HARSH GURAWALIYA

harshgurawaliya3@gmail.com | Born 24.03.2002| +49 17655738830

Silvanster 4 , Munich, 81927 LINKEDIN GITHUB_GITHUB-work

EDUCATION

Deggendorf Institute of Technology 10/2021– present

Bsc Artificial Intelligence Deggendorf, Germany

Academic Focus: Generative AI , NLP , Computer Vision , Autonomous Robotics , Deep Learning , Machine learning

Expected Graduation: September 2025

PROFESSIONAL EXPERIENCE

Onepager Software Gmbh 7/2024 - present

AI Backend Engineer Munich, Germany

CUDA | REST API | RAG ARCHITECTURE | LOCAL LLM | DOCKER | AWS | OPENVPN

- Engineered production-grade RAG application from scratch using FastAPI and open-source LLMs via Ollama, implementing custom document chunking and ChromaDB for efficient vector storage and retrieval
- Developed intelligent retrieval system **achieving 30% improvement** in document generation accuracy compared to standard ChatGPT with raw company documents, while ensuring data privacy through locally deployed models
- Architected comprehensive backend infrastructure with PostgreSQL for user management and query tracking, implementing AWS S3 for secure document storage with versioning
- Built and led development of secure RESTful APIs with JWT-based authentication, implementing core endpoints for authentication flows, document upload/processing, history tracking, and document retrieval with role-based access control
- Deployed local LLM infrastructure accessible via OpenVPN, ensuring data privacy and eliminating external API dependencies while maintaining company security requirements

B Plus Automotive Gmbh 2/2024 – 6/2024

Working Student Deggendorf, Germany

DEEP LEARNING | CNN | AUTONOMOUS VEHICLE |PYTORCH

- For a university project, I developed an advanced post-processing method to enhance deep neural network predictions, specifically **addressing label error detection in connected components within semantic segmentation** tasks.

- My contribution was to write the Python Script to implement object detection algorithm YOLOv8 using pytorch.
- Previous achievements: Grade 1 in Computer Vision university course and Kaggle certificates.

PERSONAL PROJECTS

CUSTOM FINE-TUNING OF LLAMA 3 FOR MEDICAL DOMAIN CONVERSATIONS JULY 2024

. QLoRA | TRANSFORMERS| LLAMA.CPP | GGUF FORMAT | QUANTIZATION | LM STUDIO

- Fine-tuned Llama 3 on an extensive patient-doctor conversation dataset to tailor the model for medical inquiry and advice.
- Merged the fine-tuned adapter with the base model, then converted and quantized it into Llama.cpp's GGUF format for reduced resource usage and efficient inference.
- Successfully integrated the optimized model into the LM Studio application, enabling secure, domain-specific, and locally hosted conversational AI.

FULL STACK AI OPEN SOURCE INTELLIGENCE SAAS FOR PERSONALISED MARKETING MAY 2024

. FLASK. | REACT | TAILWIND CSS | MONGO DB | NLP | WEB SCRAPPING

- Back-end development uses Python and Flask to handle data processing and machine learning tasks.
- Front-end development uses JavaScript, React, HTML/CSS, and Tailwind CSS for an interactive user interface.
- Implements Shadecn/UI as a Tailwind CSS framework.
- Implemented Google authentication, Social Searcher API for multi-source data collection, and Perplexity AI for data analysis and insight extraction.
- Stores and manages data using MongoDB database

END TO END FULLY AUTOMATED MACHINE LEARNING MODEL FULLSTACK March 2024

DATA ANALYSIS | REGRESSION | HYPER PARAMETER TUNING| CI/CD PIPELINES | DOCKER | AWS

- Utilized **NumPy, Pandas** for data manipulation, **Matplotlib, Seaborn** for visualization; led exploratory data analysis.
- Developed **automated** preprocessing pipeline: **feature categorization, one-hot encoding, feature scaling**.
- Employed **Scikit-learn** for regression model analysis; minimized RMSE, MAE, maximized R2 Score. Utilized GridSearchCV for hyperparameter tuning. Used **Flask web app**
- Implemented **CI/CD pipeline via GitHub Workflows, Docker for testing, deployment, containerization; deployed on Amazon AWS for scalability.**

FACIAL RECOGNITION-BASED COFFEE DISPENSING SYSTEM FOR UNIVERSITY CAMPUS Nov 2023

PYTHON | YOLOv5 | TKINTER | PYTORCH | TENSORFLOW

- **Facial Recognition via CNNs Trained with PyTorch and TensorFlow**: Implemented a facial recognition system using convolutional neural networks (CNNs) trained with PyTorch and TensorFlow, integrated within YOLOv5 and operated through a Python-based Tkinter GUI

Distributed Multiplayer Chess Platform: A Client-Server Implementation with Python and Pygame

PYTHON | PYGAME | SOCKET MODULE June 2023

- **Development of Real-Time Multiplayer Chess Game**: Designed and implemented using Python, integrating Pygame for game mechanics and leveraging sockets for real-time multiplayer gameplay across different systems.
- **Server-Client Architecture for Scalable Communications**: Managed robust server-client communications essential for handling multiple simultaneous games, with a focus on IoT integration and system scalability.

EXTRA CURRICULAR

Member of United AI club , Deggendorf JAN – DEC 2022

REFERENCES

Paolo Rechia

Lead Software Engineer Schwarz IT KG

www.paolorechia.com

Prof. Dr. Patrick Glauner

• Lead by CDO Magazine in the list of the world's leading professors in data

• Expert of the German Bundestag and the French National Assembly on AI

Harsh

Date : 9.12.2024

Place: Munich