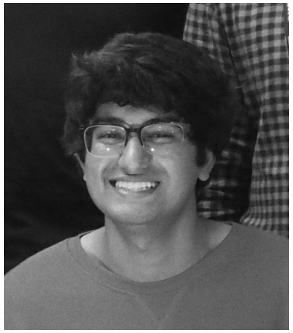
🍻 BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych



Nandan Thakur
UKP

Nils Reimers
Hugging Face

Andreas Rücklé
Amazon

Abhishek
Srivastava, UKP

Iryna Gurevych
UKP

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

https://www.ukp.tu-darmstadt.de/

🍻 What is 🔍 Information Retrieval?

🔍 Which football club does Lionel Messi play for?

natural language query

OR

🔍 Messi football club

keyword-based query

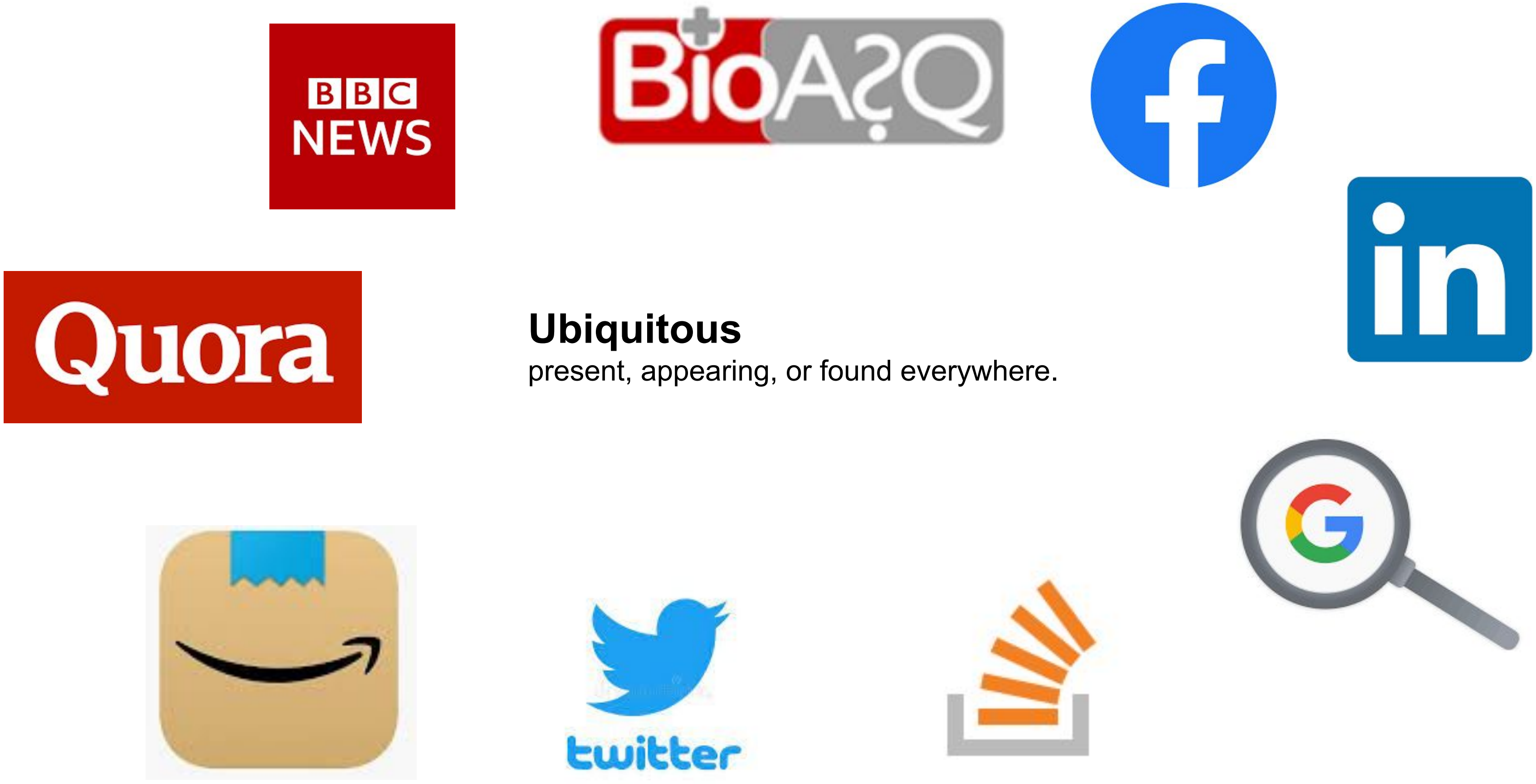
WIKIPEDIA
The Free Encyclopedia

5.5M Articles

Lionel Messi

