Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Doctoral Consortium Track

IKnowFlow: Trustworthy RAG for Sensitive Domains

Yash Saxena
KAI^2 Lab, University of Maryland, Baltimore County (UMBC)
ysaxena1@umbc.edu

Abstract

RAG systems ground LLM responses in external evidence, yet their trustworthiness remains underspecified. Retrieval, re-ranking, and generation are optimized independently with no unified account of whether the system is interpretable, grounded, and traceable. My thesis operationalizes **trustworthy RAG** through three sub-goals, each at the stage where it is most vulnerable. For traceability, REASONS benchmarks LLMs on 20K papers, showing attribution failures are systematic even under retrieval augmentation. For interpretability, IMRNNs decomposes dense embeddings via learned adapters, exposing human-readable query-document matching at retrieval. For grounding, METEORA replaces opaque re-ranking with rationale-driven selection, ensuring generation operates over explicitly justified context. I unify these under Interpretable Knowledge Flow (**IKnowFlow**), where trustworthiness is preserved across stage boundaries rather than checked at output. **IKnowFlow** will next address two longstanding blind spots: negation-awareness in retrieval, where surface similarity masks semantic inversion, and provenance-aware verification through structured source graphs.

1 Introduction

Vision. This doctoral research envisions Retrieval-Augmented Generation (RAG) architectures for sensitive domains, where each generated claim carries an auditable trail of its retrieval origin, selection rationale, and attribution evidence, enabling domain experts to verify, contest, or refine system outputs before they reach end users.

Goal. The goal of this doctoral research is to formalize trustworthiness in RAG as a system-level property composed of four verifiable dimensions: traceability, interpretability, grounding, and provenance. Each dimension is studied at the RAG stage, where it is most at risk, and the resulting methods are unified under **IKnowFlow**.

In 2025, a preregistered study found that leading RAG-based legal tools hallucinate between 17% and 33% of the time, fabricating citations even when grounded in curated databases [Magesh *et al.*, 2025]. In healthcare, RAG systems produce unsupported claims by blending retrieved passages with parametric priors over contradictory evidence [Kim *et al.*, 2025]. These are structural failures: RAG was designed for factual grounding but never engineered to be trustworthy.

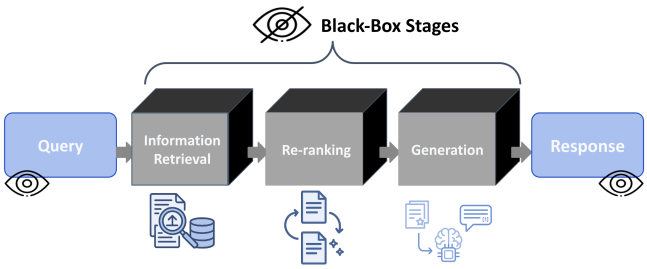

Figure 1: Traditional Black-box RAG system illustrating how unobservable retrieval, re-ranking, and generation stages limit system-level trustworthiness.

RAG [Lewis *et al.*, 2020] reduces confabulation by grounding Large Language Model (LLM) responses in external evidence but does not guarantee trustworthy generation. Three bottlenecks persist (see Figure 1). Dense retrievers match on embedding similarity without exposing *why* a document is selected. Re-rankers optimize scores without generating justifications. Citation mechanisms verify surface correspondence without tracking whether claims *causally descend* from retrieved evidence. This research addresses these bottlenecks through **IKnowFlow**, operationalizing trustworthiness as traceability, interpretability, grounding, and provenance. The following four questions guide this thesis:

1. **RQ1 (Traceability):** How can source-attribution failures be diagnosed across model families and domains?
2. **RQ2 (Interpretability):** How can dense retrievers expose interpretable query-document matching signals?
3. **RQ3 (Grounding):** How can RAG systems select explicitly justified evidence to keep generation grounded in relevant and verifiable source context?
4. **RQ4 (Provenance):** How should attribution preserve provenance structure, capturing how evidence originates, evolves, and is challenged over time?

