BORRA VISHNU SUHAAS

AI Engineer

mehersuhaas123@gmail.com Hyderabad, India
linkedin.com/in/suhaasborra github.com/Suhaas_code

SUMMARY

AI Engineer with 2+ years of industry experience and a core team member on GenX, a multitenant enterprise AI platform. Builds Python microservices for agent runtimes, hybrid RAG, multi provider model routing, data ingestion and memory. Specializes in LLM finetuning with PEFT, adapters, RLHF and hyperparameter optimization, and contributes to open source tooling for synthetic data generation.

EXPERIENCE

AI Engineer 02/2025 to Present
Nainovate Technologies Hyderabad, India

- Core team member owning LLM finetuning capabilities for open source models, covering PEFT, adapter architectures, reinforcement learning (RLHF) and hyperparameter optimization
- Built synthetic dataset generation pipelines used for model training and retrieval context
- Developed production grade Python/FastAPI services for training, inference and evaluation, and contributed React frontend components

AI Developer Intern 08/2024 to 02/2025
Brillius Technologies Pvt Ltd Hyderabad, India

- Finetuned and deployed LLMs with Hugging Face Transformers and PEFT/LoRA for enterprise use cases, and built a custom hyperparameter optimization algorithm for finetuning
- Built NLP pipelines for text classification, summarization and question answering, including dataset curation and benchmark based evaluation
- Integrated optimized models into backend systems and improved inference performance and reliability

Machine Learning Intern, Radar and Seismology 05/2023 to 05/2024
India Meteorological Department, Govt of India Jaipur, India

- Built radar analysis models using neural networks, Transformers and denoising algorithms
- Trained multiple models in parallel and combined them through ensemble learning

PROJECTS

GenX: Enterprise AI Platform Nainovate
Python, gRPC, MongoDB, Qdrant, Pinecone, LiteLLM, Docker, Kubernetes 2025 to Present

- Built backend services for a multitenant AI suite of 12 Python microservices (gRPC and HTTP) with per tenant MongoDB isolation
- Developed an agent runtime with RAG, multi agent and tool calling modes, routing across OpenAI, Anthropic, Gemini and local TGI models via LiteLLM with automatic fallback
- Built hybrid retrieval (dense plus BM25, RRF fusion, reranking), document ingestion and MCP enabled connectors for Jira, SQL, S3 and SharePoint

HDAI: Hand Detection As Input Pravaig
Computer Vision, Real Time Gesture Recognition

- Built software that turns any standard camera into a touchless input device, detecting and interpreting hand gestures to control desktop applications in real time with no extra hardware
- Pitched to investors and pilot partners with a roadmap from desktop control to a developer SDK and enterprise deployment, targeting presenters, educators and accessibility users

DocuMint AI: Legal Document Intelligence GitHub

Python, FastAPI, Google Cloud Vision, Document AI, Docker

- Built a FastAPI backend that runs a six stage pipeline: PDF to images, OCR with Google Cloud Vision, Document AI parsing, legal document classification and knowledge graph input generation
- Added background processing, live pipeline status endpoints, Docker deployment and a pytest suite

Open Source: Meta Synthetic Data Kit GitHub

Python, LLM training data generation

- Contributed to Meta's synthetic data kit (1.6k stars), a tool for generating high quality synthetic datasets for LLM finetuning

EDUCATION

Bachelor of Technology in Computer Science, Hons in Artificial Intelligence & ML 2021 to 2025

Manipal University Jaipur

- Student Excellence Award, Distinguished member of the branch

SKILLS

Python, FastAPI, gRPC, REST APIs, SSE

Agentic AI, MCP, tool calling, RAG, hybrid search, reranking, memory management

Computer vision, OCR, document AI, React, Node.js

LLMs, Hugging Face Transformers, finetuning, PEFT (LoRA, adapters), RLHF, HPO

MongoDB, Qdrant, Pinecone, LiteLLM, Docker, Kubernetes, Google Cloud, Git

Fluent in English, Telugu and Hindi, Basic in French and German

CERTIFICATIONS

- NPTEL Introduction to Machine Learning (Elite, Top 5%)
- NPTEL Computer Vision and Image Processing (Elite)
- NPTEL Introduction to Internet of Things (Elite, Silver)
- Coincent Certified Cyber Security Trainee and Ethical Hacker

LEADERSHIP

Deputy Head, Curations 06/2022 to 05/2023

MUJ ACM Student Chapter Jaipur, India

Campus Ambassador 11/2021 to 11/2023

Cisco Jaipur, India