Lionel Andrés Messi (born 24 June 1987), also known as Leo Messi, is an Argentine professional footballer who plays as a forward for Ligue 1 club **Paris Saint-Germain** and captains the Argentina national team. Often considered the best player in the world and widely regarded as one of the greatest players of all time, Messi has won a record six Ballon d'Or awards, a record six European Golden Shoes, and in 2020 was named to the Ballon d'Or Dream Team.

🍻 Why Study 🔍 Information Retrieval?

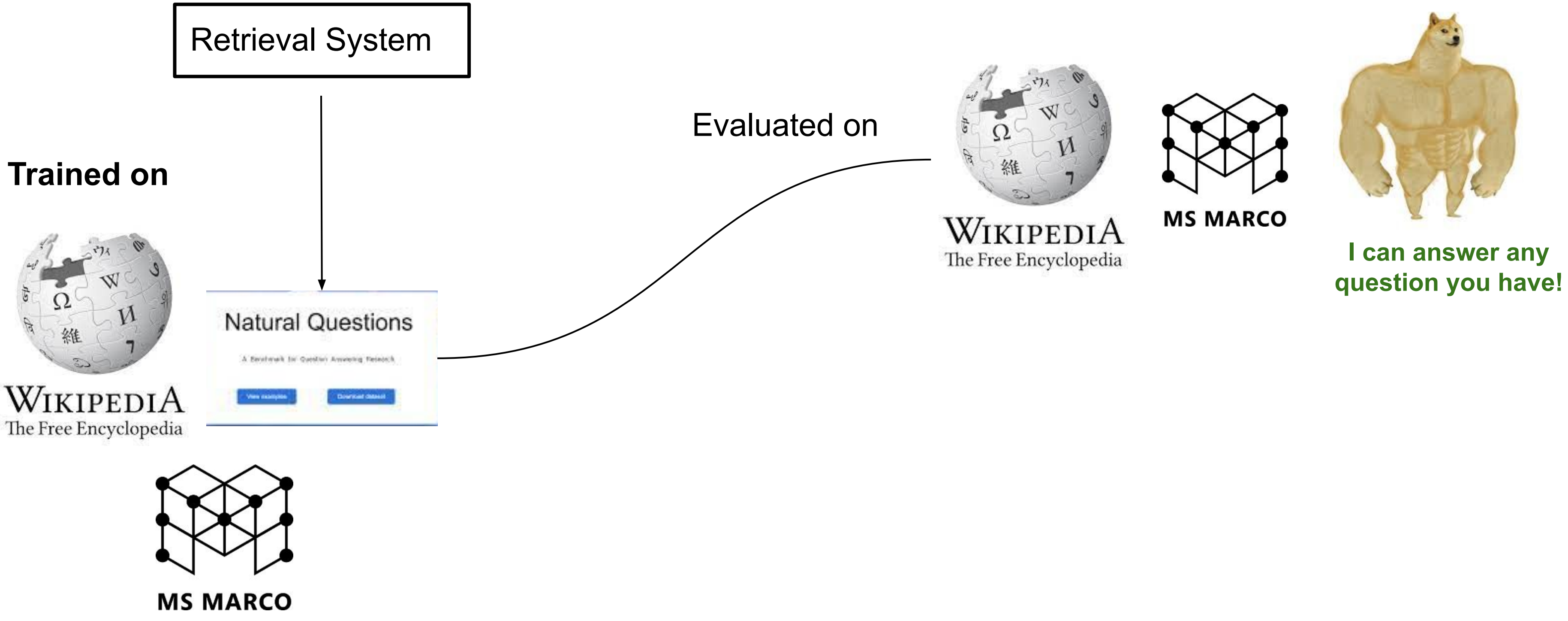


Ubiquitous

present, appearing, or found everywhere.

