HIMANSHU KUMAR SINGH



IIT Ropar | Senior Backend | AI Engineer – 5+ Years Experience

hs230998@gmail.com | +91-9877381254 | LeetCode Profile | India, Delhi
himanshutech4purpose | himanshutech4purpose | My Portfolio Profile

ABOUT ME

Senior Backend | AI Engineer with 5+ years of experience designing and deploying production LLM systems, AI agents, and distributed backend services. Hands-on expertise in Text-to-SQL, RAG architectures, document intelligence, and LLM evaluation, with real-world deployments in government and compliance-sensitive domains. Prioritizes system correctness, observability, and long-term maintainability over demos.
Conducted 100+ technical interviews and mentored freshers for Beehyv software Solutions and EB1Aexperts.

SKILL

- Languages – Python (Expert, strong in Data Structures & Algorithms), JAVA, Go, JavaScript
- AIML – Pytorch, Generative AI, LLM Applications, LangChain/LangGraph, Transformer Architectures (BERT/GPT), Diffusion Models, CNNs, Agentic AI, Multimodal AI (Text/Image/Audio/Video), RAG Systems, Reinforcement Learning
- Backend – Fastapi, Flask, Spring boot, NodeJs with SQl & noSQL
- Frontend – React, Nextjs
- Devops – GCP, AWS, Github Action, Docker, Prometheus, Grafana, Kong API Gateway, Kubernetes (EKS)

WORK EXPERIENCE

Senior Data Scientist

Tredence Inc. | July 2026 – Ongoing | Gurugram

- Developed Text to Sql agent Audience Plateform for marketing teams at Target Corporation. Handled memory management, Evaluation, Tracing, latency optimization and RAG architecture.

Senior Software Developer

Beehyv Software Solutions Private Limited | August 2022 – July 2026 | Gurugram

GIDR.ai – Research & Production AI Systems

- Evaluation & Benchmarking Systems: Developed evaluation frameworks for chatbot and text-to-SQL systems; improved SQL generation accuracy from 50% to 95%+ by implementing dynamic few-shot prompting using query logs; created custom datasets and conducted prompt fine-tuning based on research papers
- Built a B2B, voice-driven AI workflow platform (React + FastAPI) enabling businesses to configure document-to-action pipelines; implemented extraction engine that converts PDFs/images into structured JSON based on admin-defined rules, powering dynamic frontend flows for form filling, troubleshooting, and guided interactions
- Developed AI agent–driven UI (DOM manipulation + voice) for step-by-step navigation, input capture (forms, signatures, media), and task automation; integrated state-machine workflows (pytransitions) and advanced PDF parsing (tables, TOC, layout-aware chunking)

EB1A expert – Backend, GenAI, Devops

- Led the development and deployment of an AI-powered EB1A platform with production-grade FastAPI backend services, React frontend, and CRUD APIs powering internal operations and customer-facing AI applications, automating profile generation, evaluation, and verification workflows to reduce manual review effort.
- Engineered a multi-agent EB1A recommendation engine that analyzed applicant documents, enriched evidence using RAG, web-search, MCP agents, benchmarked profiles against historical successful cases, and generated criterion-wise approval probability scores with actionable recommendations for strengthening individual USCIS EB1A criteria (awards, publications, judging, memberships, patents, speaking engagements, and original contributions).
- This system was developed 0 -> 1 in 1 year and able to attract 1000+ clients and leads. Also reduced manual human effort of content writer by 70 percent
- Tech: FastAPI, Python, AWS, Grafana Stack, Kong, GitHub Actions, SpaCy, BERT, Multi-Agent Systems

Kumbh Sahayak – Data Scientist

- Handled user base of 10 million+ users and 10k+ concurrent requests for Kumbh Mela 2025. Achieved accuracy of 90%+ on evaluation dataset with llama3:8b. Developed a multilingual chatbot by optimizing LLM performance through structured prompt tuning (TextGrad, DSPy) and fine-tuned BGE embeddings for improved semantic retrieval; implemented response safety using AWS Guardrails and custom rule-based systems. Designed and implemented a deepfake video pipeline for CM speech synthesis using audio cloning (ElevenLabs, Coqui-TTS) and Wav2Lip-based lip sync, collaborating with the Samagra team.

Beehyv internal products – Backend Engineer/AI engineer

- Built the evaluation system for multi-tenant distributed AI agent workflow leveraging the deepeval library
- Beecoach takes adaptive assessment from user based on topic and concept set by admin using LLM. Tech used: FastApi, Postgres, docker compose, llamma, Openai, React

EasyKnock (Real Estate business) – Senior Data Engineer

- Processed and transformed data 40+ sources. Reduced failure rate by 90% and enabled notification for faiure on slack and email. Automated schema evaluation and backfilling of data through custom operators.
- Built data pipeline and batch automation task for business (portfolio valuation and portfolio acquisition), sales (Salesforce, Rentvine, Zendesk, Gmail data), accounting (Netsuite data automation), marketing (Google, facebook, bing ads with ML to prioritize leads) and Data Analyst (microservice data sync and views for reporting purpose) team. Migration and Deployment of Machine leaning and airflow related code base from GCP to AWS. CICD related to Airflow and DBT using Github Action
- Tech used: *Python*, *Airflow*, *Github Actions*, *Docker*, *Kubernetes*, *AWS*, *GCP*, *Nodejs*

BluBracket (Acquired by HashiCorp)) – Backend Developer

- Built software bill of material python module for security analysis and contributed Multi-Task Scheduler code base.

Data Scientist (Financial Intelligence Unit of India)

LTIMindtree — May 2021 – August 2022 — Delhi

- Built the application Data Quality Report in python framework. It takes batch data from 17 different types of financial institution like Banks, exchange house etc in different format and checks the quality of data based on 32 rules like duplicate reports, unique id validation rules, external validation of attributes from 8 different institution via API, ML rules etc. Hands-on Data Modelling and View creation involving multiple tables for consumption of user interface, network visualization using Neo4j tool, implementation of Risk score, Data preprocessing and EDA using python library, SQL queries

PERSONAL PROJECTS

Personal full stack projects - Visit my portfolio

Full-stack engineering, AI experimentation

- Built and deployed **MycoLoop**, an e-waste management platform connecting users with pickup agents for recycling/disposal of electronics. Also developed a booth management system with demographics analytics and heuristic-based fake voter detection using ECI voter-list PDFs for civil society.
- Armour Design with Deep Learning – B.Tech one-year coursework project. Our team applied LSTM and neural network models on simulation datasets of armor bullet impact. Through literature review, we found researchers mostly used simple neural networks. Our work focused on reducing the number of simulations by leveraging intermediate time-series simulation data using LSTM, making the approach more computationally efficient.
- Image classification with CNN and Grad-CAM for evaluation, Hindi-to-English translation using attention models, implementation of RNN, LSTM, WGAN, and VGAN, and trajectory prediction of chaotic double-pendulum systems using deep learning.

CERTIFICATIONS

- **AWS** – Udemy
- **Distributed Systems** – NPTEL
- **Apache Kafka** – Udemy
- **MLOps** – Coursera
- **Apache Airflow** – Udemy
- **CKAD – In Progress**

EDUCATION

IIT Ropar B.Tech (Hons) in Mechanical Engineering — July 2017 – May 2021 — Major CGPA: 7.61/10 • Minor in Physics: 8.0/10

Intermediate **CBSE** — July 2014 – May 2016 — Score: 90%