# Advanced Information Retrieval

**Koç University, 1st June 2023**

**Nandan Thakur**
PhD Student
Current: Part time Student Researcher @ Google Research [Remote]

**David R. Cheriton School of Computer Science**
**University of Waterloo**

# My Journey till now (Roadmap)

- **Current:** Second-year PhD student at the University of Waterloo, Canada
- **Current:** Research Internship at Google Research, Remote.
- **Previous:** Research Assistant (RA) at the UKP Lab, TU Darmstadt.

**2018**

**Undergraduate**

Studied **Electronics** at BITS Pilani, India. Additionally graduated with a minors in **Finance**.

**2019**

**Industry**

Worked as a Software Developer in a Startup in Bangalore, India. Developed a Content Search Product for Companies.

**2020**

**Research (UKP)**

Developing my curiousity for research under the supervision of **Nils Reimers** and **Prof. Iryna Gurevych** at UKP Lab, TU Darmstadt

**2021**

**PhD (UWaterloo)**

Started my PhD in Fall 2021 under the supervision of **Prof. Jimmy Lin**. Focus on efficient and generalized Neural Information Retrieval.

**2022**

**Google Research**

Currently I am working with Daniel Cer and JIanmo Ni on diverse challenges faced in Cross-lingual IR

# What is Information Retrieval?

Google

Information retrieval

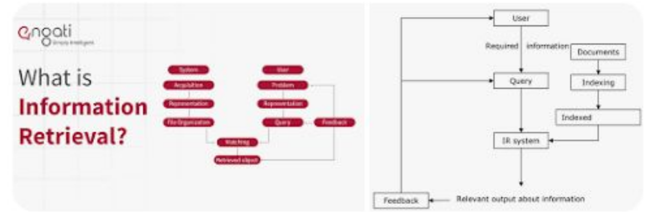
All | Images | Books | Videos | News | More | Tools

About 346,000,000 results (0.43 seconds)

Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds.

Wikipedia
https://en.wikipedia.org › wiki › Information_retrieval

Information retrieval - Wikipedia




About featured snippets • Feedback

## People also ask

What is an example of information retrieval?

What is information retrieval main purpose?

What is the basic concept of information retrieval?

What are the three types of information retrieval?

Feedback

Stanford University
https://nlp.stanford.edu › IR-book › information-retrie...

Introduction to Information Retrieval - Stanford NLP Group

The book aims to provide a modern approach to **information retrieval** from a computer science perspective. It is based on a course we have been teaching in ...

Boolean retrieval · Irbook.html · Resources · CS 276 / Ling 286

GeeksforGeeks
https://www.geeksforgeeks.org › what-is-information-...

What is Information Retrieval?

Jul 3, 2022 — It is A process of identifying and retrieving the data from the database, based on the query provided by user or application. Retrieves ...

## Information retrieval

Information retrieval in computing and information science is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing. Wikipedia

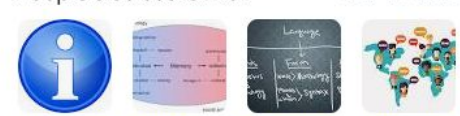
Brain

Types

Field

Features

People also search for — View 10+ more



Information | Memory | Semantics | Language

Feedback

# Formal Definition of the Retrieval Task

**Query (Natural language)**

**Which football club does Lionel Messi play for?**

**Query (Keyword)**

**OR**

**Messi football club**

**Document**

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 is 🔍 Information Retrieval Important?