Existing Neural IR Setups

In-domain (Training data is available)



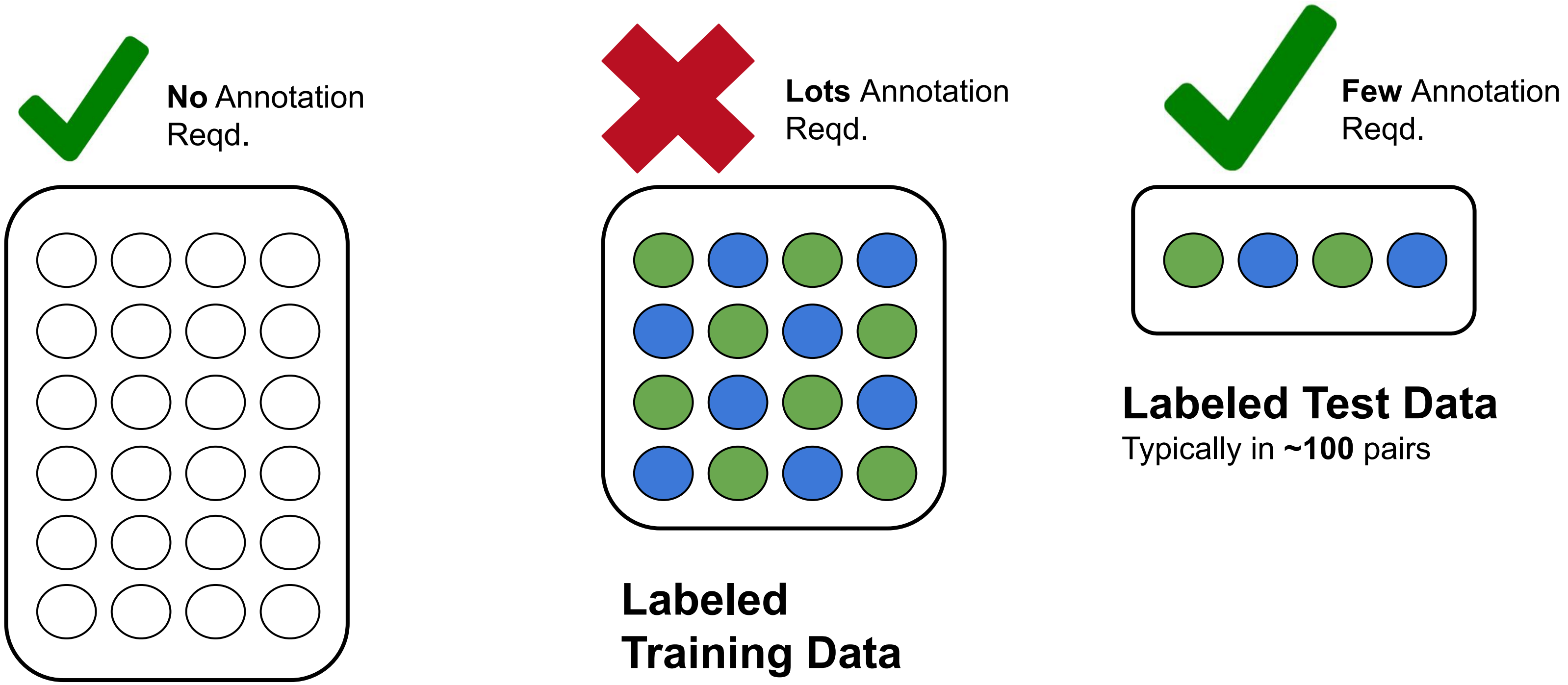
Retrieval System

Trained on

Evaluated on

I can answer any question you have!

