

# Recall Is Not Completeness: Retrieving All the Evidence a Question Needs

**Sebastien Kawada**  
MIT CSAIL  
kawada@mit.edu

**Adhiban Arulselvan**  
MIT CSAIL  
adhiban@mit.edu

**Manolis Kellis**  
MIT CSAIL  
manoli@mit.edu

## Abstract

Questions in law, policy, the life sciences, and the humanities often require evidence from many sources, all of which retrieval must find. We prove that recall can be high even when few questions have all of their gold items retrieved, that is, are complete. We demonstrate that completeness strongly affects answer correctness. On HotpotQA questions where Quorum, our learned fusion of diverse retrievers, is the only retriever to retrieve all of the supporting paragraphs, our reader’s exact match rises from 0.212 to 0.553. We find that recall hides poor completeness. On Housing Statute QA questions that cite four or more statutes, Quorum’s top-ten results include an average of 39.3% of the gold items, yet are complete for only 2.3% of the questions. Recall exceeds completeness by 17- to 81-fold on this subset for all seven systems we evaluate. The candidate pools contain all of the gold items for 65% of these questions, meaning that the reranker causes more failures of completeness than the candidate retrieval. Quorum significantly outperforms the baseline, improving completeness@5 over dense retrieval from 0.453 to 0.670 on MuSiQue and accuracy on the agent-memory benchmark LongMemEval-V2 from 0.385 to 0.530. Benchmarks where multiple items must be retrieved should report completeness conditioned on the number of gold items.