**Ubiquitous**
present, appearing, or found everywhere.

# IR Tasks: Architecture

# What Happens in a Ad-hoc Retrieval System?

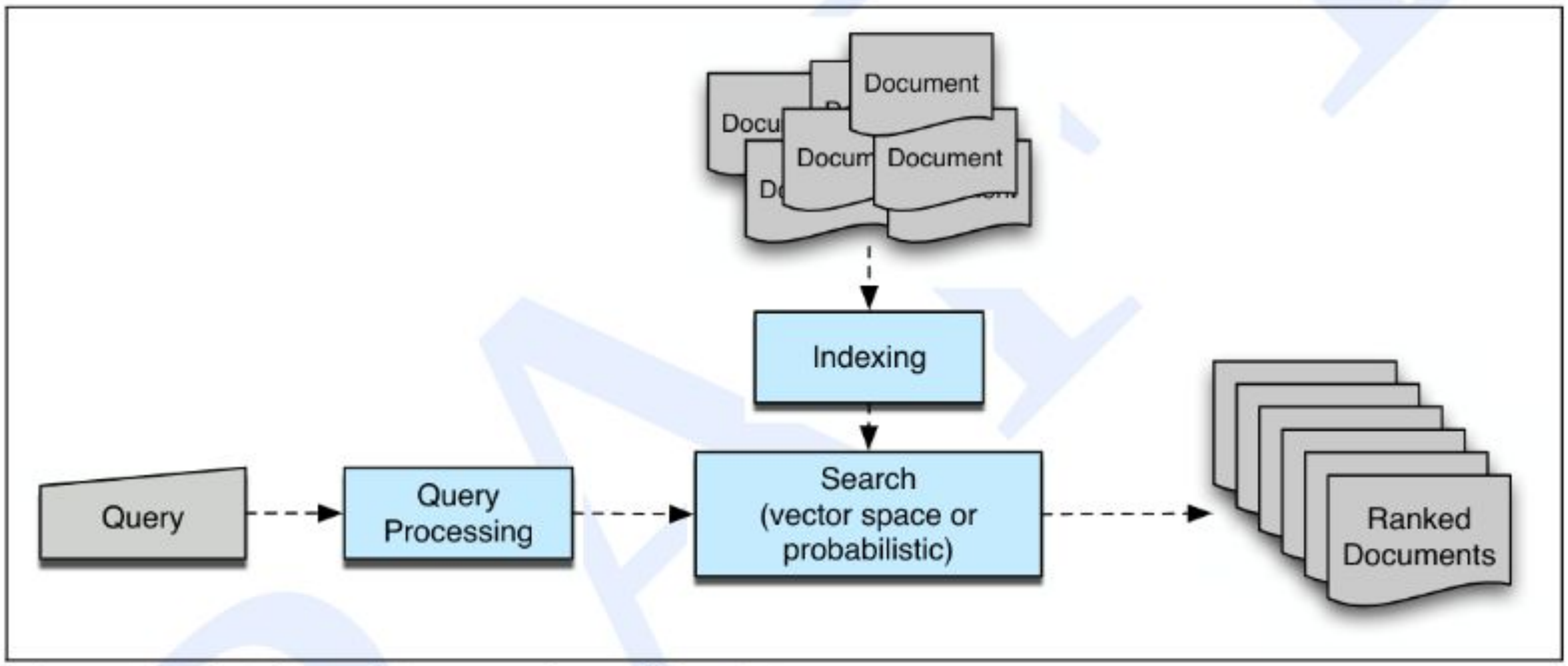


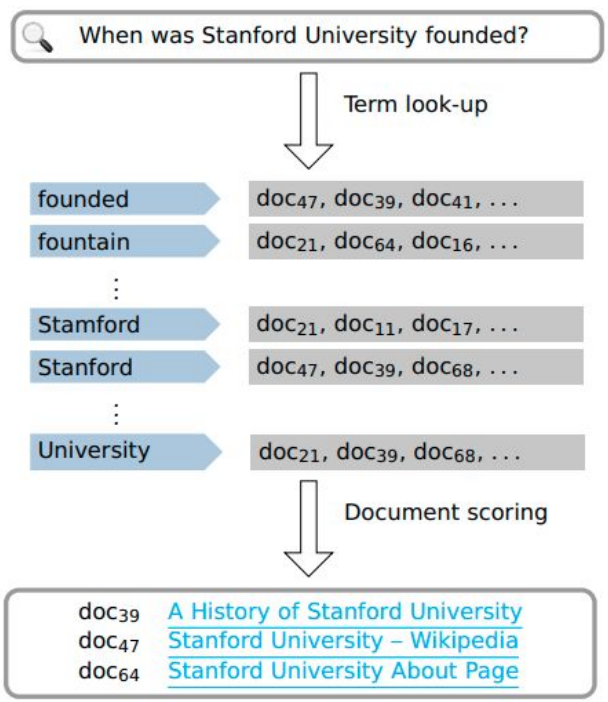

**Figure 23.2** The architecture of an ad hoc IR system.

Figure taken from Speech and Language Processing, 2nd Edition by Dan Jurafsky and James H. Martin.