Annotating Training Data is expensive!



No Annotation Reqd.

Lots Annotation Reqd.

Few Annotation Reqd.

Unlabeled Data

Typically in ~Millions

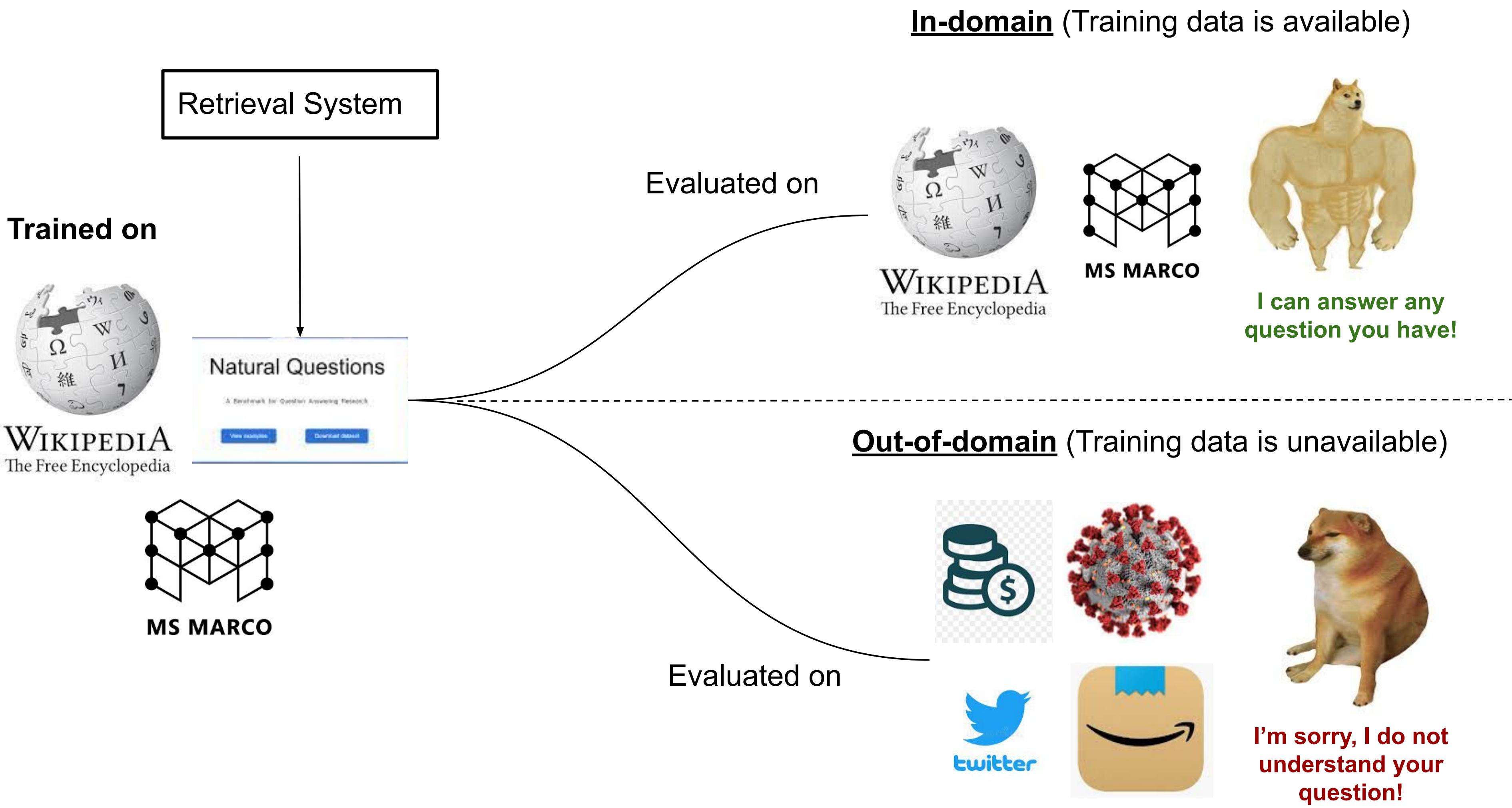
Labeled Training Data

Typically in **~100k** pairs

Labeled Test Data

Typically in **~100** pairs

Do these retrieval models generalize?



