

# Yash Saxena

PHD STUDENT, COMPUTER SCIENCE, UMBC | TRUSTWORTHY AI & LLM ATTRIBUTION

ADVISOR: DR. MANAS GAUR | KAI2 LAB

[✉ ysaxena1@umbc.edu](mailto:ysaxena1@umbc.edu) | [🏡 yashsaxena21.github.io/Portfolio/](https://yashsaxena21.github.io/Portfolio/) | [🔗 YashSaxena21](https://www.linkedin.com/in/yash-saxena-a18b251bb/) | [✉ yash-saxena-a18b251bb](https://scholar.google.com/citations?user=QzXzgAAJAAQ&hl=en) | Scholar

## Research Focus

---

Research on trustworthy Large Language Models (LLMs), focusing on interpretable Retrieval-Augmented Generation (RAG), neurosymbolic retrievers, and evaluation of citation fidelity. My work studies how architectural choices at retrieval and generation time affect faithfulness, robustness, and verifiability of LLM outputs.

## Skills

---

**Research Areas** Trustworthy AI, Information Retrieval, RAG, LLM Attribution, Neurosymbolic AI

**Methods** Preference Optimization (DPO, PPO), Retrieval Modeling, Evaluation Design

**Systems** PyTorch, Hugging Face, Slurm, FAISS

## Education

---

### University of Maryland, Baltimore County

Baltimore, Maryland, USA

PHD, COMPUTER SCIENCE

Jan 2025 - Present

- Research Assistant | GPA: 4.0/4.0
- Advisor: Dr. Manas Gaur

### Galgotias University

Greater Noida, Uttar Pradesh, India

B.TECH, COMPUTER SCIENCE AND ENGINEERING WITH SPECIALIZATION IN ARTIFICIAL  
INTELLIGENCE AND MACHINE LEARNING

Nov 2020 - May 2024

- CGPA: 8.75/10

## Professional Experience

---

### Knowledge-Infused AI & Inference (KAI2) Lab, University of Maryland, Baltimore

Baltimore, MD, USA

County

PHD RESEARCH ASSISTANT, ADVISED BY DR. MANAS GAUR

Jan 2025 – Present

- **Project: Interpretable Dense Retrieval via Embedding Modulation (IMRNNs).** Developed interpretable dense retrieval methods via embedding-space modulation for transparent and stronger information retrieval.
- **Project: LLM Attribution and Citation Fidelity.** Studied **generation-time vs post-hoc citation** and evaluated how attribution mechanisms affect faithfulness and verifiability.
- **Project: Ranking-Free RAG for Sensitive Domains.** Investigating evidence **selection** as a replacement for reranking to improve robustness in RAG pipelines.
- **Outcomes: EACL 2026 (IMRNNs); IEEE Intelligent Systems 2026 (Neurosymbolic Retrievers for RAG); NeurIPS 2025 LLM Evaluation Workshop (LLM Attribution); ICLR 2026 (Ranking Free RAG, under review).**

**Knowledge-Infused AI & Inference (KAI2) Lab, University of Maryland, Baltimore County**

Baltimore, MD, USA

**REMOTE RESEARCH INTERN, ADVISED BY DR. MANAS GAUR**

Aug 2024 – Jan 2025

- **Project: Trustworthy RAG with Preference-Tuned Rationale Generation.** Fine-tuned **LLaMA-3.1 70B** with **Direct Preference Optimization (DPO)** and integrated it into a retrieval-augmented generation (RAG) pipeline for **evidence-grounded answering**.
- Designed and executed experiments (ablations, evaluation protocols) to validate model behavior and performance; supported data and evaluation design for a benchmark effort in the lab.
- **Outcome:** Contributed to the development of the **REASONS** benchmark and associated experimental results (manuscript in progress / internal).

**Stareout Games**

Remote

**AI ENGINEER INTERN**

Jan 2024 – Mar 2024

- **Project: LLM + Image Generation Pipeline for Game Prototyping.** Built an end-to-end workflow combining large language models and image generation models to accelerate game concept creation and iteration.
- Fine-tuned language and image generation models; developed lightweight tooling and demos using **LangChain**, **Streamlit**, and supporting Python libraries.
- **Outcome:** Delivered an internal prototype pipeline used for rapid idea-to-asset generation and iteration.

**Artificial Intelligence Institute, University of South Carolina (AIISC)**

Columbia, SC, USA

**REMOTE RESEARCH INTERN, ADVISED BY DR. AMIT SHETH**

Aug 2023 – Jul 2024

- **Project: LLM Applications and Dataset Construction.** Developed LLM-based components using **LangChain** and **Llamaindex** for research prototypes and experiments.
- Built a web scraper to construct the primary dataset used in ongoing research; supported data collection, cleaning, and organization.
- **Outcome:** Enabled downstream modeling experiments by delivering a structured dataset and working LLM prototypes.

