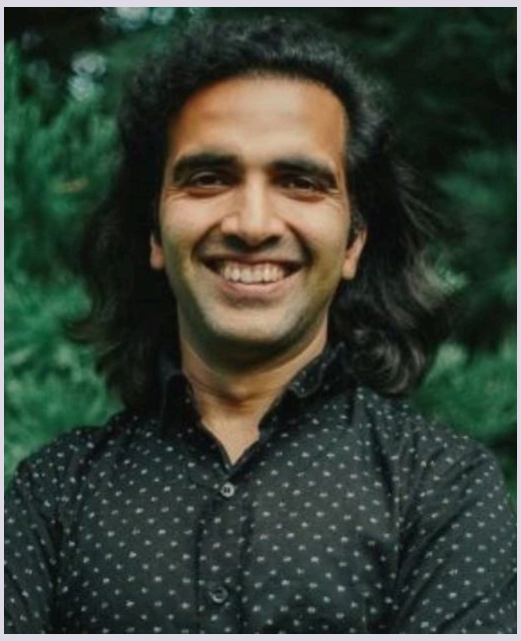
OG-RAG: Ontology-Grounded Retrieval-Augmented Generation using Knowledge Hypergraphs For Large Language Models

BIOGRAPHY



Dr. Peeyush Kumar is a research scientist at Microsoft Research with extensive experience in building large-scale intelligent systems and conducting applied research across disciplines, including large language models (LLMs), energy, food and agriculture, and healthcare. Peeyush's research philosophy combines people, processes, and technology to create regenerative and reproducible solutions. He is an entrepreneur and an artificial intelligence scientist by background. Peeyush has authored 15 patents and over 30 peer-reviewed journal articles, with his work featured in major media outlets such as CNN, Techmonitor, IEEE Spectrum, and GeekWire. Prior to his role at Microsoft, Dr. Kumar founded Engooden Health, a healthcare AI and services company that he scaled to a market value of ~$80 million. Additionally, he serves on the board of Sustainable Seattle, leveraging his expertise in grassroots community development and strategic programming.

ABSTRACT

This seminar introduces OG-RAG, an Ontology-Grounded Retrieval Augmented Generation framework that enhances large language models (LLMs) by grounding retrieval in structured domain knowledge. While LLMs have revolutionized tasks like search and question answering, they falter in specialized domains without costly fine-tuning or naive retrieval techniques. OG-RAG bridges this gap by integrating domain-specific ontologies into the retrieval process. It constructs a hypergraph of factual clusters, grounded in ontological structures, and uses an optimization algorithm to retrieve the minimal, most relevant set of facts. This enables precise, context-rich generation while preserving complex entity relationships. Evaluated across domains such as healthcare, law, agriculture, and investigative research, OG-RAG significantly improves fact recall, response accuracy, and reasoning quality over traditional RAG models—boosting fact recall by 55% and correctness by 40%. Join us to explore how structured knowledge can unlock smarter, more reliable LLM outputs for critical knowledge-driven applications.