



MoR: Better Handling Diverse Queries with a Mixture of Sparse, Dense, and Human Retrievers

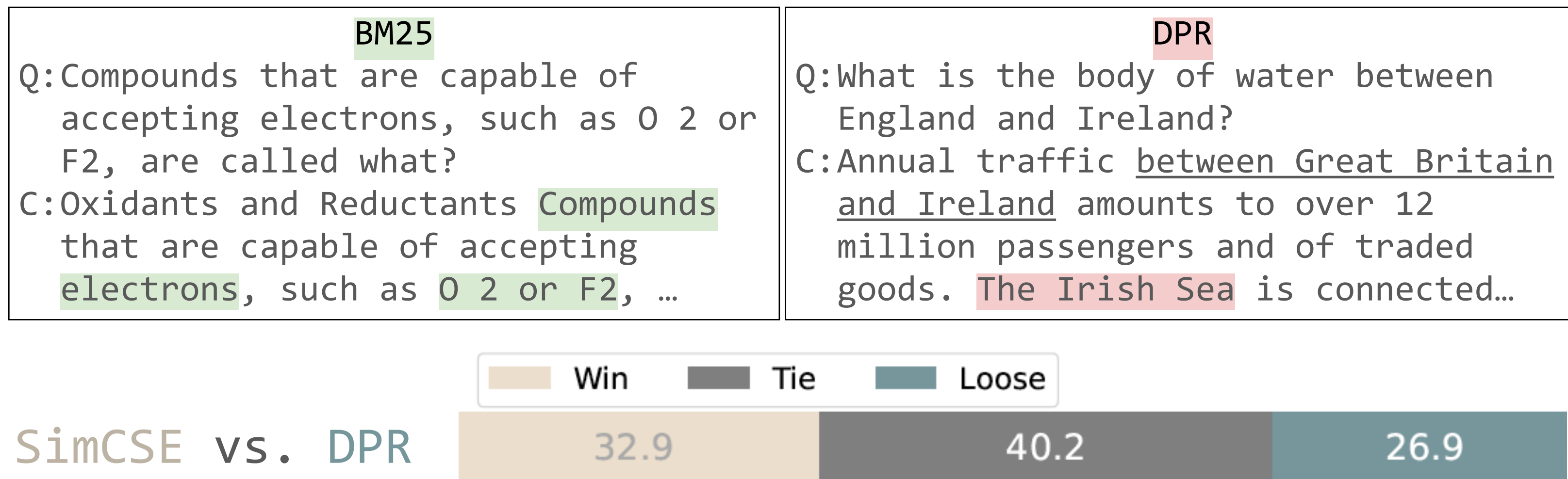
Jushaan Singh Kalra*, Xinran Zhao*, To Eun Kim, Fengyu Cai, Fernando Diaz, Tongshuang Wu

Goal: Dynamically select and integrate the best retrievers for each query

Motivation:

Different retrievers are good at different kinds of queries

Select retrievers that wins for each query to always "win"!



