

Ishani Kathuria

AI Engineer • LLM Systems • RAG • Agentic AI • AWS

ishani@kathuria.net | +12193680435 | [LinkedIn/ishani-kathuria](https://www.linkedin.com/in/ishani-kathuria) | [GitHub@ikathuria](https://github.com/ikathuria) | [Portfolio](#)

EDUCATION

Master of Science in Applied Artificial Intelligence, GPA: 4.0/4.0

Purdue University Northwest, Hammond, Indiana, USA

August 2025 – Expected May 2027

- Coursework: Generative AI, Ethical AI, Applied Machine Learning, Data Visualization

Bachelor of Technology in Artificial Intelligence, CGPA: 9.09/10

Amity University, Noida, Uttar Pradesh, India

July 2019 – July 2023

- Coursework: Deep Learning, Computer Vision, Natural Language Processing

SKILLS & CERTIFICATIONS

- **AI:** Generative AI (LLMs, RAG, prompt engineering, multi-agent systems), PyTorch, Tensorflow, HuggingFace, Langchain, Ollama, Scikit-learn
- **Programming & Development:** Python, Golang, TypeScript, MATLAB
- **Cloud/MLOps:** AWS (Bedrock, Lambda, ECS, Step Functions, CloudFormation, CloudWatch)
- **Data Analytics:** Pandas, NumPy, Tableau, Matplotlib, Seaborn
- **Web:** Flask, Django, HTML, CSS, Bootstrap
- **Certifications:** AWS Certified AI Practitioner, AWS Certified Cloud Practitioner

EXPERIENCE

Center for Cybersecurity, Purdue University

AI Research Assistant

September 2025 – Present

- Conducting research on **Retrieval Augmented Generation (RAG)** pipelines, focusing on retrieval quality, hallucination reduction, and latency optimization for LLM-based systems.

Amazon Web Services

Software Development Engineer

March 2025 – July 2025

- Contributed to internal AI-assisted debugging workflows leveraging **LLM-based log summarization and retrieval** techniques, enabling faster root-cause identification of known issues and reducing average debugging time from **2.5 hours to 30 minutes**.
- Built internal tools to simulate customer interactions, enabling proactive bug detection in microservices and **reducing external customer reported issues by 30%**.
- Contributed to internal AI-assisted debugging workflows leveraging LLM-based log summarization and retrieval techniques.
- Designed and deployed AI-powered chatbots for internal tools, guiding operators on command usage and troubleshooting steps; improved **operational efficiency by 40%** in resolving routine queries.

System Development Engineer

July 2023 – March 2025

- Automated region-build processes for **15+ AWS OpenSearch services**, reducing manual intervention by **80%** and cutting deployment timelines by **3 weeks** per launch using Python-based automation.
- Refactored and optimized domain scripts, cutting CPU and memory utilization by greater than **50%**, lowering infrastructure costs by **\$15,000 monthly savings** and improving query performance for end customers.
- Collaborated with cross-functional teams to ensure rigorous pre-launch testing of new features, reducing downtime incidents during regional expansions by **25%**.

Get It Done

Founder

October 2021 – June 2023

- Founded and led a student-focused **AI startup**, mentoring **200 students** and organizing workshops leading to **78%** achieving **placements**.

PUBLICATIONS

- [Conversational AI for Supporting Children with Autism](#) – Springer 2025
- [Temperature Based Food Recommendation using AI](#) – IEEE 2023
- [Deep Learning in Healthcare](#) – Springer 2023
- [PM2.5 Prediction using Azure ML Studio](#) – IEEE 2022

PROJECTS

TrustworthyRAG

- Designed a novel RAG system called QALF (Query-Adaptive Learned Fusion) that dynamically routes queries to optimal modalities (Vector, Graph, Keyword) based on 4D complexity and intent analysis, significantly outperforming standard RRF baselines.
- Integrated **Neo4j** with **LangChain** and **Llama 3** to ingest and link unstructured text, images, and tables into a unified **multimodal knowledge graph**, enabling complex **multi-hop retrieval**.

AutoRedTeam

- Engineered a multi-agent adversarial simulation framework to autonomously stress-test LLM safety and robustness against prompt injection attacks.
- Orchestrated a 3-agent feedback loop (Attacker, Target, Judge) where a "Red Team" model iteratively refines attacks based on "Blue Team" responses, using **Chain-of-Thought** (CoT) reasoning to optimize jailbreak strategies.
- Implemented a provider-agnostic LLM interface supporting Google Gemini (1.5 Pro), OpenAI (GPT-4), and Groq (Llama 3), enabling cross-model benchmarking and latency comparison.
- Developed a deterministic "Judge" agent with structured JSON output enforcement to automatically score defense performance and detect secret key leakage.
- Built a React/TypeScript frontend for real-time visualization of agent thought processes and attack vectors, integrating client-side API calls to eliminate backend overhead.
- Enabled systematic evaluation of LLM safety defenses, simulating real-world red-teaming workflows used in enterprise AI governance.

DeepFakeGuard

- Built an end-to-end AI forensic tool to detect synthetic audio, evolving the architecture from a FastAPI microservice to a fully client-side Edge AI solution.
- Implemented in-browser machine learning using Transformers.js, enabling the application to run complex deep learning models directly in the user's browser without external dependencies.
- Designed the streaming inference logic, handling real-time audio input, feature extraction (spectrogram/waveform analysis), and continuous probability updates.
- Created a forensic dashboard with Recharts to visualize detection confidence levels and temporal anomalies in uploaded audio files.
- Eliminated server-side inference, improving privacy guarantees and reducing inference latency to sub-second performance.

Autbot – Educational Chatbot

- Developed a conversational AI chatbot for children with autism using **DistilRoBERTa & TensorFlow**, improving engagement by **30%** in pilot tests.
- The model is based on **speech and text based emotional recognition** to converse with children and give responses based on how they are feeling.
- Our proposed model achieved over **90% accuracy** in speech based and **73% accuracy** in text based emotional recognition.

LEADERSHIP & VOLUNTEERING

Vice President, SIREN

Volunteer, Amazon Global Volunteers

Chair, IEEE Student Branch

Student Coordinator, Amity Technical Placement Centre

September 2025 – Present

September 2023 – July 2025

May 2022 – January 2023

May 2022 – July 2023