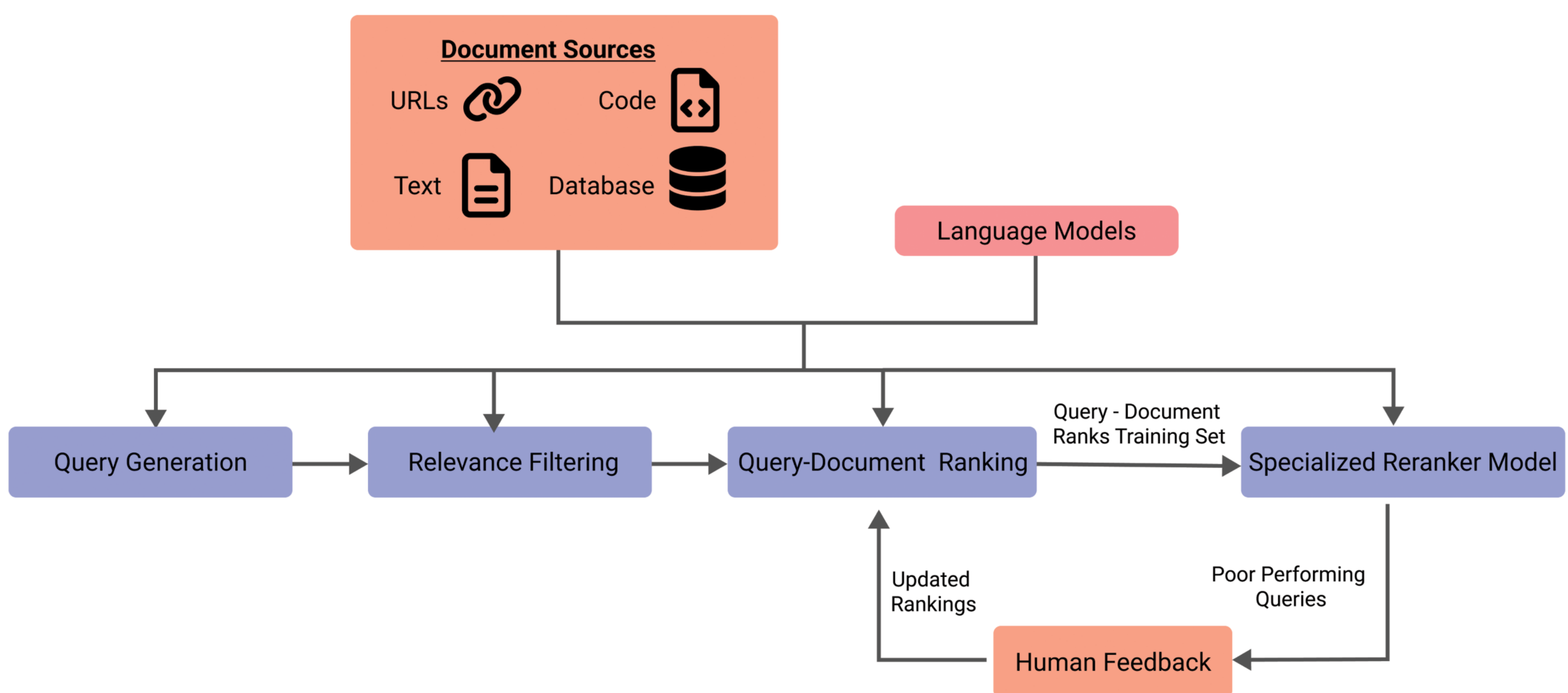# ARIA
APPLIED RESEARCH IN ACTION

Enhancing Document Retrieval with Large Language Models and Human Feedback

Using generative AI to make retrieval with human feedback more efficient

Ilan Gofman

Scott Sanner
ACADEMIC SUPERVISOR

Paul (Jiapeng) Wu, Raunaq Suri, and Guangwei Yu
INDUSTRY SUPERVISORS

PROJECT SUMMARY

At one of Canada's largest consumer banks, with over a century's worth of documents, we leverage the growing capabilities of Large Language Models (LLMs) to improve information retrieval. Given the expensive nature of incorporating human feedback for training document ranking models, we introduce a novel human-in-the-loop framework using LLMs to significantly reduce the need for human intervention. By employing LLMs, we automate the data generation pipeline: query generation, relevance, and ranking, effectively reducing the human workload and improving retrieval. While our approach integrates a human feedback loop to handle uncertainties, the primary benefit lies in the enhanced efficiency brought about by the LLMs themselves. Recognizing the challenges of managing large amounts of sensitive materials in a regulated environment, our solution incorporates the necessary governance and control beyond the limitations of generic, off-the-shelf products. By combining LLM capabilities with expert feedback, this research opens doors to a more efficient and context-aware retrieval system, with potential applications in Retrieval Augmented Generation (RAG) and specialized chatbot development.

layer6
AI at TD

Computer Science
UNIVERSITY OF TORONTO | Master of Science in Applied Computing