RATNESH KUMAR KUSHWAHA | BE, ME (Computer Science Engineering)

Email: rat.kush@gmail.com | Phone: +91-7898617389

PROFESSIONAL SUMMARY

Senior Software Engineer with 6+ years of experience in scalable microservices, Python, Data Science, Generative AI, and FastAPI/Django. Hands-on expertise with Huggingface, OpenAI, AWS, Agentic AI, RAGA, MCP and vector databases. Delivered LLM-powered projects using LangChain/LangGraph for retrieval-augmented generation, multi-agent orchestration, and tool integration. At Pace Wisdom, designed FinTech Kafka pipelines reducing latency by 30% and built fault-tolerant, high-performance data and API systems on AWS/GCP. Skilled in multi-AI agent architectures, technical leadership, code reviews, and mentoring teams for scalable delivery.

SKILLS

Languages/Frameworks: Python, Flask, Django, FastAPI, StreamLit, Gradio
ML & GenAI: TensorFlow, Scikit, HuggingFace, LangChain, PEFT, RAG, OpenAI, LangGraph
Data Tools: Pandas, NumPy, Postgres, MongoDB, Redis, ChromaDB, FAISS
Cloud & DevOps: AWS (Bedrock, IAM, Lambda, EC2, S3, RDS, Sagemaker), Gemini ADK, Docker, Git, GitHub Actions, CI/CD
ETL & Data Engineering: Scalable ETL pipeline design, data ingestion, transformation, Spark SQL
Others: Agile/Scrum, JIRA, Mentoring, Statistical analysis

EXPERIENCE

Star Co., Hyderabad | *Lead LLM / GEN AI Engineer* | **Aug 2025 – Dec 2025**
Worked in designing and implementing document automation solutions, RAG-based chatbots, and end-to-end generative AI pipelines. Developed and optimized workflows to handle and process transcript and PDF documents as well as base64-encoded data (Transcript, Invoices, MTR), including robust parsing, chunking, and retrieval mechanisms. Led prompt engineering, Text-to-SQL fine-tuning, and RAGAS-based evaluation, and developed internal tooling (TOON), Hugging Face, Cohere, OpenAI–based solutions, and Streamlit/Gradio applications to enhance accuracy, reliability, and user experience across AI products.

Pace Wisdom Pvt. Ltd., Bengaluru | *Senior Software Engineer* | **Apr 2022 – May 2025**
Led the development of LLM-powered agent-based chatbots using LangChain, Ollama, AWS Bedrock for intelligent query handling. Designed and reviewed RAG pipelines and NLP workflows integrated with Postgres and FastAPI. Architected FinTech Kafka pipelines, reducing processing latency by 30%. Delivered scalable microservices using Flask/Django, mentored a team of 5, and drove a 55% boost in engineering productivity. Dev and UAT reviews before Prod deployment.

Azilen Technologies Pvt. Ltd. | *Contract Application Developer* | **Jul 2021 – Dec 2021**
Migrated healthcare analytics platform from Scala to Python. Developed RESTful APIs and optimized data processing pipelines for extensibility, performance, and TDD-based architecture. Multiple MR code reviews before merging.

Infobeans, Indore | *Associate Software Engineer* | **Feb 2021 – Jul 2021**
Built secure NLP-based document classification APIs with email/OTP authentication. Implemented preprocessing pipelines and Hive-backed storage to improve data accuracy.

High IQ, Hyderabad | *Solution Architect* | **Dec 2019 – Feb 2021**
Engineered an OCR + CNN-powered PDF (Mortgage, Invoice)classifier on AWS SageMaker. Built scalable batch classification workflows using Parquet/JSON and optimized inference accuracy to 90%.

Emorphis, Indore | *Software Engineer* | **Sep 2018 – Nov 2019**
Developed multithreaded Python pipelines for IoT signal analysis. Built TensorFlow-based predictive services (Audio classification) with REST APIs, used AutoEncoder, mel-spectrogram

Saras InfoTech (Freelance) | *Independent Consultant* | **Dec 2014 – Aug 2018**
Delivered full-stack municipal portals (GRS/UMC) with backend APIs, dashboards, and database solutions. Designed custom ETL pipelines and Java-based reporting interfaces.

