PROFILE

Results-driven Software Engineer with proven experience in developing and deploying robust GenAI, Machine Learning and traditional Computer Vision software solutions both for edge hardware and cloud platforms. Displaying expertise in ADAS, multi-modal fusion and embedded deployment, passionate about innovation and committed to continuous learning, effective problem-solving, and delivering high-quality, scalable solutions in agile environments.

WORK EXPERIENCE

Software Engineer, *Aumovio (Previously Continental AG)*, Oct. 2022 – Present Ulm

- Developed a **CNN** framework for traffic object detection using **ONNX** in **C++** and Python for **LiDAR** data, contributing to a **mid-level fusion** schema and achieving over 85% precision in urban traffic scenarios, thereby enhancing real-time detection accuracy and contributing to improved road safety.
- Designed and implemented a **LiDAR Odometry** algorithm compatible with sensors including HRL131, Innovusion, P64, and P128.
- **Filed a Patent** on **a Long-range Camera and LiDAR-based height optimization** strategy for **out-of-distribution** objects thereby, improving the accuracy of height estimation by 20% compared to a LiDAR-only solution.
- Engineered a unique KPI framework to compare **multi-modal Odometry** signals without the need for explicit extrinsic **calibration** between sensors.
- Implemented **CI/CD pipelines** using **Jenkins and GitHub Actions** for automated testing and deployment in production environments, ensuring SiL-HiL consistency. Reduced the **release cycle from 2 sprints to 1 sprint**.

Working Student (Master Thesis), *Aumovio (Previously Continental AG)*, Nov. 2021 - July 2022 Neu-Ulm

- Conducted research and development of a **domain mapping-based transfer learning** approach for point cloud adaptation between optomechanical and microelectromechanical systems (MEMS) LiDARs, enabling downstream tasks such as **semantic segmentation and object detection**.
- Used open-source datasets like **Cirrus and KITTI** for **pre-training** and inferring pseudo-labels for proprietary internal HRL131 LiDAR data.

PERSONAL PROJECTS

AI-Guided Robot Mobilization with ROS 2 and VLM

- Simulated a custom 4-wheeled in **ROS2** with front-facing camera fed to a **Vision-Language Model** every **~1.5s**, a cadence determined empirically to **prevent frame drops** due to inference delays.
- Designed the VLM **reasoning pipeline** around a structured spatial grid with explicit lane and obstacle identification, **returning an intent in structured output** rather than raw coordinates, eliminating jitter and leveraging the model's semantic strengths.
- Created a mapping between the intent from VLM to **low level controls** for the Robot.
- Integrated **Langfuse** observability to trace VLM’s decision making process.

Transformer Architecture Implementation from Scratch

- Developed a complete Transformer architecture from scratch in **PyTorch** and converted it to a **TensorRT-optimized engine** for accelerated inference.
- Benchmarked performance across multiple **NVIDIA GPUs (T4, L4, A100), edge hardware (Jetson Orin Nano Super)** and **AMD MI300X**, achieving **1.8x to 2.8x** inference speedup compared to standard PyTorch.
- Addressed numerical stability problems using **FP16 precision** - demonstrating deep understanding of both model internals and hardware constraints.

Llama-2 7B Implementation from Scratch

- Developed a complete **Llama-2 architecture** from scratch in PyTorch and executed inference on Jetson Orin Nano Super and T4 GPU**.**
- Implemented **Rotary Positional Embeddings** (RoPE), **Grouped Query Attention** (GQA) and **KV Cache**.

Serverless AI-Enabled Resume Chatbot on AWS

- Developed a fully serverless digital resume hosted on AWS **S3** and deployed globally via **CloudFront**.
- Designed and implemented the backend using AWS **Lambda** instead of EC2 and **API Gateway** to provide interactive and scalable functionality. Reducing the cost from an estimated 8 Euro/month to 0.1 Euro/month.
- Used **AWS CDK** to deploy infrastructure as a code (IaC).
- Implemented a **Retrieval-Augmented Generation (RAG**) workflow with vector embeddings in a database stored in S3 for cost efficiency and LLM integration for AI-powered resume insights.
- Containerized Python runtime with Docker stored on **ECR** for Lambda to support custom dependencies.
- Deployed the chatbot on personal website with frontend made with **Node.js.**

Fine-tuned Personalized LLM (Digital Guru)

- **Fine-tuned Llama-3.2-3B-Instruct** on over 100,000 anonymized personal chat messages from WhatsApp and Instagram to develop an AI assistant that replicates individual writing style and communication patterns with integrated guardrails.
- Devised an **ETL pipeline** for the consumption of WhatsApp and Instagram datasets.
- Implemented **QLoRA (4-bit quantization + LoRA adapters)** for memory-efficient training on personal laptop.
- Deployed the model under the brand of "Digital Guru" on **HuggingFace Spaces**.

SKILLS

Languages: Python, C++, CUDA

ML/AI: PyTorch, TensorRT, ONNX, OpenCV, ROS/ROS2

Cloud: AWS CDK, Terraform, S3, Lambda, API Gateway, ECS, EC2, CloudFront, AWS Bedrock, Azure (Basic)

Collaboration: GitHub, GitLab, Jira, Confluence, SCRUM, Agile Development

CI/CD & Build Systems: Jenkins, Docker, Kubernetes, CMake

Agents SDK: OpenAI Agents SDK, LangGraph, LangChain, MCP, CrewAI

Databases: PostgreSQL, MongoDB

Testing: pyTest, GoogleTest

EDUCATION

Master of Science, Robotics, *Delft University of Technology*, Sept. 2020 - Sept. 2022

GPA: 8.22 (~ 1.8 German Scale)

- Awardee of **Erasmus Scholarship** – Aug 2021: Competitive scholarship for internships across Europe.
- Awardee of **Holland Scholarship** – Sept 2020: Merit-based scholarship for top-performing international students (limited to 3-4 students per year).
- Thesis: Transfer Learning for 3D Point Cloud Using Domain Mapping for Motorized Optomechanical And Microelectromechanical Systems LiDARs

Bachelor of Technology, Mechanical and Automotive Engineering, *Delhi Technological University*, June 2016 - June 2020

GPA: 9.09 (~ 1.45 German Scale)

LANGUAGES

- **German: B1** (Exam May 2026)
- **English: C1**

CERTIFICATES

- **AWS Certified Cloud Practitioner, Amazon Web Services**
- **AI Builder with n8n, Udemy**
- **LLM Engineering, Udemy**
- **Agentic AI, Udemy**