Harsh Dwivedi

New York, United States | +1 (680) 216-2032 | edwivediharsh@gmail.com | LinkedIn | GitHub | Portfolio

WORK EXPERIENCE

iConsult Collaborative, *Syracuse, New York* **Mar 2025 – Present**

Team Lead & Software Developer

- Leading an Agile team, streamlining sprint planning, backlog grooming, and PR workflows in **Jira** and **GitLab** to boost delivery velocity and improve architecture decision cycles.
- Developed and deployed a scalable, SEO-optimized web app (**Next.js**, **TypeScript**, **React**, **Tailwind CSS**) with lazy loading and SSR, achieving a Lighthouse SEO score of **95+**.
- Implemented client-side routing and server-rendered navigation via **Next.js App Router** and **Context API**, improving load speed and in-app UX.
- Built secure **RESTful backend APIs** and **microservices** for **AWS DynamoDB** with **Cognito**-based authentication, session management, and role-based access control.
- Deployed containerized **CI/CD workflows** with **GitHub Actions**, **Docker**, and **AWS EC2**, cutting deployment time by **40%** and enabling zero-downtime releases with automated rollback.
- Developing a high-speed API layer for **AWS ElastiCache (Redis)** to support real-time queries and planning **Apache Kafka** integration for event-driven streaming.

NEXIS AI Lab, *Syracuse, New York* **Jan 2025 – Present**

AI Researcher

- Designed token-level saliency visualizations for **RAG pipelines** using **LangChain**, **OpenAI API**, and **PyTorch**, improving explainability for technical and non-technical audiences.
- Applied embedding-based semantic similarity search using **FAISS** and **Hugging Face Transformers** to enhance entity matching and improve document retrieval accuracy.
- Experimenting with interpretable NLP methods (**attention rollouts**, **attribution mapping**, **SHAP**) to boost model traceability and decision transparency.

Tata Consultancy Services (TCS), *Mumbai, India* **Jun 2023 – Jul 2024**

Assistant System Engineer

- Delivered backend support for European IT operations, integrating **Python** scripting into enterprise systems to improve uptime, security, and compliance checks.
- Automated ticket resolution and infrastructure tasks with **Python** scripts integrated into **ServiceNow REST APIs**, reducing manual workload by **4 hours/day** and cutting resolution time by **30%**.
- Built monitoring scripts with **PowerShell**, **Python**, and **Azure Monitor APIs** to track uptime, detect anomalies, and speed up incident response; implemented unit testing and integration testing to validate functionality.
- Created internal tooling for **Azure Active Directory** and **RBAC** policy management, strengthening authentication workflows and security governance.
- Partnered with DevOps teams to manage servers and **Azure** resources via **CLI** and **IaC** practices, improving deployment speed, scalability, and ensuring compliance with security protocols.

EDUCATION

M.S. in Computer Engineering, Syracuse University, Syracuse, NY **Aug 2024 – May 2026**

Relevant Coursework: Advanced Data Structures & Algorithms, Object-Oriented Design, Application Programming, DBMS, NLP

TECHNICAL SKILLS

Programming Languages: Python, Java, C, C++, C#, TypeScript, JavaScript

Web & Frameworks: React, Tailwind, Node.js, Next.js, Express, Flask, FastAPI, Spring Boot, Selenium, JUnit, Apache Spark

Cloud & DevOps: AWS (EC2, S3, DynamoDB, RDS), Azure, Docker, Kubernetes, RBAC, REST APIs, GraphQL

Databases: PostgreSQL, MySQL, MongoDB, Firebase, DynamoDB, Redis, Kafka, Scala

AI & ML: Scikit-learn, PyTorch, TensorFlow, Apache Spark, BERT, Hugging Face, OpenAI, LangChain, RAG pipelines, XAI

Tooling & Platforms: Git, Jira, Postman, Poetry, VS Code, ManageEngine, Power BI, Power Automate, Cursor, Windsurf

PROJECTS

RAG-Based Capability Matrix Analysis System **Mar 2025 – Present**

Tech Stack: Python, LangChain, FAISS, Chroma, FastAPI, Docker

- Designed and implemented a **Retrieval-Augmented Generation (RAG)** pipeline for intelligent document comparison and capability analysis across technical specifications, research papers, and compliance reports.
- Developed backend services using **LangChain, OpenAI API, and FAISS/Chroma vector** stores to retrieve contextually relevant document segments with reasoning, confidence scores, and supporting evidence.
- Optimized **text chunking**, **semantic search**, and **vector store caching** to reduce retrieval latency and improve **LLM** accuracy on large-scale document collections.