# Traditional Search Systems

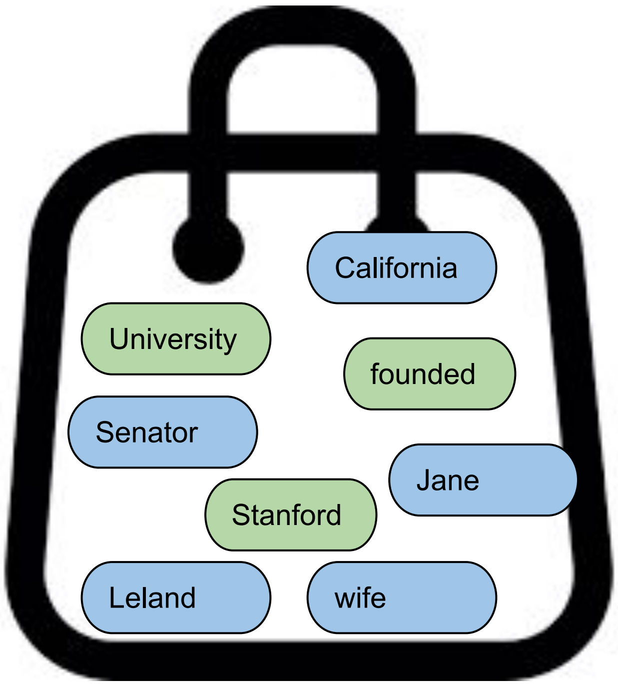
# TF-IDF (Bag-of-Words Model)

## Keyword based Search: Exact Match of Words




**Q:** When was Stanford University founded?

**Doc:** Stanford University was founded in 1885 by California senator Leland Stanford and his wife, Jane.




Ref: Christopher G Potts, ACL-IJCNLP 2021 keynote address
https://web.stanford.edu/~cgpotts/talks/potts-acl2021-slides-handout.pdf

# TF-IDF Intuition and Example

Corpus D

Doc 1: A quick brown **fox** jumps over the lazy dog. What a **fox**!

Doc 2: A quick brown **fox** jumps over the lazy **fox**. What a **fox**!

Doc 3: A quick brown dog jumps over the lazy dog. What a dog!

**TF**: Frequency of any “term” in a given document.

**IDF**: Ratio of documents which include the “term”.

First, let’s compute Term Frequency (TF) and Inverse Document Frequency (IDF) for **“fox”**:

TF(“fox”, Doc 1): 2 / 12 = 0.17, TF(“fox”, Doc 2): 3 / 12 = 0.25, TF(“fox”, Doc 3): 0 / 12 = 0

IDF(“fox”, D) = log(3/2) = 0.18

**TF-IDF score = TF x IDF**

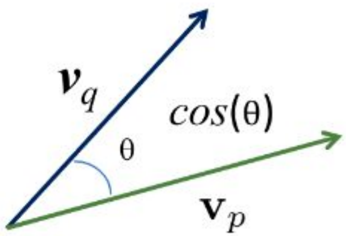
TF-IDF(“fox”, Doc 1) = 0.03 TF-IDF(“fox”, Doc 2) = **0.045** TF-IDF(“fox”, Doc 3) = 0

