# Suryansh Singh Rawat

+1 2064965588 suryansh@uw.edu suryansh-singh-rawat xsuryanshx snareyansh suryansh.space

## Education

**University of Washington, Seattle** **Sept 2025 – Mar 2027**
*Master of Science (M.S), Data Science* *Seattle, WA*

**Birla Institute of Technology and Science (BITS), Pilani** **Aug 2018 – July 2022**
*Bachelor of Engineering (B.E), Electronics and Instrumentation* *Goa, India*

## Professional Experience

**Ascentt (Toyota Motors North America)** **March 2024 – Sept 2025**
*Principal AI Engineer* *Remote, US*

- Led the development of **ToyotaGPT**, an **enterprise-grade** chat platform (ChatGPT equivalent), scaling to **74,000+ DAUs**.
- Designed scalable **REST APIs** in Python and integrated user data management with **DynamoDB**, building a robust **RAG pipeline** capable of extracting, ingesting, and retrieving **13+ complex unstructured file types**.
- Optimized RAG pipelines for chatbots across **4 enterprise products**, implementing lossless data preprocessing, chunking and indexing in **Amazon OpenSearch Vector DB**, and leveraged **RAGAS evaluations** to benchmark index performance, leading to **25%** gain in **Precision@10** and **12%** gain in **Recall@10**.
- Built a metadata-extracting chatbot using **LangChain** and **Pydantic** for precise car model searches in **PGVector DB**, researched and implemented a **Dynamic RAG approach** to intelligently route queries based on type and complexity.
- Built an **Agentic RAG chatbot** with **multi-source retrieval** across **OpenSearch**, **PostgreSQL** and other enterprise datasources using **LangGraph**, handling complex user queries over five enterprise datasets **(1TB+ data)**.

**Deloitte USI** **July 2022 – March 2024**
*Data Scientist/ML Engineer (AI Center of Excellence - CyberSecurity Pod)* *Bangalore, India*

- Developed custom **Autoencoder architectures** along with **Graph Network** based preprocessing with **Neo4J** to identify anomalies for detecting **Zero-Day Cyber Threats** using realtime cloud network flow data.
- Leveraged **Deep Learning NLP** techniques to develop *Deloitte IRL* tool (multi-document to single-document requirements library automation tool). Created **multiclass ensemble classification** models and **unsupervised clustering** models with **HDBSCAN and t-SNE**, finetuned document embeddings with **Metric Learning** using Sub-center ArcFace Loss.
- Constructed a **Text-Summarization Model** for generating cyber threat reports using **SecBERT**. Incorporated **NER Tagging and Masking** with **SpaCy annotation**. Developed model pipeline leveraging **Flask and Docker** for efficient deployment.
- Built **Generative AI** solutions for a regulatory compliance tool by creating **RAG powered Entity Extraction and Summarization models** using LLM models like **Llama 2 and GPT-4**, orchestrated pipelines using **Langchain** and deployed models using **Streamlit**.

**Scienaptic AI** **July 2021 – Dec 2021**
*Data Scientist Intern* *Bangalore, India*

- Utilized **XGBoost and Logistic Regression** algorithms to create **Credit Underwriting ML Models** to forecast the probability of credit accounts that may default, enabling an **11.5% increase** in application approvals.
- Designed a Unified Credit Underwriting model framework leveraging **100MM+ raw credit records**, deployed to deliver baseline performance metrics for credit unions across diverse risk segments.

## Projects

**OpenProbe** | *Python, Langchain, LangGraph* **May 2025 – June 2025**

- Built OpenProbe, an **open-source deep research agent to answer complex queries** that works with any LLM.
- It works using a state-based orchestration of agents/tools like **websearch, coding and logical reasoning** via **LangGraph**.
- OpenProbe, paired with DeepSeek-R1 and Qwen3-32B, **outperforms** OpenAI's GPT-4o-Search on the challenging multi-hop **FRAMES benchmark (DeepMind)**, achieving **67.1% accuracy** (+1.5%).

**Detecting GAN Generated DeepFake Images** | *Python, Tensorflow, Deep Learning* **Jan 2021 – June 2021**

- Researched on developing a custom CNN architecture to efficiently detect GAN generated DeepFake images.
- Achieved precision and recall matching state-of-the-art methods, with **97.77% accuracy** on the **StyleGAN** dataset.

## Technical Skills

**Languages**: Python, SQL, C++, Java, TypeScript, Bash, MATLAB, R
**Developer Tools**: Git, AWS, Microsoft Azure, Docker, S3, DynamoDB, PostgreSQL, MongoDB, Redis, SQLite
**Frameworks**: PyTorch, TensorFlow, Keras, SpaCy, HuggingFace, CUDA, LangChain, LangGraph, Streamlit, FastAPI, Selenium
**Deployment & MLOps**: Airflow, Amazon EKS, Grafana, DataDog, LiteLLM, vLLM
**Relevant Courses**: Data Structures and Algorithms, Advanced Calculus, Linear Algebra, Neural Networks and Deep Learning, Statistical Machine Learning, Applied Statistics and Probability, Scalable Data Systems and Algorithms