Kushagra Tiwari

+91-8127328573 • kushagracse.techai@gmail.com • /kushagra-tiwari-a695a0268/ • /Kushagracse2803

Summary

Full Stack AI Engineer building production-grade GenAI systems end-to-end — from custom transformer architectures and multi-LLM orchestration to cloud-native deployment on Azure Kubernetes. Specialized in agentic RAG pipelines, QLoRA fine-tuning, and full-stack delivery with React and FastAPI.

Education

•Pranveer Singh Institute of Technology, Kanpur *2022 – 2026*
Bachelor of Technology in Computer Science and Engineering (AI) CGPA: 7.7 / 10.0

Professional Experience

•XprtCode *July 2025 – Nov 2025*
Artificial Intelligence Intern Remote

- Performed **data collection** from multiple sources and structured it into usable datasets for ML and RAG workflows.
- Carried out **data cleaning** and preprocessing with **Pandas** and **NumPy** on datasets exceeding 100,000 records, improving training efficiency.
- Built **RAG chatbots** using **LangChain** and **OpenAI**, reducing query resolution time by **40%**.
- Developed chatbots and exposed AI services through **FastAPI** endpoints.

Projects

•AI COPILOT FOR STOCKOUT RISK PREDICTION *GitHub*
Stockout Prediction & Replenishment System for a Bookstall Business Sep 2026 – Oct 2026

- Built an **XGBoost** model that predicts which products may run out of stock in the next 7 days across 3 branches.
- Created a rules engine that suggests how much to order and when to transfer stock between branches; every order stays a `draft` until a human approves it.
- Used **RAG** (**ChromaDB**) with **Llama 3.2** and **MCP** tools to explain each risk in plain English with policy citations.
- **Technologies:** Python, XGBoost, ChromaDB, Llama 3.2, Anthropic MCP, FastAPI, Docker, Kubernetes (AKS), React

•AI-POWERED INVESTOR INTELLIGENCE PLATFORM *GitHub*
Enterprise RAG Platform for Financial Document Intelligence & KPI Analytics Aug 2026 – Sep 2026

- Architected an enterprise-grade **RAG pipeline** to ingest annual reports and extract financial KPIs using **Azure OpenAI**.
- Built a modular **FastAPI** backend with ingestion, vector search, and KPI-storage services powering a chatbot and analytics dashboard.
- Containerized with **Docker** and deployed on **Azure Kubernetes Service (AKS)** for scalable, production-grade serving.
- **Technologies:** Azure OpenAI, Azure AI Search, FastAPI, LangChain, PostgreSQL, Docker, Kubernetes (AKS)

•ELYSIUM AI *GitHub*
GenAI Platform with Custom GPT-style SLM and Multi-Model Routing Oct 2025 – June 2026

- Built **Elysium SLM**, a GPT-style small language model trained from scratch in PyTorch, implementing custom tokenization, token embeddings, causal self-attention with masking, and stacked transformer blocks.
- Engineered an intelligent **multi-LLM router** to dynamically switch between Groq, Gemini, and the in-house **Elysium SLM**.
- Deployed multimodal GenAI tools (Quiz Generator, Summarizer, Resume Analyzer) with ATS scoring and skill-gap analysis.
- **Technologies:** PyTorch, Flask, LangChain, Hugging Face Transformers, Diffusers, Groq, Gemini, REST APIs

•PSYCHETRACK *GitHub*
Multimodal Mental Health Screening in Social Media Aug 2024 – Sep 2024

- Fine-tuned a lightweight **Phi-3.5 SLM** using **QLoRA** in PyTorch to support depression-risk screening and generate explainable reasoning for detected signs; intended as decision support, not a clinical diagnosis.
- Engineered a multimodal feature-fusion pipeline integrating text NLP, audio MFCCs, and visual CNNs, reaching **92% text, 90% audio, and 88% visual accuracy** on held-out test data.
- **Technologies:** Python, PyTorch, Hugging Face, QLoRA, Phi-3.5, TensorFlow, Keras, Librosa, OpenCV, Scikit-learn

Technical Skills

Languages: Python, Java, JavaScript, SQL, HTML, CSS
Libraries: PyTorch, TensorFlow, Hugging Face, LangChain, Diffusers, Librosa, OpenCV, Scikit-learn, NumPy, React.js
AI & Model Engineering: LLM/SLM Fine-tuning, QLoRA, Agentic RAG, MCP, LangGraph, Human-in-the-Loop
Inference & Deployment: Flask, FastAPI, Streamlit, REST APIs, Docker, Kubernetes (AKS), Git Control
Databases: MySQL, SQLite, PostgreSQL; **Vector DBs:** FAISS, ChromaDB
Tools: Git, GitHub, VS Code, Jupyter Notebook, Google Colab, LangSmith
Coursework: Data Structures & Algorithms, DBMS, Operating Systems, Computer Networks, Software Engineering

Coding Profiles

LeetCode: 250+ problems solved • **HackerRank:** 5-star in Data Structures & Algorithms