Rajeev Atla

Machine Learning Engineer | GenAI, RAG & Search Systems Specialist

US Citizen | 732-209-3995 | rajeev.atla@gmail.com | linkedin.com/in/rajeev-atla | github.com/RajeevAtla | rajeevatla.com

Education

Rutgers University - School of Engineering — Sep 2021 — May 2025

***Bachelor of Science** (Triple Major) in Computer Engineering, Computer Science, and Data Science* — *New Brunswick, NJ*

Recipient of the Eleanor and Samuel Sneath Endowed Merit Scholarship for Engineering Students

Coursework: AI, ML, Distributed Deep Learning, Data Science, Robotics and Computer Vision, Info and Network Security

Skills

- **Programming Languages**: Python, R, SQL, Java, C/C++/CUDA, JavaScript/TypeScript, Rust, Bash
- **AI/ML**: NumPy, PyTorch, JAX, TensorFlow, Keras, Pandas, Scikit-Learn, OpenAI API, LangChain/LangGraph, OpenCV, DSPy, RAG, HuggingFace (Transformers, Tokenizers, Datasets, Diffusers), vLLM, pgvector, Pydantic, FastAPI, NLTK, spaCy
- **Data Visualization**: Matplotlib, Seaborn, Plotly, Tableau
- **Cloud & DevOps**: AWS, Microsoft Azure, OCI, GCP, GitHub Actions (CI/CD Pipeline), Docker, Kubernetes, Slurm
- **Tools & Databases**: Jupyter, PySpark, Kafka, Git, Linux (Ubuntu), PostgreSQL, MongoDB, Jira, ROS2, Codex, Claude Code

Certifications

- **AWS**: Certified Cloud Practitioner, Certified Machine Learning Specialist, Certified AI Practitioner
- **Oracle (OCI)**: AI Foundations Associate, Generative AI Professional, Data Science Professional, Vector AI Search Professional

Work Experience

AI Engineering Intern — May 2024 — Sep 2024

Atlait Inc. — *Remote*

- Developed a Python-SQL compression script for form data, **reducing storage costs by 7%** for enterprise clients
- **Accelerated mean response time by 96 milliseconds** by integrating PyTorch inference models into Kafka microservices
- Created a $>$ **1TB** RAG-PySpark system, utilizing A/B testing to evaluate and optimize AI-powered search accuracy
- Optimized CI/CD pipeline to **speed up build times by 13%** in an Agile environment, ensuring efficient development cycles

Projects

raceformer — https://bit.ly/raceformer

- Engineered a high-fidelity “Real-to-Sim” validation pipeline processing **30GB of multimodal sensor data** (LiDAR, camera, radar) on 4x A100s, utilizing JAX-based vision-language model to generate ground truth scenarios for critical edge case simulation
- Achieved a **95% pass rate on safety metrics** by leveraging geometric priors to fine-tune RL policies, establishing clear performance baselines and **outperforming standard models by 35%** in collision avoidance testing

dexMCP — https://bit.ly/dexmcp

- Engineered Model Context Protocol (MCP) server exposing **5+ reusable tools** and **5+ Pydantic models**
- Implemented parameter validation across **20+ typed fields** and **100% of tool inputs**
- Built asynchronous clients using DSPy and LangChain to auto-discover tools and execute multi-step requests

DocuMint — https://bit.ly/DocuMint

- Built a 5-agent LangGraph + Gemini API doc-modernizer with Gradio, achieved **90%+ modernization coverage** on sample docs, **cut manual edit time by 50%** with a **4-tab UX**, hardened with **8 deterministic pytest cases** and network-safe skips
- Authored a modular multi-agent system with structured prompts and severity-prioritized research, **lifting modernization accuracy by 35%** and **trimming LLM API spend by 20%**

SuperconGAN — https://bit.ly/3z7JaqZ

- Built a PyTorch-based GAN to create synthetic superconductivity data of various materials, enhancing generative AI applications
- Extracted and processed **80,000+ dataset entries** from the UCI ML Repository using Pandas efficiently
- Released Python package on PyPI, achieving over **80,000 downloads** and widespread adoption