

 CANLEGALRAGBENCH:
Evaluating Retrieval-Augmented Generation on Canadian Case Law

Ethan Zhao^{1,2} Maksym Taranukhin^{1,2} Wei Cui⁴ Moira Aikenhead⁴ Vered Shwartz^{1,2,3}

¹ Department of Computer Science, University of British Columbia ² Vector Institute

³ CIFAR AI Chair ⁴ Peter A. Allard School of Law, University of British Columbia

{ethanz01, maksymt, vshwartz}@cs.ubc.ca

Abstract

RAG-based legal assistants have been growing in popularity, but LLM hallucinations remain a key issue and potentially undermines justice. While benchmarks have been developed to evaluate progress, many rely on synthetic queries rather than realistic legal scenarios. Moreover, Canadian law remains underrepresented in existing evaluations. To address this gap, we introduce CANLEGALRAGBENCH, a Canadian legal QA benchmark based on realistic queries and expert-annotated answers grounded in case law. Our evaluation shows that retrieval performance is sensitive to design choices and that open-source embedding models are competitive with closed source models. However, it also reveals the limitation of automatic evaluations that penalize systems for retrieving alternative relevant documents. We also find that generated answers often diverge from gold responses, either with hallucinations or by producing overly detailed or irrelevant content, with 8-29% of claims not being supported by the retrieved documents. We hope this benchmark will help drive continued progress in addressing limitations of legal RAG systems. Code:  <https://github.com/NLP-UBC/CanLegalRAGBench>.