

Nikhil Obuleni

Tel: (571) 478-2007 | Email: nikhil.obuleni@gwu.edu | [LinkedIn](#) | [Github](#) | [Portfolio](#) | Washington, DC

Education

George Washington University

Master of Science in Data Science | GPA: 3.75

Jan 2025 – Dec 2026

Washington, DC

- **Relevant Coursework:** Machine Learning, Deep Learning, Natural Language Processing, Statistical Modeling, Data Mining, Data Visualization

Professional Experience

Center for Global Mental Health Equity (CGMHE)

AI Research Engineer

Mar 2026 – Present

Washington, DC

- Shipped an LLM agent pipeline (**LangChain, OpenAI API**) that automates counselor competency scoring across **26K+ multilingual clinical transcripts** in 25 countries, cutting per-session review time from **45 min to 4 min** and replacing a fully manual WHO/UNICEF evaluation process.
- Designed a production **RAG system (FAISS, LangChain, FastAPI)** grounding LLM outputs in EQUIP/ENACT clinical rubrics through chunking, reranking, and a hallucination-flagging layer; reduced factual error rate by **~40%** on held-out multilingual benchmarks.
- Ran LLM evaluation experiments (**RAGAS, DeepEval**) comparing GPT-4 and Claude against human-annotated ground truth across **14 languages**; identified a **23% calibration gap** on low-resource languages, with findings feeding into a joint OpenAI collaboration on clinical LLM deployment.

Data Science for Sustainable Development (DSSD)

ML Engineer (Part-time)

Feb 2025 – May 2026

Washington, DC

- Automated ingestion and structuring of unstructured sustainability reports using **NLP pipelines (spaCy, Hugging Face)** and LLM-assisted extraction (OpenAI API, LangChain) for NGO clients, reducing manual data prep time by **~60%** per project cycle.
- Developed open-source **REST APIs (FastAPI, Python)** and modular data pipelines for nonprofit and government partners, integrating third-party data feeds with automated schema validation to improve reliability of SDG-aligned reporting tools.
- Deployed ARIMA, LSTM, and XGBoost ensemble models on **200K+ humanitarian records** across 18 countries, achieving **78% weighted F1** on disaster impact forecasting for field teams with limited data infrastructure.

Asteria Aerospace

Data Science Intern

May 2023 – Aug 2023

Bangalore, India

- Deployed **computer vision models (YOLOv5, Faster R-CNN)** for real-time drone object detection, enhancing surveillance efficiency by **30%**; built anomaly detection with Autoencoders and Isolation Forest, improving fault detection by **40%**.
- Deployed predictive maintenance models (LSTMs, Random Forest) forecasting equipment failures and reducing unexpected breakdowns by **25%**; conducted geospatial analysis (GIS, LiDAR, satellite imagery) improving navigation accuracy by 20%.

Technical Skills

Languages & ML: Python, SQL, R · PyTorch, TensorFlow, scikit-learn, XGBoost, LightGBM, Hugging Face, spaCy, BERT
LLM & Agents: OpenAI API, Anthropic (Claude) API, LangChain, LangGraph, RAG, Prompt Engineering, Multi-Agent Systems
RAG, Infra & Eval: FAISS, Pinecone, pgvector · FastAPI, Docker, AWS (Lambda, EC2, S3, SageMaker), GitHub Actions, MLflow · RAGAS, DeepEval, LangSmith
Data Engineering: Apache Airflow, BigQuery, PostgreSQL, MongoDB, PySpark, ETL Pipelines, REST APIs
Methods: NLP, Computer Vision, Statistical Modeling, Time Series Forecasting, Anomaly Detection, SHAP Explainability, A/B Testing

Projects

medrag-toolkit – [GitHub](#) / *LangChain, LlamaIndex, FAISS, Qdrant, vLLM, Ollama, FastAPI, Docker*

2026

- Developed a modular RAG framework for medical QA with pre-built connectors to **PubMed (36M+ abstracts)**, RxNorm, and OpenFDA; every response enforces citation grounding and passes through an automated hallucination-detection layer.
- Implemented **hybrid retrieval (dense FAISS + sparse BM25 over Qdrant)** with iterative query refinement; benchmarked against MedQA/MedMCQA and ships with a streaming FastAPI layer supporting local (Ollama, vLLM) and cloud LLM backends via Docker.

Clinical RAG Evaluation System – [GitHub](#) / *LangChain, Pinecone, GPT-4, FastAPI, AWS Lambda, RAGAS*

2026

- Built a RAG pipeline over **5,000+ WHO and NIMH clinical guidelines** (LangChain, Pinecone, semantic chunking), deployed on AWS Lambda via FastAPI; achieved **91% response relevance** and **38% hallucination reduction** vs. GPT-4, benchmarked via RAGAS.
- Integrated prompt injection guardrails and LLM evaluation logging via LangSmith; cut clinician response-verification time **45%** in usability testing with 3 clinical reviewers.

PlanMoE – AI Floor Plan Designer – [GitHub](#) / *MoE, GCN, HouseGAN++, PyTorch, FastAPI, HF Spaces*

2026

- Designed a custom **Mixture of Experts (MoE)** with a 3-layer GCN encoder and transposed-convolution MaskDecoder; each expert specializes in a room-type layout, with a learned gating network routing room-constraint graphs to the right expert at inference time.
- Deployed as a HuggingFace Space (FastAPI backend); applied INT8 post-training quantization reducing inference latency by **52%**, enabling real-time floor plan generation on a single GPU.