**Celebal Technologies**

Jaipur, Rajasthan, India

**REMOTE DATA SCIENCE INTERN**

May 2023 – Jul 2023

- **Project: NLP Feature Engineering for Client Use Cases.** Worked with large datasets, performed preprocessing, and extracted task-relevant features for client-facing applications.
- Built NLP components using **spaCy** and related tooling to support production-facing pipelines.
- **Outcome:** Delivered processed datasets and reusable NLP components for downstream application development.

**HCLTech**

Noida, Uttar Pradesh, India

**INTERN**

Nov 2022 – Jan 2023

- **Project: Resume Parsing and Filtering Web App.** Developed a Python web app using **spaCy** and **Streamlit** to extract and structure information from resumes.
- Implemented resume filtering features to match user-defined requirements and streamlined candidate shortlisting.
- **Outcome:** Delivered an end-to-end prototype that automated resume parsing and rule-based filtering.

**Publications**

---

**PUBLISHED/ACCEPTED**

**Yash Saxena**, Ankur Padia, Kalpa Gunaratna, Manas Gaur. “IMRNNs: An Efficient Method for Interpretable Dense Retrieval via Embedding Modulation”. **Accepted EACL 2026**. (Acceptance Rate: 16.10%)

**Yash Saxena**, Manas Gaur. “Neurosymbolic Retrievers for Retrieval-Augmented Generation”. **Accepted IEEE Intelligent Systems 2026**. (Acceptance Rate: ~ 23.00%)

Deepa Tilwani, **Yash Saxena**, Ankur Padia, Srinivasan Parthasarathy, Manas Gaur. “Neurosymbolic AI for Legal AI-TRISM: Trustworthy, Reliable, Interpretable, Safe Models”. Book chapter **Neuro-symbolic AI: Foundations and Applications**.

**Yash Saxena**, Raviteja Bommireddy, Ankur Padia, Manas Gaur. “Generation-Time vs. Post-hoc Citation: A Holistic Evaluation of LLM Attribution”. Accepted at **NeurIPS 2025 LLM Evaluation Workshop**. [Link](#)

Seyedreza Mohseni, Seyedali Mohammadi, Deepa Tilwani, **Yash Saxena**, Gerald Ketu Ndawula, Sriram Vema, Edward Raff, Manas Gaur. “Can LLMs Obfuscate Code? A Systematic Analysis of Large Language Models into Assembly Code Obfuscation”. **Proceedings of the AAAI Conference on Artificial Intelligence 2025**, 39(23), 24893-24901. (Acceptance Rate: 23.40%). [Link](#)

**Yash Saxena**, Sarthak Chopra, Arunendra Mani Tripathi. “Evaluating Consistency and Reasoning Capabilities of Large Language Models”. **2024 Second International Conference on Data Science and Information System (ICDSIS)**, Hassan, India, 2024, pp. 1–5. [Link](#)

**Yash Saxena**, Aman Kumar Mishra, Daksh Arora, Runumi Devi. 2023. “Emotion Based Mental Health Classifier for NCR Based Engineering Students”. IEEE 6th **International Conference on Contemporary Computing and Informatics (IC3I)**, Page(s):285-290. [Link](#)

## UNDER REVIEW

**Saxena et al.** “Ranking Free RAG: Replacing Reranking with Selection in RAG for Sensitive Domains”. (Communicated to **ICLR 2026**) [Link](#)

## IN PREP

**Saxena et al.** “Attribution in Scientific Literature: New Benchmark and Methods”.

## Presentations

---

**Yash Saxena**, Ankur Padia, Swati Padhee, Manas Gaur and Srinivasan Parthasarathy. June 2025. Title: “RASOR: Contextual Legal Intelligence via Rationalized Selection and Refinement in RAG”. Venue: **Bloomberg Law, Language, and AI Symposium**.

**Yash Saxena** and Manas Gaur. October 2025. Title: “Building Trustworthy LLM Agents for Academia through Structured, Interpretable Knowledge Retrieval and Source Attribution”. Venue: **UMBC Library 2025 AI Symposium**.

## Recognitions, Fellowships, & Grants

---

### RECOGNITIONS

- 2022 Winner UNESCO India–Africa Hackathon, Ministry of Education (India).
- 2022 Winner Smart India Hackathon (Team Lead), Ministry of Education (India).

### SERVICES

- 2026 Served as a reviewer, AAAI
- 2025 Served as a reviewer, NeurIPS 2025 LLM Evaluation Workshop
- 2024 Served as a reviewer, ACM transactions on computing for healthcare journal