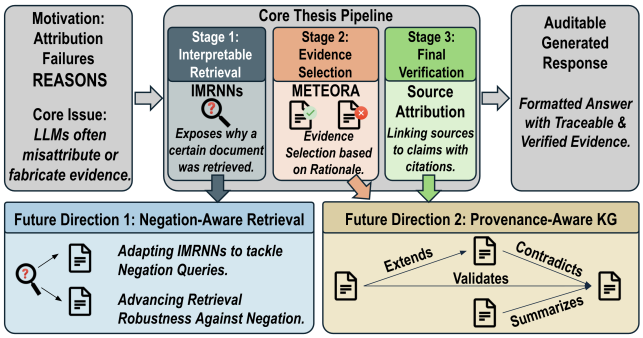

Figure 2: Interpretable Knowledge Flow (`IKnowFlow`)

2 Current Progress and Future Work

My current research addresses RQ1 and RQ2. To diagnose traceability failures (RQ1), REASONS [Saxena *et al.*, 2025b] benchmarks LLMs on 20K scientific papers across 12 domains, revealing that even state-of-the-art models with retrieval augmentation fabricate citations in scientific text, meaning the attribution problem propagates through reranking and generation. To address interpretability (RQ2), IMRNNs [Saxena *et al.*, 2026b] augment dense retriever with two learned adapters: one conditions document embeddings on the query, the other refines query embeddings using corpus-level feedback. This bidirectional modulation exposes human-readable keyword-level matching signals, yielding +7.14% recall across seven benchmarks. For sensitive applications, IMRNNs gives domain experts the ability to audit *why* a specific document was retrieved; a prerequisite absent from most existing dense retrievers.

To address RQ3, METEORA [Saxena *et al.*, 2026a] replaces score-based re-ranking with rationale-driven selection, where a preference-tuned LLM generates query-conditioned rationales, an evidence chunk selection engine pairs these with passages using adaptive cutoffs, and a verifier filters poisoned content before generation. Across six datasets spanning scientific, legal, and finance, METEORA achieves 33.34% higher answer accuracy using 50% fewer chunks, and raises adversarial defense F1 from 0.10 to 0.44. Extending grounding further, my work on Neurosymbolic Retrievers [Saxena and Gaur, 2026] integrates symbolic reasoning with neural retrieval. Evaluations with licensed psychologists show that neurosymbolic grounding improves clinical relevance in mental health risk assessment over purely neural approaches. Together, these contributions make RAG auditable in domains where unexplained evidence selection invites legal liability and clinical risk.

The completed work substantially advances RQ1, RQ2, and RQ3, with RQ3 continuing to evolve through ongoing neurosymbolic extensions (see Figure 2). My comparison of generation-time and post-hoc citation [Saxena *et al.*, 2025a], showing post-hoc verification achieves 78% versus 69% human-evaluated correctness, begins addressing RQ4. Two directions remain. First, I will study negation-aware retrieval as a safety-critical blind spot in dense retrievers. In clinical settings, confusing "patient denies suicidal ideation" with "patient reports suicidal ideation" is not just a retrieval error; it can change the meaning of the evidence entirely. I will examine whether IMRNNs can be adapted to capture such negation at the embedding level in an interpretable manner. Second, provenance-aware verification (RQ4) will construct a claim-level ontology with typed relations such as *extends*, *validates*, *contradicts*, and *summarizes*, enabling verification over dynamic source graphs. In law, healthcare, and science, knowing that a claim was later contradicted is as important as knowing it was ever stated.

References

[Kim *et al.*, 2025] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir Tulebaev, and Cynthia Breazeal. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, 2025.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[Magesh *et al.*, 2025] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.

[Saxena and Gaur, 2026] Yash Saxena and Manas Gaur. Neurosymbolic Retrievers for Retrieval-Augmented Generation . *IEEE Intelligent Systems*, 41(01):96–104, January 2026.

[Saxena *et al.*, 2025a] Yash Saxena, Raviteja Bommireddy, Ankur Padia, and Manas Gaur. Generation-time vs. post-hoc citation: A holistic evaluation of llm attribution, 2025.

[Saxena *et al.*, 2025b] Yash Saxena, Deepa Tilwani, Ali Mohammadi, Edward Raff, Amit Sheth, Srinivasan Parthasarathy, and Manas Gaur. Attribution in scientific literature: New benchmark and methods, 2025.

[Saxena *et al.*, 2026a] Yash Saxena, Ankur Padia, Mandar S Chaudhary, Kalpa Gunaratna, Srinivasan Parthasarathy, and Manas Gaur. Ranking free rag: Replacing re-ranking with selection in rag for sensitive domains, 2026.

[Saxena *et al.*, 2026b] Yash Saxena, Ankur Padia, Kalpa Gunaratna, and Manas Gaur. IMRNNs: An efficient method for interpretable dense retrieval via embedding modulation. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6324–6337, Rabat, Morocco, March 2026. Association for Computational Linguistics.