In-domain (Training data is available)

Retrieval System

Trained on

Evaluated on

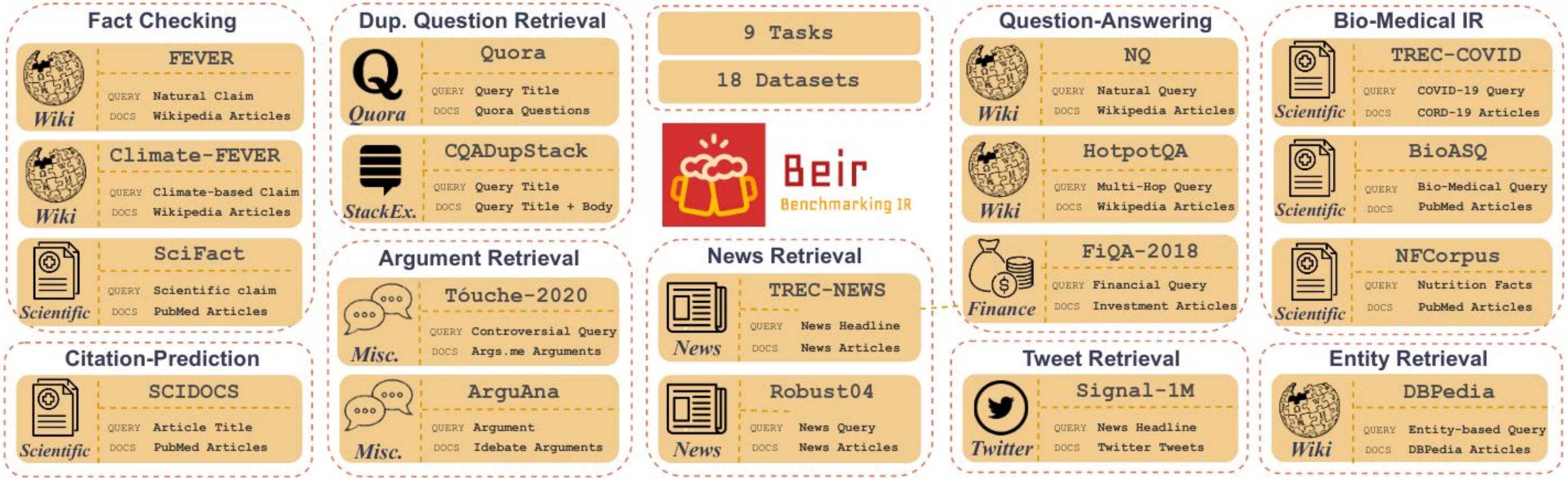
I can answer any question you have!

Out-of-domain (Training data is unavailable)

Evaluated on

I'm sorry, I do not understand your question!