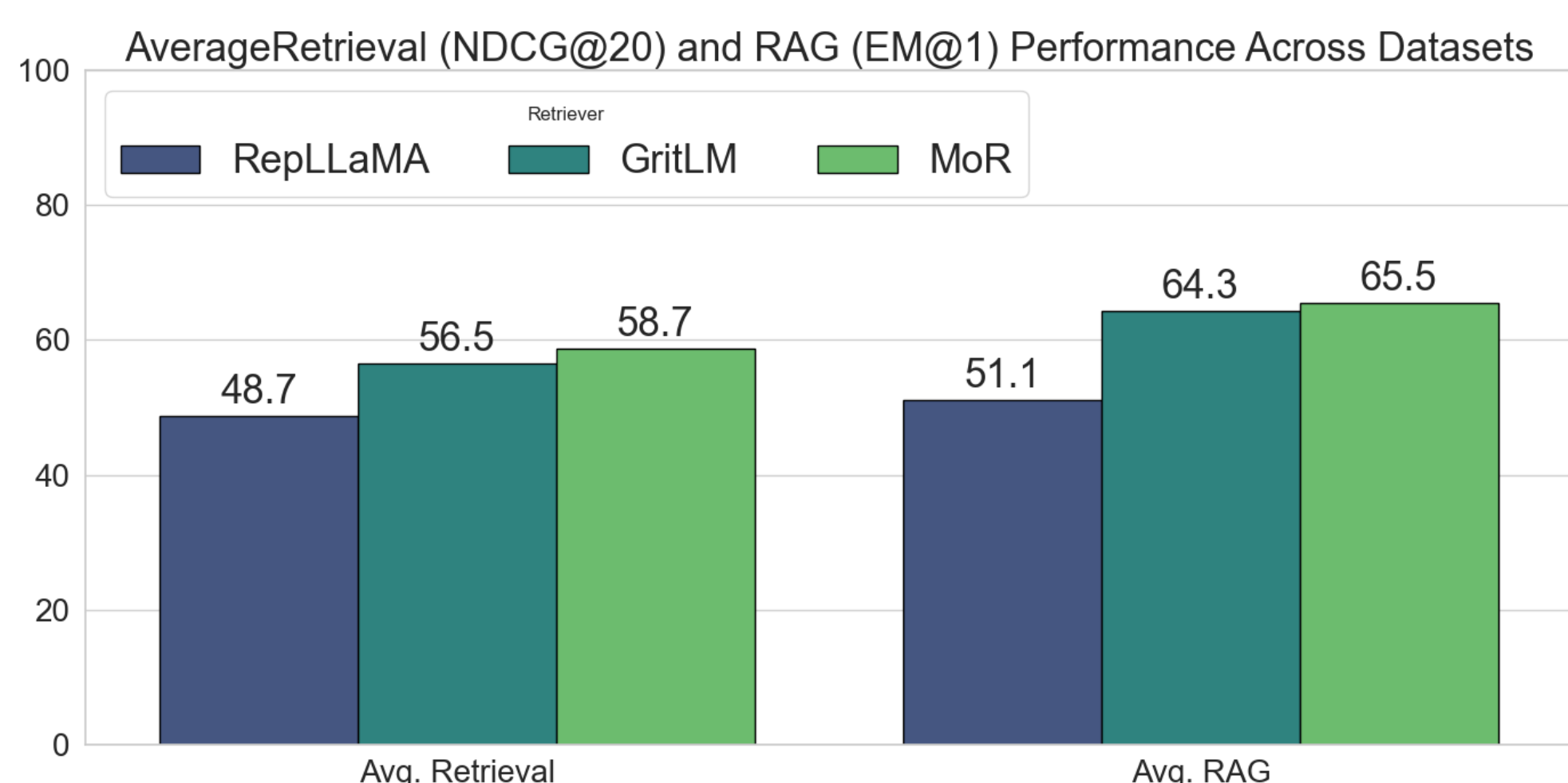
Pipeline: Mixture of Retrievers, handle diverse queries with a zero-shot, weighted combination of heterogeneous retrievers



Key Design: (1) per query-doc weight; (2) multi-granularity considered

Signal Source: (1) pre-retrieval: a *when-to-retrieve* problem (Vectorized Cluster. Distance); (2) post-retrieval: *query-performance-prediction* (Moran Index & Doc-Thrust)

