

Yash Saxena

PHD STUDENT, COMPUTER SCIENCE, UMBC | LLM RESEARCH, INFORMATION RETRIEVAL & TRUSTWORTHY AI
ADVISOR: DR. MANAS GAUR | KAI² LAB

✉ y.saxena1@umbc.edu | 🏠 yashsaxena21.github.io/Portfolio/ | 📧 YashSaxena21 | 🔗 LinkedIn | 📄 Google Scholar

Research Focus

I study trustworthy Retrieval-Augmented Generation (RAG) through **Interpretable Knowledge Flow (IKnowFlow)**, exposing how retrieved sources support generated claims across high-stakes domains such as law, finance, and scientific research.

Skills

Research Areas	Large Language Models, Information Retrieval, Trustworthy AI, Neurosymbolic AI, Scientific Discovery, Natural Language Processing
Methods	Preference Optimization (DPO, PPO), Retrieval Modeling, Knowledge Graphs, Evaluation Design
Systems	Python, PyTorch, Hugging Face, FAISS, Slurm, LangChain, LlamaIndex

Education

University of Maryland, Baltimore County

Baltimore, Maryland, USA

PHD, COMPUTER SCIENCE

Jan 2025 – Jan 2028

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

Professional Experience

TIB, Leibniz Information Centre for Science and Technology

Hannover, Germany

VISITING RESEARCH INTERN, HOSTED BY DR. SÖREN AUER

May 2026 – Aug 2026

- Built **Dynamic Cognitive Knowledge Graphs (DCKG)** using structured scholarly knowledge from the **Open Research Knowledge Graph (ORKG)** to support **TIB Assistant** research-lifecycle workflows.
- Represented papers through fine-grained entities and AI-agent-generated inter-paper relations for interpretable literature analysis, research-gap discovery, hypothesis generation, and experiment planning.

Knowledge-Infused AI & Inference (KAI2) Lab, UMBC

Baltimore, MD, USA

PHD RESEARCH ASSISTANT, ADVISED BY DR. MANAS GAUR

Jan 2025 – May 2026

- Proposed **IKnowFlow**, a doctoral framework that operationalizes trustworthy RAG through traceability, interpretable retrieval, grounding, and provenance.
- Developed **IMRNNs**, improving dense-retrieval recall by **+7.14% across seven benchmarks** while exposing human-readable matching signals; released an **open-source Python package** with pretrained assets and reproducible training and evaluation workflows.
- Developed **METEORA** for rationale-driven evidence selection in sensitive-domain RAG, improving answer accuracy by **+33.34%** with **50% fewer evidence chunks**.
- Studied LLM citation fidelity through **REASONS** and generation-time vs post-hoc citation evaluation; post-hoc citation reached **78%** human-evaluated correctness vs **69%** for generation-time citation.

Knowledge-Infused AI & Inference (KAI2) Lab, UMBC

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 RAG pipeline for **evidence-grounded answering**.
- Designed ablations and evaluation protocols and contributed to the **REASONS Dataset**, accepted at **IEEE DSAA 2026**; part of IARPA's Rapid Explanation, Analysis, and Sourcing ONLINE System project.

Artificial Intelligence Institute, University of South Carolina (AIISC)

Columbia, SC, USA

REMOTE RESEARCH INTERN, ADVISED BY DR. AMIT SHETH

Aug 2023 – Jul 2024

- Developed LLM research prototypes with **LangChain** and **LlamaIndex** and built web-scraping pipelines that produced structured datasets for downstream modeling experiments.

