Utkarsh Upadhyay

☎ +91 8824848425 ✉ utkarsh.upadhyay.in@gmail.com in utkarsh-upadhyay utkarsh-ls

Experience

Serri — **Nov 2024 – May 2025**
Machine Learning Engineer (LLM Deployment, AI/ML, Backend Development) — *Remote (India)*

- Developed backend for an agentic chatbot using LangGraph, enabling natural language tool-calling and autonomous next-action suggestions, with response times under $< 20s$.
- Optimized token/memory usage, reducing per-conversation token consumption from 30K to 3-5K (80-90% reduction).
- Built an agentic RAG-based Q&A chatbot employing hierarchical chunking and metadata tagging for precise retrieval.
- Deployed LLMs on Google Vertex AI and Cloud Run, benchmarking latency and performance for production workflows.

Samsung R&D — **Aug 2023 – Sep 2024**
Software Engineer (Language AI framework & MDE) — *Bangalore, India*

- Integrated generative AI with Bixby, enhancing user interaction and boosting model accuracy by 5% over previous rule-based approaches.
- Led the development, delivery, and maintenance of 3 Bixby modules deployed in Galaxy Z Fold 6 and Z Flip 6 devices.
- Proposed a PoC on diverse LLM prompting techniques, increasing accuracy for reasoning tasks by 15%.
- Mentored a team of 4 college students, guiding the team through the research paper publication process.

Validus Techfin Services Pvt Ltd — **May 2022 – Jul 2022**
Software Developer Intern (FullStack Developer) — *Hyderabad, India*

- Engineered a web application to manage Alternative Investment Funds (AIFs), streamlining investment processes for the client company's Fund Managers.
- Formulated an asynchronous back-end structure for data handling, boosting data processing speed by 50%, and devised a UI for data capture.
- Orchestrated the deployment of services on AWS using independent Docker containers within an EC2 instance.

Blue Lit Solutions LLP — **Jan 2021 – May 2021**
Software Developer Intern (ML, FullStack Developer) — *Remote (India)*

- Collaborated with a 4-member team on developing a cloud-based system on AWS for uploading and classifying waste object images using a micro-service architecture and a UI to visualize ML model responses.
- Leveraged Docker for the deployment of 6 micro-services, ensuring scalable and consistent environments.

Projects

Wikipedia Search Engine | *Python, Search Engine* — **Mar 2023**

- Developed an efficient search and indexing engine for English and Hindi Wikipedia dumps, implementing blocked sort-based indexing and TF-IDF ranking, resulting in a 50% increase in query speed.

Intent Detection Model | *Python, React, FastAPI, ML, Transformers* — **Nov 2022**

- Built a neural network upon pre-trained models for user intent and slot identification.
- Created a UI adopting the ML model for speech input and output, achieving a response time of under 3 seconds.

Anaphora Resolution Model | *Python, ML, Transformers* — **Nov 2021**

- Leveraged an anaphora resolution system for user-generated text data retrieved from Twitter conversations.
- Employed 2 binary classifiers: the Mention model and the Pair-Score model, to determine whether a pair of mentions constitutes an anaphora-antecedent pair.

Education

International Institute of Information Technology, Hyderabad — **Aug 2019 – Jul 2023**
B.Tech in Computer Science — **GPA:** 8.93/10.00 — *Hyderabad, India*

- Ranked in the top 5% on the college Merit List for academic excellence.
- Taught as a Teaching Assistant for 2 semesters in Probability and Statistics.

Skills

Languages & Tools: Python, C++, Bash, Git, Docker, AWS, GCP, Azure, Atlassian, C, Java, R, SQL.
ML & Automation: Prompt Engineering, RAG, Token Optimization, Transformers, Vector Databases.
Backend Development: Django, Flask, FastAPI, PostgreSQL, MongoDB.

Extracurricular/Achievements

Competitive Ratings: Codeforces — 1607, Codechef — 1930, Atcoder — 623.
JEE Advanced: 1617 AIR, NTSE and KVPY Scholar (top 1%).