# Modern (Neural) Search Systems
# Part 1: Dense Retrieval

# Limitations with Traditional Systems

## Why do we need modern (neural) search systems?

**Huge Memory Indexes**: Sparse vectors are big and can be quite inefficient to store!




$d_1 >> d_2$

sparse repr: $[0\ldots\ 1\ \ldots.\ 1\ \ldots0..1] \in \mathbb{R}^{d_1}$

dense repr: $[1.03, -5.72, 6.42, .., 9.91] \in \mathbb{R}^{d_2}$

**Unable to handle Synonyms**: Won't understand "*bad guy*" and "*villain*" are similar in meaning!

dense

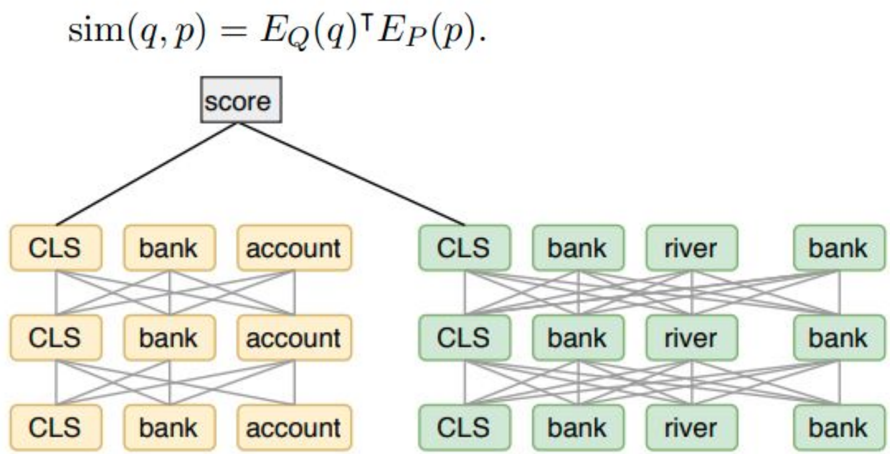
"Who is the **bad guy** in lord of the rings?"

*Sala Baker is an actor and stuntman from New Zealand. He is best known for portraying the* ***villain*** *Sauron in the Lord of the Rings trilogy by Peter Jackson.*

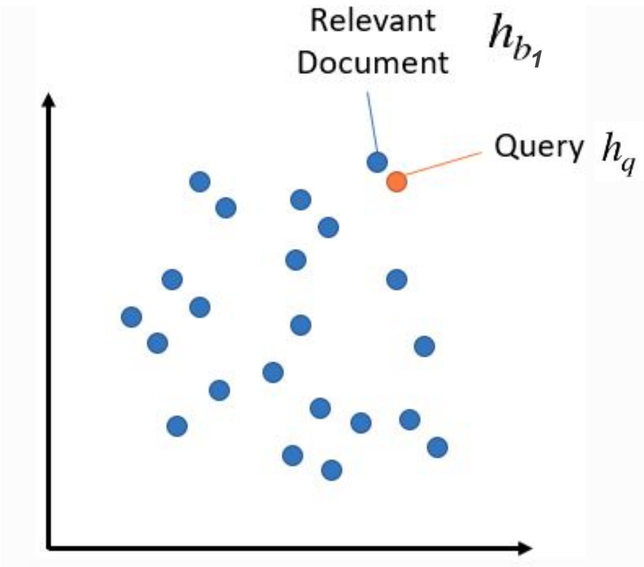
Ref: Danqi Chen, ACL 2020 OpenQA Tutorial
https://github.com/danqi/acl2020-openqa-tutorial/blob/master/slides/part5-dense-retriever-e2e-training.pdf

# Dense Retrieval with Bi-Encoders

## Mapping Individual Text to a fixed dimensional embedding!




(b) Dense Retrievers (e.g., DPR)

