SHREYAS S K

GENERATIVE AI SPECIALIST

skshreyas714@gmail.com | +917019667673 | LinkedIn | Substack | GitHub | Youtube | Portfolio

SUMMARY

6+ years of broad experience designing and delivering production-grade LLM systems, RAG pipelines, and agentic workflows across enterprise environments. Strong background in NLP, Computer Vision, Reinforcement Learning and ML infrastructure. Proven track record leading cross-functional initiatives, owning end-to-end delivery from architecture and implementation to documentation

CORE SKILLS

Programming : Python, PyTorch, Numpy, Pandas, Langchain, Langgraph
Technologies : Machine learning, NLP, Computer Vision, Reinforcement Learning
Cloud Expertise : AWS, GCP : ECS, EC2, S3, Sagemaker, Cloud build, Cloud Run
Generative AI : GPT-4, Gemini-Pro, Mistral 7B, Llama 7B, DPO, LoRA, DoRA
MLOps/LLMOps : Dagster, Argilla, GitHub CI, Docker, Kubernetes, WandB, DeepSpeed, vLLM, accelerate, Nvidia Triton, TensorRT

WORK EXPERIENCE

Ford Motors Pvt Limited, Bangalore - Generative AI Specialist

AI DevOps Agent - Product Owner/ Tech Anchor Jul 2024 - Present

- Led a team of 5 to build an AI-powered DevSecOps automation platform integrating Terraform, CI/CD and security/compliance (FOSSA, 42Crunch, SonarQube, Cycode), onboarding 150+ internal customers.
- Designed a supervisor agentic workflow enabling natural-language to complete deployment and compliance via MCP, Browser agent, instrumented tracing/monitoring with Arize Phoenix.
- Achieved 30–60% reduction in engineering effort by automating pipelines and applying Ford-specific compliance guardrails.
- Drove product GTM internally: docs, forum demos, onboarding sessions

Owner's Manual Chatbot - Individual Contributor

- Evaluated Nvidia Aegis, Llama Guard, Prompt Guard, and commercial LLMs for jailbreak/prompt-injection resilience; integrated best-performing stack.
- Fine-tuned an SLM (prompt-guard-86M) on adversarial datasets tailored to domain threats; improved safety metrics by ~10% vs base.
- Established red-teaming and safety benchmarks, integrated runtime policy checks

Khoros India R&D Pvt Ltd, Bangalore - Machine Learning Engineer

Helix (RAG-powered Khoros Bot) - Individual Contributor Jan 2022 - July 2024

- Built analytics engine using Taxonomy/Ontology and Topic Modeling orchestrated with Dagster; extracted entities as metadata for fine-grained answers.
- Fine-tuned retriever for domain terminology and abbreviations; fine-tuned reranker for answer relevancy.
- Reduced agent handoff rate by 38% through improved retrieval precision and content routing.

Agent Auto Assist - Individual Contributor

- Fine-tuned Mistral 7B on 1.5M call-center conversations to power smart compose; collected accept/reject feedback and applied DPO for preference alignment.
- Reduced agent response time by 61% while maintaining brand tone and compliance.

Meta, Hyderabad - ML Engineer (HCL Client Project) Apr 2020 - Jan 2022

Test Automation CLI: Developed for Facebook's AR/VR/VA (Oculus, Ray-Ban Stories)

- Implemented image binarization with SciPy/Numba to cut OCR compute by **97%.**
- Deployed Detectron2 model for AR effects detection in WhatsApp/Messenger video calls; owned data pipeline, training, and deployment.
- Delivered DistilBERT-based textual similarity with high accuracy and low latency; CLI reduced execution time by **50%.**

HCL Technologies, Bangalore - Lead Engineer Apr 2020 - Jan 2022

- COVID-19 X-ray classification: modified DenseNet-121 achieved 96% accuracy; quantized for edge with 50–75% size reduction and 5–15% accuracy loss.
- Android app to benchmark ARM Mali GPU inference; improved GPU inference speed by 30% (FP32) and 80% (FP16) via TFLite optimization.

PUBLICATIONS

- Shreyas and Kameshwar (2021). "**Diagnostic Decision Support for Medical Imaging and COVID-19 Image Classification on ARM Mali GPU**". IEEE Global Communications Conference 2021 - Edge-AI-IoT - Published on 7th December 2021 (link)
- Shreyas and Dey, (2019). "**Application of soft computing techniques in Tunneling and Underground excavations: State of the art and future prospects**". Innovative Infrastructure Solutions. Springer Publications. Published on 25th September 2019 (link)
- Shreyas and Dey, (2019). "**Comparative Efficacies of Ensemble methods and Hybrid Intelligent model in Predicting Penetration Rate of Tunnel Boring Machine**". International IACMAG Symposium, IIT Gandhinagar 2019. Published on 7th March 2019

EDUCATION

Indian Institute of Technology, Guwahati Aug 2017 - June 2019

M.Tech in Civil Engineering

- Final CGPA: 9.79 / 10
- Secured 1st place in the merit list

P.E.S University, Bangalore July 2013 - May 2017

B.Tech in Civil Engineering

- Final CGPA: 9.14 / 10
- Secured 7th place in the merit list

ACHIEVEMENTS

- **Selected in Finals** (Out of 470 teams): **Bengaluru Mobility Hackathon 2024** (Organized by IISc, IUDX, BTP) - Vehicle Re-identification across multiple cameras for O-D Flows.
- **3rd Place** (out of 1200 teams): **SEBI Hackathon** (Global Fintech Fest 2023) - Developed an AI model to detect misleading claims by financial influencers using Generative AI.
- **Top 25%** (out of 400 teams): **Autonomous Agents Hackathon** (Lablab.ai, Weaviate)
- **4th Place** (out of 60 teams): **Generative AI Hackathon** (WhatFix, AWS) - Contributed to an AI solution for social good using Generative AI.
- **Special Mentions** at **Vista Hackathon** for demonstrating Content Genie - An all-in-one tool for social media post creation