Yash Saxena

PhD Student, Computer Science, UMBC | Trustworthy AI, RAG & Neurosymbolic AI
Advisor: Dr. Manas Gaur | KAI^2 Lab

ysaxena1@umbc.edu | yashsaxena21.github.io/Portfolio/ | YashSaxena21 | yash-saxena-a18b251bb | G Scholar

Research Focus

I study trustworthy RAG through **Interpretable Knowledge Flow (IKnowFlow)**, making the evidence trail from retrieved sources to generated claims explicit across high-stakes domains such as law, finance, and academia.

Skills

Research Areas	Grounded Generation, Information Retrieval, Neurosymbolic AI, Trustworthy AI, NLP
Methods	Preference Optimization (DPO, PPO), Retrieval Modeling, Knowledge Graphs, Evaluation Design
Systems	PyTorch, Hugging Face, FAISS, Slurm, LangChain, LlamaIndex

Education

University of Maryland, Baltimore County *Baltimore, Maryland, USA*

PhD, Computer Science *Jan 2025 - Present*

- Research Assistant | 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 Prof. Sören Auer *May 2026 – Present*

- Building **Dynamic Cognitive Knowledge Graphs (DCKG)** on **ORKG** structured scholarly knowledge to support **TIB AIssistant** research-lifecycle workflows.
- Representing 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**, lightweight embedding-modulation adapters for dense retrievers, improving recall by **+7.14% across seven benchmarks** while exposing human-readable matching signals.
- 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 and executed ablations and evaluation protocols to validate model behavior and support attribution-benchmark development.
- **Outcome:** Contributed to the development of the **REASONS** benchmark and associated experimental results (manuscript in progress / internal). This was a 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*

- **Project: LLM Applications and Dataset Construction.** Developed LLM-based components using **LangChain** and **LlamaIndex** for research prototypes and experiments.
- Built web-scraping workflows to construct structured datasets for downstream modeling experiments.
- **Outcome:** Enabled downstream modeling experiments by delivering a structured dataset and working LLM prototypes.

Publications

Published/Accepted

Yash Saxena. "IKnowFlow: Trustworthy RAG for Sensitive Domains". **IJCAI-ECAI 2026 Doctoral Consortium**, Bremen, Germany. Travel grant awarded.

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**. 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

Yash Saxena, Manas Gaur. "Neurosymbolic Retrievers for Retrieval-Augmented Generation". **IEEE Intelligent Systems 2026.** (Acceptance Rate: 23%). Link

Deepa Tilwani, **Yash Saxena**, Ankur Padia, Srinivasan Parthasarthy, Manas Gaur. "Neurosymbolic AI for Legal AI-TRISM: Trustworthy, Reliable, Interpretable, Safe Models". Book chapter **Neuro-symbolic AI: Foundations and Applications.** edited by Prof. Pradeep Ravikumar (CMU) and Alvaro Velasquez (CU/DARPA). Link

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. Link

Yash Saxena, Sarthak Chopra, Arunendra Mani Tripathi. "Evaluating Consistency and Reasoning Capabilities of Large Language Models". IEEE 2nd **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

Deepa Tilwani, Yash Saxena et al. "REASONS: A Benchmark for Retrieval and Automated Citations of Scientific Sentences using Public and Proprietary LLMs". (Communicated to **KDD 2026**). Benchmark spans **20,000 scientific papers** across **12 domains**.

Presentations

July 2026. Upcoming poster: "Ranking Free RAG: Replacing Re-ranking with Selection in RAG for Sensitive Domains (METEORA)". **ICML 2026.**

January 2026. Virtual poster: "IMRNNs: An Efficient Method for Interpretable Dense Retrieval via Embedding Modulation". **EACL 2026.**

December 2025. Poster: "Generation-Time vs. Post-hoc Citation: A Holistic Evaluation of LLM Attribution". **NeurIPS 2025 LLM Evaluation Workshop.**

October 2025. Talk: “Building Trustworthy LLM Agents for Academia through Structured, Interpretable Knowledge Retrieval and Source Attribution”. **UMBC Library 2025 AI Symposium.**

June 2025. Talk: “RASOR: Contextual Legal Intelligence via Rationalized Selection & Refinement in RAG”. **Bloomberg Law, Language, and AI Symposium.**

Mentoring

Aug–Nov 2025. Mentored **Raviteja Bommireddy** (B.Tech CSE, IIITDM Kancheepuram) 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) on AI agents for legal-case simulation and reasoning.

Recognitions, Fellowships, & Grants

Recognitions

2026 **Summer research scholarship**, TIB.

2026 **Travel grant recipient**, IJCAI-ECAI 2026.

2022 **Winner, UNESCO India–Africa Hackathon**, Ministry of Education (India).

2022 **Winner and team lead, Smart India Hackathon**, Ministry of Education (India).

Services

2026 **Served as a reviewer**, AAAI, ACL

2025 **Served as a reviewer**, NeurIPS 2025 LLM Evaluation Workshop

2024 **Served as a reviewer**, ACM Transactions on Computing for Healthcare