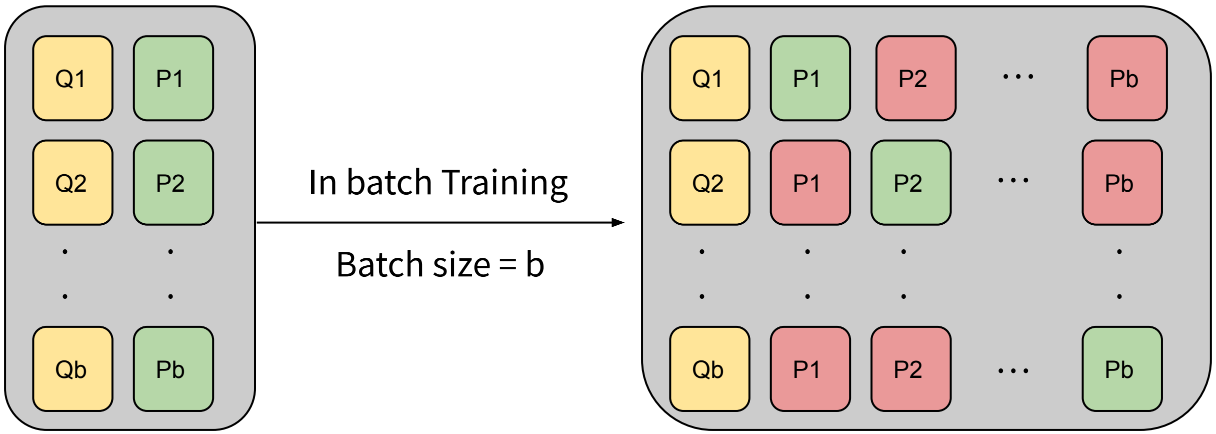



2D Vector Space

- Passage Embeddings can be precomputed using BERT and stored!
- Fast and efficient at runtime, ideal for a practical system!

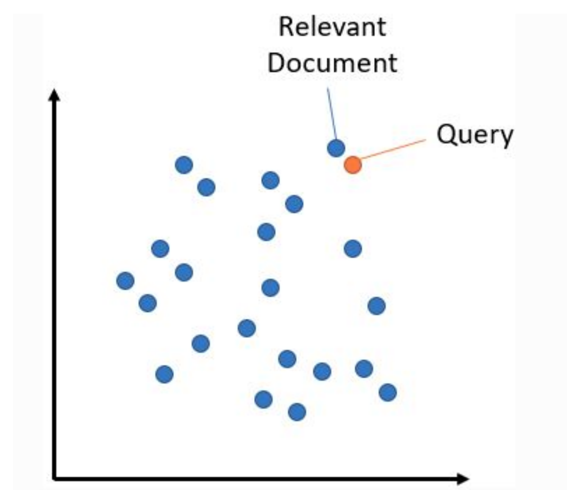
# How to train the Dense Retriever model?