The BEIR Benchmark



Fact Checking

FEVER — Wiki — QUERY Natural Claim — DOCS Wikipedia Articles

Climate-FEVER — Wiki — QUERY Climate-based Claim — DOCS Wikipedia Articles

SciFact — Scientific — QUERY Scientific claim — DOCS PubMed Articles

Citation-Prediction

SCIDOCS — Scientific — QUERY Article Title — DOCS PubMed Articles

Dup. Question Retrieval

Quora — Quora — QUERY Query Title — DOCS Quora Questions

CQADupStack — StackEx. — QUERY Query Title — DOCS Query Title + Body

Argument Retrieval

Tóuche-2020 — Misc. — QUERY Controversial Query — DOCS Args.me Arguments

ArguAna — Misc. — QUERY Argument — DOCS Idebate Arguments

9 Tasks

18 Datasets

Beir Benchmarking IR

News Retrieval

TREC-NEWS — News — QUERY News Headline — DOCS News Articles

Robust04 — News — QUERY News Query — DOCS News Articles

Question-Answering

NQ — Wiki — QUERY Natural Query — DOCS Wikipedia Articles

HotpotQA — Wiki — QUERY Multi-Hop Query — DOCS Wikipedia Articles

FiQA-2018 — Finance — QUERY Financial Query — DOCS Investment Articles

Tweet Retrieval

Signal-1M — Twitter — QUERY News Headline — DOCS Twitter Tweets

Bio-Medical IR

TREC-COVID — Scientific — QUERY COVID-19 Query — DOCS CORD-19 Articles

BioASQ — Scientific — QUERY Bio-Medical Query — DOCS PubMed Articles

NFCorpus — Scientific — QUERY Nutrition Facts — DOCS PubMed Articles

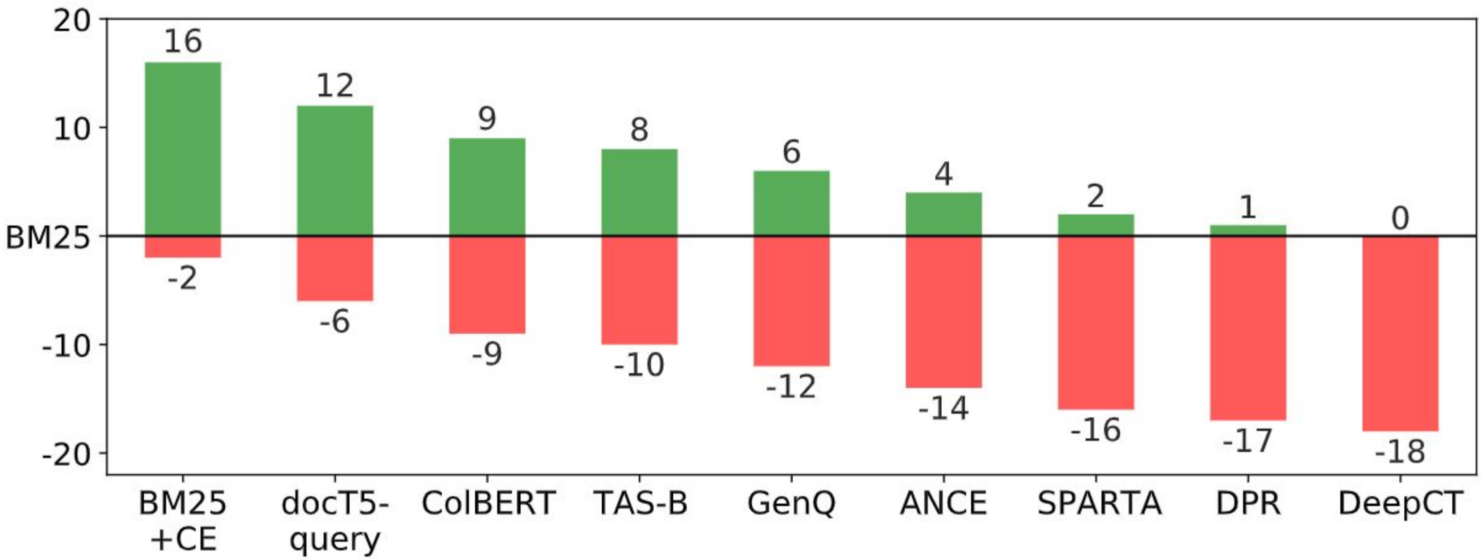
Entity Retrieval

DBPedia — Wiki — QUERY Entity-based Query — DOCS DBPedia Articles

- Robust evaluation of a retrieval system across 18 diverse datasets and 10+ domain types
- Contains datasets covering broad topics (like Wikipedia) and specific domains (COVID-19)
- Contains datasets with different corpus sizes (3k to 15Mil), query and document sizes, and different text types (Tweets vs. News articles)