Main Result: MoR improves RAG

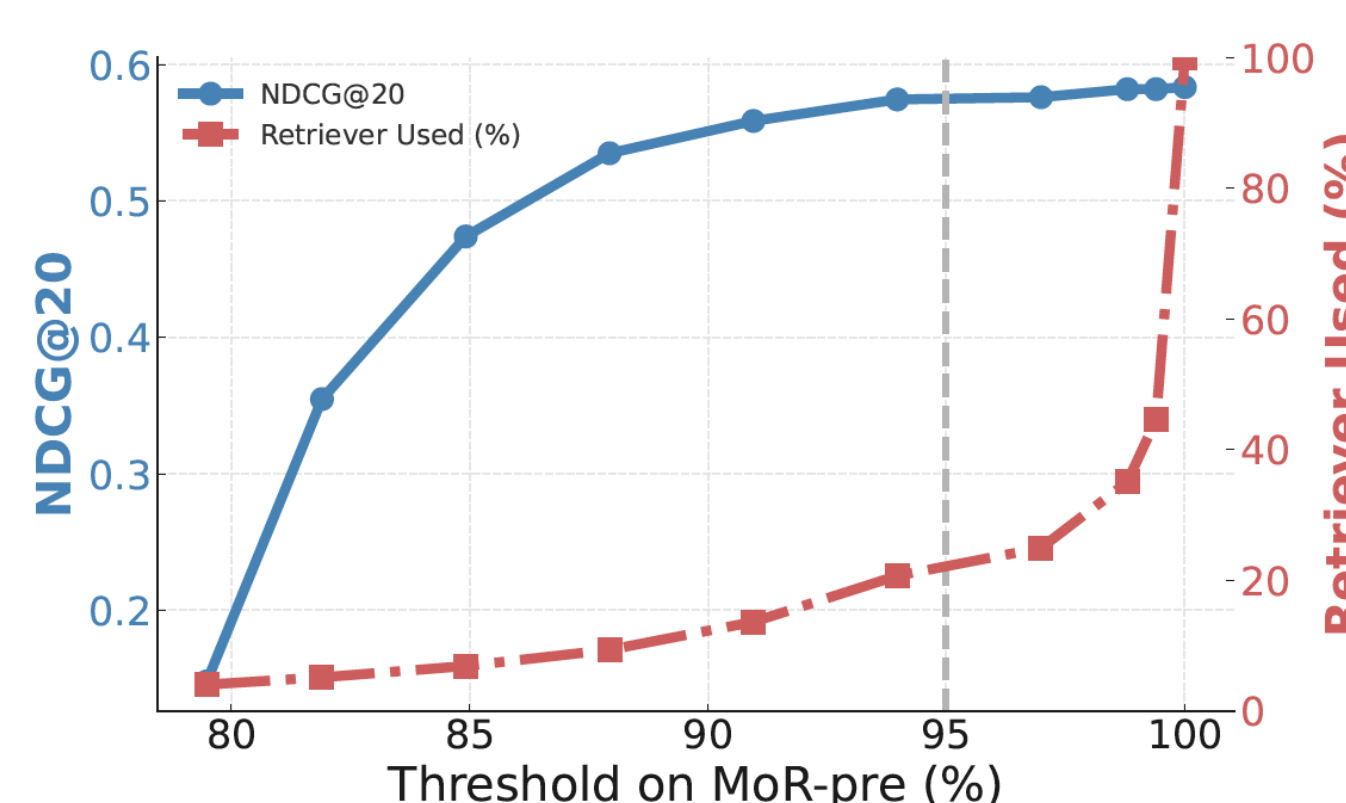


Experiment: 8 retrievers (BM25, SimCSE, DPR, TAS-B, MPNet, etc); 4 Tasks: NFCorpus, SciDocs, SciFact, SciQ

MoR achieves better performance than 7B retrievers with ~10% parameters

Consistent downstream gain on fact-check, QA

Analysis: Efficiency & Human Retriever



MoR is efficient!

Maintains performance with only 20% retrieval usage

Method	NDCG@20
GritLM	38.3
MoR	38.2
Human (sim.)	57.8
MoR + Human	91.8

Simulating human experts as retrievers

Human are retrievers!

MoR can incorporate fuzzy human info sources

When to use MoR?

Diverse questions, No Single Optimal Retriever
Need Efficiency
Human Experts can help

Check our paper!



Check our repo!