## Method 1: Inbatch Fine-tuning with Random Negatives



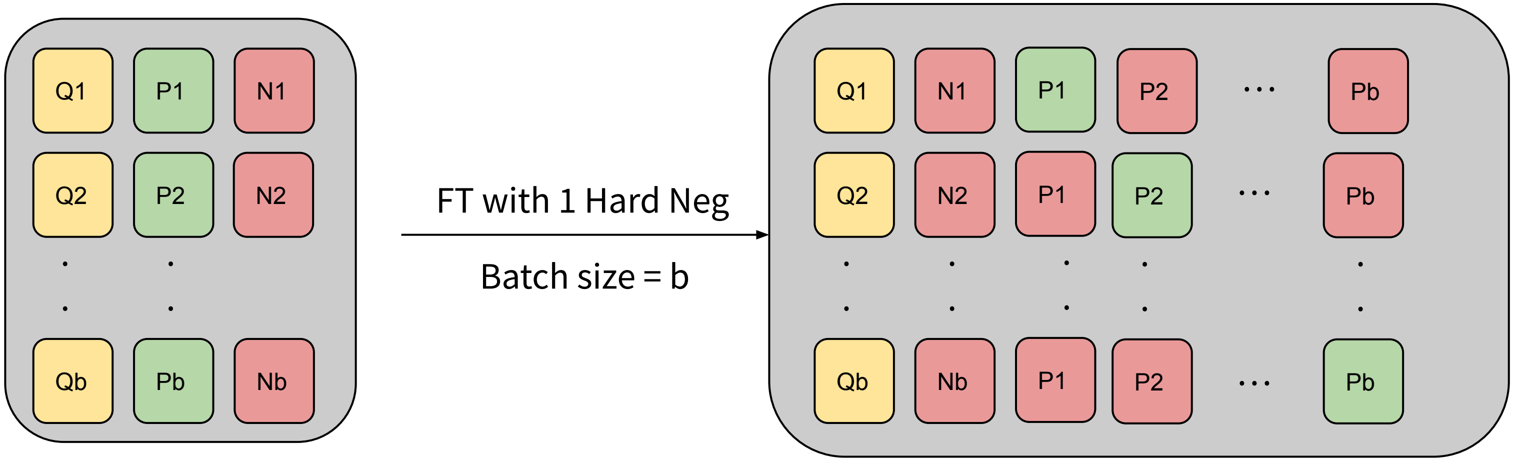

Cross-Entropy loss function

$$
\begin{aligned}
& L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) \\
= \quad & -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}
\end{aligned}
$$



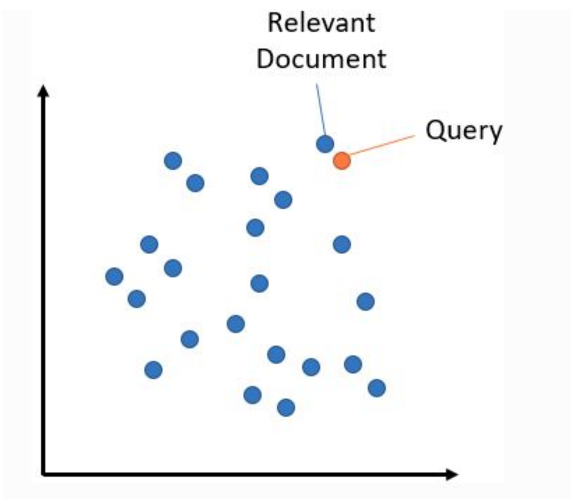
# How to train the Dense Retriever model?

## Method 2: Inbatch Fine-tuning with 1 Hard Negative



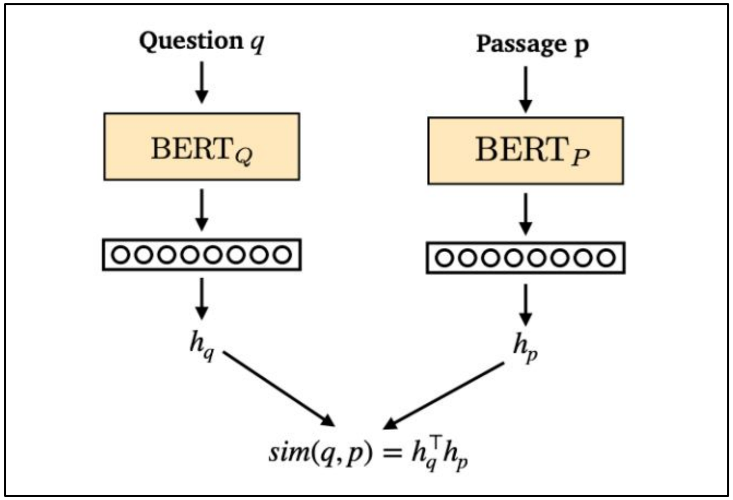

Cross-Entropy loss function

$$
\begin{aligned}
& L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) \\
= & -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}
\end{aligned}
$$



