Shubham Srivastava

Lead AI / ML Engineer 1

☎ +91 9170647101 ✉ shubhams.careers@gmail.com in LinkedIn

Over the past five years, I've driven the end-to-end development and deployment of production-ready AI solutions specialising in Large Language Models, Computer Vision, and MLOps. In my role as Senior ML/AI Engineer, I've led cross-functional teams to deliver 10+ client projects on schedule, collaborated with pre-sales to define technical requirements and timelines, and mentored junior engineers through model development and training. I am proficient in scaling AI/ML systems to address intricate business requirements and am now keen to apply this skill in a Senior ML Engineer or Senior AI Engineer role.

Skills

Python, Pandas, Numpy, Transformers, TensorFlow, MLFlow, WandB, NLP, CV2, FastAPI, RAG, Agentic AI, STT, TTS, Git, AWS (S3, EC2, Bedrock), Langchain, LlamaIndex, LLMs, VLLM, Fine-tuning with LoRA

Experience

Lead AI / ML Engineer - 1

GeekyAnts India Pvt Ltd — Apr 2024 – Present

- Led a team of 5 engineers and oversaw the organization's AI/ML programs.
- Delivered 4 client projects as tech lead with only 5% deviation from estimated timelines.
- Spearheaded development of RAG pipelines and an Agent Voice Interview Platform.
- Trained advanced models for business-specific documents, including layout-aware and table extraction models using VLLM, achieving over 90% accuracy.
- Closely collaborated with clients and internal teams for requirement gathering, solutioning, and time estimation.

Senior AI / ML Engineer - 2

GeekyAnts India Pvt Ltd — Apr 2023 – Mar 2024

- Designed solution architecture for complex microservices and scalable data pipelines.
- Mentored interns and junior engineers, planned their upskilling roadmap, and monitored execution.
- Delivered 3 major client projects end-to-end with strong focus on quality and deadlines.
- Developed advanced document chunking, retrieval, and re-ranking algorithms for a production-grade RAG system.
- Built object tracking and classification modules for an AI-driven exercise monitoring system.

Senior Software Engineer - 1

GeekyAnts India Pvt Ltd — Oct 2021 – Mar 2023

- Collaborated across multiple teams to architect and build complex AI-enabled systems.
- Delivered 2 key projects in e-commerce and computer vision domains.
- Developed a QR-based access control system for a workspace provider, achieving 100ms latency and integrating security compliance to prevent false access.
- Built a CV model to convert floor maps into Canva JSON format, enabling automated floor mapping for rapid inventory listing.
- Trained a model with 95% precision to classify the results of the pregnancy test kit, improving the efficiency and reliability of human diagnosis.

Software Engineer

GeekyAnts India Pvt Ltd — Jan 2020 – Sep 2021

- Delivered 6 diverse projects across React, mobile apps, computer vision, and speech-to-text domains.

- Developed a customized speech-to-text model based on open-source architecture, incorporating audio preprocessing techniques to achieve 10–12% WER on a dataset of 5,000 users.
- Engineered a high-throughput data pipeline for a medical IoT device, handling 1.5k records per minute per device.
- Built an advanced learning marketplace platform connecting students and teachers.
- Designed and implemented a machine learning-based camera filter for real-time image enhancement.

Projects

Advanced Document Extraction Pipeline

- Developed and fine-tuned custom models to automate data extraction from complex, business-specific documents, including long-form and table-heavy formats.
- Leveraged DocLayout and InternVL models for high-precision table layout and information extraction.
- Built a complete ML pipeline using Python, MLflow (for MLOps), and WandB, with custom data labeling handled via Label Studio.
- Deployed scalable inference infrastructure on GCP for production-grade use.
- Achieved 95%+ extraction accuracy, reducing the the turnaround time from 24 hours to less than 15 minutes.

Interview Voice Agent

- Built an AI-powered voice agent to autonomously conduct interview rounds, significantly reducing the processing time from days to hours.
- Integrated an agentic AI system to generate dynamic interview flows based on candidate resumes and job descriptions.
- Implemented advanced cheat detection and automated report generation mechanisms.
- Technologies used: LlamaParse (resume parsing), Langchain, OpenAI GPT-4o/4o-mini (interview + report generation), ElevenLabs (TTS), and Google STT.
- Achieved 98% accuracy in autogenerated reports with near-real-time interview performance on a highly scalable architecture.

Education

Post Graduation in ML and AI from IIIT Bangalore

Certifications / Courses

- Self-Driving Car Engineer – Udacity
- Intro to TensorFlow – Coursera
- Convolutional Neural Networks – Coursera
- Deep Learning Explained – edX
- Introduction to Machine Learning – NPTEL

Publications / Talks

- Paper on - Integrating mechanical driving function with custom circuitry - link
- Tech talk on - Agent Architectures & Design Patterns - link