🍻 Zero-shot Results on the BEIR Benchmark

Lexical	**Sparse**	**Dense**	**Late-Interaction**	**Reranking**
BM25 (Anserini)	DeepCT, SPARTA, DocT5query	DPR, ANCE, TAS-B, GenQ	ColBERT	BM25+ CE (MiniLM)



To Summarize

Motivation for creating the BEIR Benchmark

- Existing neural information models have been studied in limited or narrow settings.
- To robustly evaluate model generalization, we propose a zero-shot retrieval benchmark.
- The BEIR benchmark contains over 18 publicly available datasets for evaluation, spanning across 10 different retrieval tasks and domains.

Experimental results of diverse retrieval architectures on BEIR

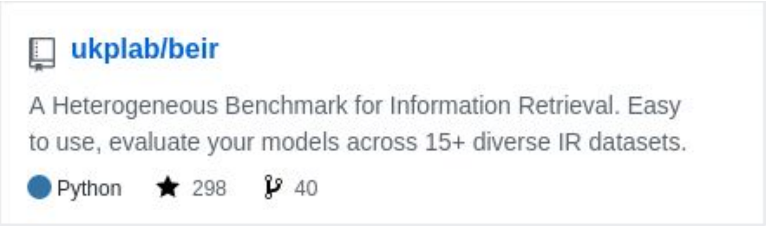
- Generalization with models is quite a difficult task and there is no free lunch!
- In-domain performances cannot be a good indicator for zero-shot performances.
- BM25 is a robust baseline, and performs competitively across several zero-shot datasets.
- Cross-Encoders or rerankers achieve the best zero-shot performances, but are slow at inference.
- Dense retrievers and sparse models suffer from out-of-distribution generalization.

Thank You!

TECHNISCHE
UNIVERSITÄT
DARMSTADT

- If you liked our work on BEIR benchmark, you can find more details in the GitHub repository.
- We actively maintain a leaderboard with diverse models and their zero-shot retrieval scores.
- For more interesting results, we would suggest you to read our NeurIPS publication.

I look forward to meet you virtually and answer your questions at NeurIPS'21

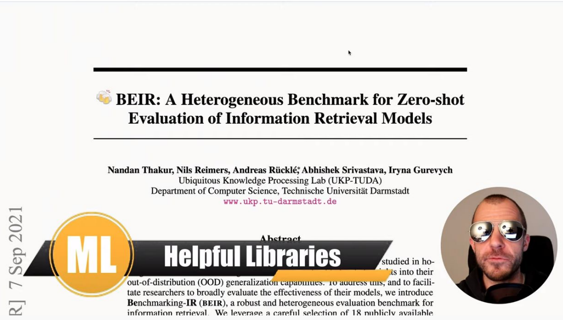


ukplab/beir

A Heterogeneous Benchmark for Information Retrieval. Easy to use, evaluate your models across 15+ diverse IR datasets.

● Python ★ 298 40

deepset



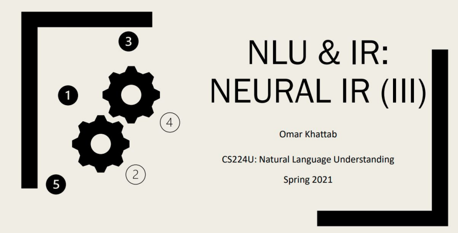
BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models

Nandan Thakur, Nils Reimers, Andreas Rücklé*, Abhishek Srivastava, Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Abstract

studied in ho- ts into their out-of-distribution (OOD) generalization capabilities. To address this, and to facilitate researchers to broadly evaluate the effectiveness of their models, we introduce Benchmarking-IR (BEIR), a robust and heterogeneous evaluation benchmark for information retrieval. We leverage a careful selection of 18 publicly available

Beir
Benchmarking IR



NLU & IR:
NEURAL IR (III)

Omar Khattab

CS224U: Natural Language Understanding

Spring 2021

Google colab

https://colab.research.google.com/drive/1HfutiEhHMJLXiWGT8pcipxT5L2TpYEdt?usp=sharing