# DPR: Dense Passage Retriever (kharpurkin et al. 2020)

DPR Model Architecture

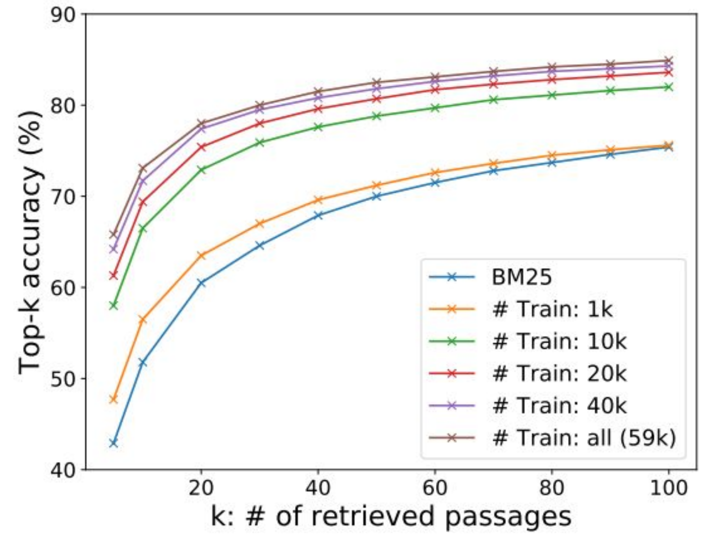



Training Loss Function

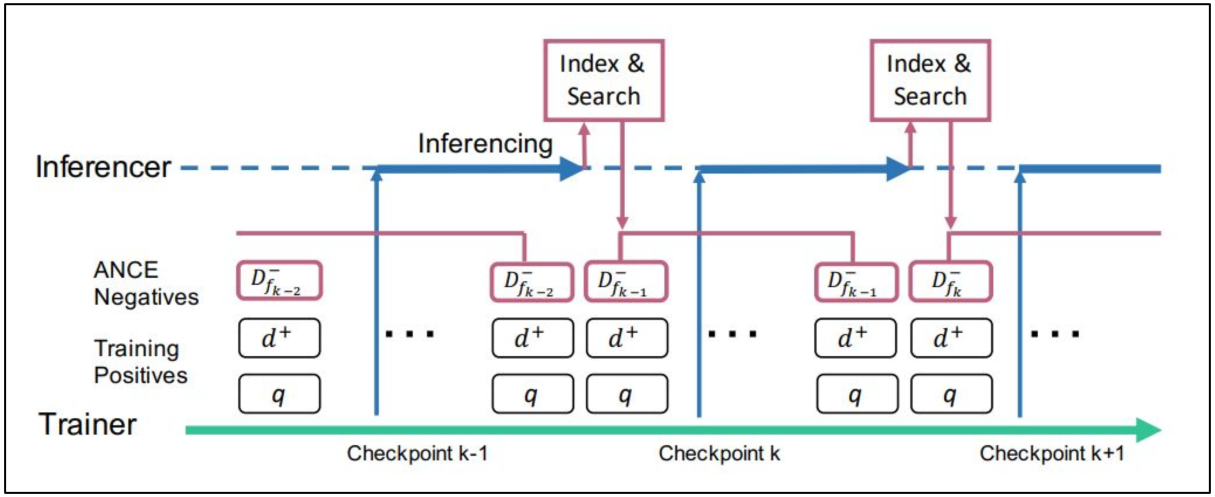
$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}$$

DPR can outperform a traditional IR system (such as BM25) using ~1k train examples.

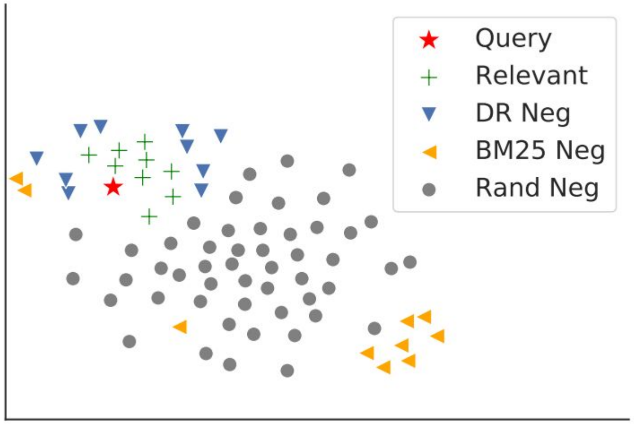
Natural Questions (Kwiatkowski et al., 2019)