CERTIFICATIONS

- Introduction to Model Context Protocol (Anthropic): MCP Client/Server, Claude, Prompt
- 5-Day AI Agents Intensive Course with Google: Agent AI, Multi agents, Session, Memory
- Building RAG Agents with LLMs (NVIDIA): LLM, LangChain
- Big Data Hadoop & Spark (Udemy): PySpark, Hadoop, SQL (IISc Bangalore)
- 5-Day Gen AI Course (Kaggle): Vertex AI, Google Gemini
- Problem Solving (HackerRank): Python, SQL
- Android Workshop (Techfest, IIT Bombay): Android SDK, Java

KEY PROJECTS

Text-to-SQL Chatbot: Implementing prompt-engineered and fine-tuned LLMs to convert natural language queries into SQL over a contextualized database schema. Developed a RAG-based pipeline with systematic RAGAS evaluation, achieving 83% accuracy on query-to-SQL generation with schema-aware context injection. Integrated the solution with a Gradio interface for interactive querying, testing, and rapid iteration with business users.

Transcript processing project: Implementing prompt-driven extraction of structured data from PDF transcripts using Cohere and GPT models, with specialized handling for table detection and parsing . Designed and developed an end-to-end asynchronous pipeline that converts raw PDFs into clean, validated JSON responses, supporting concurrent processing for large document batches. Implemented streaming-style responses for a responsive user experience, enabling real-time visibility into extraction progress and incremental data delivery .

Voice AI Agent (Whisper, SST, TTS, Multimodal): This project provides a complete solution for processing audio files through a conversational AI agent. It features a FastAPI backend for processing, a Streamlit frontend for user interaction, and uses a locally-run Ollama model to ensure privacy and avoid paid APIs. Hands-On experience in creating MCP servers and AI agents (Anthropic, Praison agents)

Social Good Website (LLM, RAG, FastAPI, Postgres, Ollama): Designed and built RAG pipeline using Ollama and HuggingFace with 95% accurate regulatory query handling. Integrated finetuning, caching, and fault-tolerant with evaluation (pytest, RagChecker) and monitoring implementations (LangGraph).

Leave Management System (OpenAI, HuggingFace, Django): Rule-based engine + LangChain chatbot for natural language leave queries. Prompt-tuning, REST APIs, and continuous quality tracing and evaluation with different metrics (faithfulness, relevantness, LLM as Judge, pytest)

Appreciate App (FinTech, Flask, ML, Kafka, MongoDB): Built ML pipeline for churn/engagement. Scalable Flask APIs, Kafka ETL jobs, LangChain NLP-based insights. Delivered 65% increase in client engagement.

Intelligent Document Classification (SageMaker, CNN, Tesseract): OCR + CNN pipeline for PDF classification (85% accuracy). Used Pandas, NumPy, and multithreading. Deployed using SageMaker with batch monitoring. Evaluated using a confusion matrix.

IoT Anomaly Detection (TensorFlow, Pyspark, Autoencoder): Real-time MQTT/TCP stream ingestion with Autoencoder-PCA models. Used FFT feature extraction, multithreaded ingestion, and applied various statistical techniques. Done anomaly detection on edge device.

EDUCATION

M.E. (Computer Engineering) 2011-13 – SGSITS, Indore (Govt. Autonomous) | *CGPA: 7.56*
Thesis: Semantic Web Service Composition on Cloud (QoS View)

B.E. (Computer Science) 2006-10 – UEC, Ujjain (Govt. Autonomous) | *Percentage: 65.59*
Major Project: Smart City

IMP URLs

https://ollama.com/imratnesh/ratnesh-bot

https://linkedin.com/in/ratneshkushwaha

http://github.com/imratnesh

http://kaggle.com/ratnesh88

http://stackoverflow.com/users/2094351/ratnesh-kushwaha