Publications

- Yash Saxena**, Ankur Padia, Mandar Chaudhary, Kalpa Gunaratna, Srinivasan Parthasarathy, Manas Gaur. “Ranking Free RAG: Replacing Re-ranking with Selection in RAG for Sensitive Domains”. **ICML 2026**. (Acceptance Rate: 26.6%). [Link](#)
- Yash Saxena**, Ankur Padia, Kalpa Gunaratna, Manas Gaur. “IMRNNs: An Efficient Method for Interpretable Dense Retrieval via Embedding Modulation”. **EACL 2026**. (Acceptance Rate: 16.1%). [Link](#)
- Deepa Tilwani*, **Yash Saxena***, Seyedali Mohammadi, Ankur Padia, Edward Raff, Amit Sheth, Srinivasan Parthasarathy, Manas Gaur. “Abstention vs. Hallucination: Benchmarking LLM Source Attribution for Sentence-level Scientific Citations”. **IEEE DSAA 2026, Dataset & Benchmark Track (Oral)**. (Acceptance Rate: 24%). ***Equal contribution**.
- Yash Saxena**, Manas Gaur. “Neurosymbolic Retrievers for Retrieval-Augmented Generation”. **IEEE Intelligent Systems 2026**. (Acceptance Rate: 23%). [Link](#)
- Yash Saxena**. “IKnowFlow: Trustworthy RAG for Sensitive Domains”. **IJCAI-ECAI 2026 Doctoral Consortium**. [Link](#)
- Syedreza 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: 23.4%). [Link](#)
- Yash Saxena**, Raviteja Bommireddy, Ankur Padia, Manas Gaur. “Generation-Time vs. Post-hoc Citation: A Holistic Evaluation of LLM Attribution”. **NeurIPS 2025 LLM Evaluation Workshop**. [Link](#)
- Deepa Tilwani, **Yash Saxena**, Ankur Padia, Srinivasan Parthasarathy, Manas Gaur. “Neurosymbolic AI for Legal AI-TRISM: Trustworthy, Reliable, Interpretable, Safe Models”. Book chapter in **Neuro-symbolic AI: Foundations and Applications**. [Link](#)
- Yash Saxena**, Gollam Rabby, Sören Auer, Manas Gaur. “Dynamic Cognitive Knowledge Graph for Scholarly Research: Benchmark, Method, and Evaluation Protocol”. **In prep**.

Talks and Presentations

- August 2026. Doctoral Consortium presentation: “IKnowFlow: Trustworthy RAG for Sensitive Domains”. **IJCAI-ECAI 2026 Doctoral Consortium, Bremen, Germany**.
- July 2026. Poster: “Ranking Free RAG: Replacing Re-ranking with Selection in RAG for Sensitive Domains (METEORA)”. **ICML 2026, Seoul, South Korea**.
- January 2026. Virtual poster: “IMRNNs: An Efficient Method for Interpretable Dense Retrieval via Embedding Modulation”. **EACL 2026, Rabat, Morocco**.
- December 2025. Poster: “Generation-Time vs. Post-hoc Citation: A Holistic Evaluation of LLM Attribution”. **NeurIPS 2025 LLM Evaluation Workshop, San Diego, CA, USA**.
- October 2025. Talk: “Building Trustworthy LLM Agents for Academia through Structured, Interpretable Knowledge Retrieval and Source Attribution”. **UMBC Library 2025 AI Symposium, Virtual**.
- June 2025. Talk: “RASOR: Contextual Legal Intelligence via Rationalized Selection & Refinement in RAG”. **Bloomberg Law, Language, and AI Symposium, New York, NY, USA**.

Mentoring

- Aug–Nov 2025. Mentored **Raviteja Bommireddy** (B.Tech CSE, IIITDM Kancheepuram, India) on LLM attribution and citation evaluation; co-authored a NeurIPS 2025 workshop paper.
- May 2026–Present. Mentoring **Shaunjit Chakrabarti** (B.Tech CSE, Manipal Institute of Technology Jaipur, India) on AI agents for legal-case simulation and reasoning.

Honors & Service

HONORS & MEDIA

- 2026 **Summer research scholarship**, TIB.
- 2026 **Travel grant recipient**, SIGAI-supported IJCAI-AIJ Doctoral Consortium.
- 2026 **Featured interview**, “[How is external knowledge used in AI systems?](#)”, Alhub.
- 2022 **Hackathon winner**, UNESCO India–Africa and Smart India Hackathons.

SERVICE

- 2024–26 **Reviewer**, AAAI, ACL, NeurIPS LLM Evaluation Workshop, and ACM TCH.