

Nitin Singh Rathore

+1(817) 819-8146 | nxr3560@mavs.uta.edu | www.linkedin.com/in/nitin-singh-rathore | github.com/Nitin3560

Profile Summary

Results-driven AI & Software Engineering professional with 2+ years of experience in building LLM-based applications, RAG pipelines, and scalable cloud systems using LLaMA, AWS, Docker, and GCP. Skilled in Python, TensorFlow, Flask, Streamlit, and full-stack development, supported by a strong foundation in data structures and system design. Currently pursuing an MS in Computer Science at UT Arlington, researching telecom-specific RAG systems and LLM evaluation methods to develop domain-adaptive AI solutions for real-world industry applications.

Education

THE UNIVERSITY OF TEXAS - ARLINGTON, TX

Jan 2025 - Dec 2026

Master of Science in Computer Science

Relevant Coursework: Machine Learning • Deep Learning • NLP • Cloud Computing • Data Structures & Algorithms • Information Retrieval

Technical Skills

- **Programming Languages:** Python • C++ • C • JavaScript • TypeScript • Java
- **Frameworks & Libraries:** Node.js • React • Flask • Streamlit • TensorFlow • LangChain • Next.js
- **Databases & Caches:** MySQL • AWS RDS • SQLite • FAISS • Pinecone • AWS S3
- **Cloud Services:** AWS (EC2, ECS, S3, IAM, CloudWatch, CodePipeline) • GCP (Cloud Run, Cloud Build) • Docker • CI/CD Pipelines
- **Project Management Systems:** Git • Jira • VS Code • Postman • Linux/Ubuntu • API Integrations

Ongoing Research

- Built a question-answering pipeline using Meta's LLaMA 2 model hosted locally on an AWS EC2 GPU instance for efficient inference.
- Integrated AWS S3 for uploading and storing multiple PDF documents, enabling scalable and organized document management.
- Extracted text from uploaded PDFs using PyMuPDF and chunked the content using LangChain's RecursiveCharacterTextSplitter.
- Designed prompt templates to combine user queries with relevant document chunks for context-aware answer generation.
- Implemented FAISS-based vector search to retrieve the most relevant document chunks efficiently using cosine similarity.
- Evaluated responses using custom scoring metrics and LLM-as-a-judge to measure retrieval accuracy and reduce hallucinations.

Experience

The University of Texas - Arlington

Graduate Teaching Assistant — CSE Department, UT Arlington

Aug 2025 - Present

- Supported 100+ students in mastering C/C++ and debugging techniques, designing practical examples and walkthroughs that raised overall assignment success rates.
- Improved course efficiency by helping automate grading workflows and restructuring Canvas content, enabling smoother communication between students and faculty

WERBOOZ Pvt. Ltd., Indore, India

Junior Software Developer

Sept 2023 - Oct 2024

- Designed and developed scalable applications and integrations using Java and Apex; implemented Salesforce automation, Apex triggers optimized for bulk operations, and optimized SOQL/SQL queries.
- Integrated external systems through REST and SOAP APIs, improving data flow and interoperability across platforms.
- Executed system integration and regression testing by authoring 500+ test cases (Tosca, Postman, JUnit) with data mocks; this reduced post-release defects by 30% and caught breaking changes earlier.
- Collaborated with cross-functional teams and clients to convert requirements into user stories (Jira), produced low/high-level designs, led UAT demos, managed dependencies, and delivered sprints with minimal rework.

Software Developer Intern

Feb 2023 - Sept 2023

- Assisted in developing and testing code for healthcare management systems using industry-relevant tools and languages, contributing to a 15% improvement in module-level efficiency through optimized logic and cleaner implementations.
- Conducted comprehensive unit, integration, and system testing, identifying and helping resolve over 20 bugs before deployment, which improved system reliability and reduced post-release issues by 25%.
- Maintained clear, well-documented code and deployment guides using Git for version control, enabling smoother collaboration across a 5-member team and reducing onboarding time for new contributors by 30%.

Projects

Retrieval-Augmented Generation for Telecom Optimization (Research Thesis)

- Fine-tuned LLaMA models to telecom datasets using domain-specific embeddings and custom chunking strategies to enhance RAG performance.
- Designed a JSON-based scoring framework & LLM-as-a-Judge evaluation method to compare retrieval strategies and fine-tuning approaches.
- Conducted ablation studies comparing embedding models (BERT vs Sentence-T5 vs LLaMA) and quantified retrieval improvement by 23%.
- Implemented evaluation dashboard with Next.js + Python APIs to visualize retrieval confidence, hallucination rate, and KPI accuracy.
- Explored research extensions in hallucination reduction and domain adaptation for telecom verticals.

Cloud-Native Microservices Pipeline for AI Inference

- Containerized AI services using Docker and deployed via AWS ECS + GCP Cloud Run with autoscaling & load balancing.
- Built CI/CD pipelines using AWS CodePipeline & GCP Cloud Build to automate zero-downtime deployments (blue/green strategy).
- Hardened infrastructure with IAM roles, S3 encryption, VPC flow logs, and CloudWatch monitoring.
- Converted LLaMA-based inference into scalable REST microservices capable of serving telecom document queries.
- Reduced latency by 32% and improved inference reliability with caching & resource tuning.