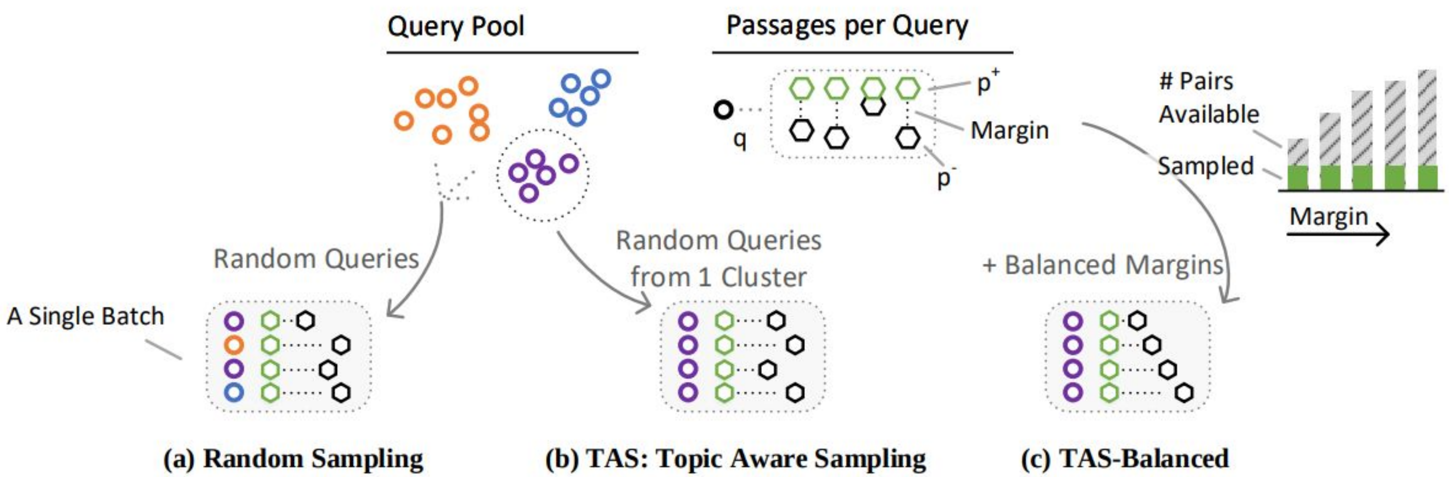
# ANCE: Approximate Nearest Neighbor Negative Contrastive Learning (Xiong et al. 2021)

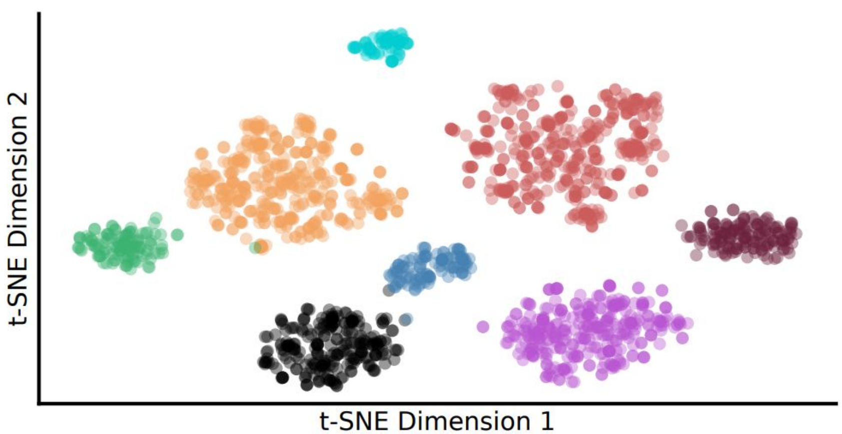



$$\theta^* = \text{argmin}_\theta \sum_q \sum_{d^+ \in D^+} \sum_{d^- \in D^-_{\text{ANCE}}} l(f(q, d^+), f(q, d^-)),$$



# TAS-B: Topic-Aware Query and Balanced Margin Sampling Technique (Hofstätter et al. 2021)




$$\mathcal{L}_{Pair}(Q, P^+, P^-) = \text{MSE}(M_s(Q, P^+) - M_s(Q, P^-), M_t(Q, P^+) - M_t(Q, P^-))$$

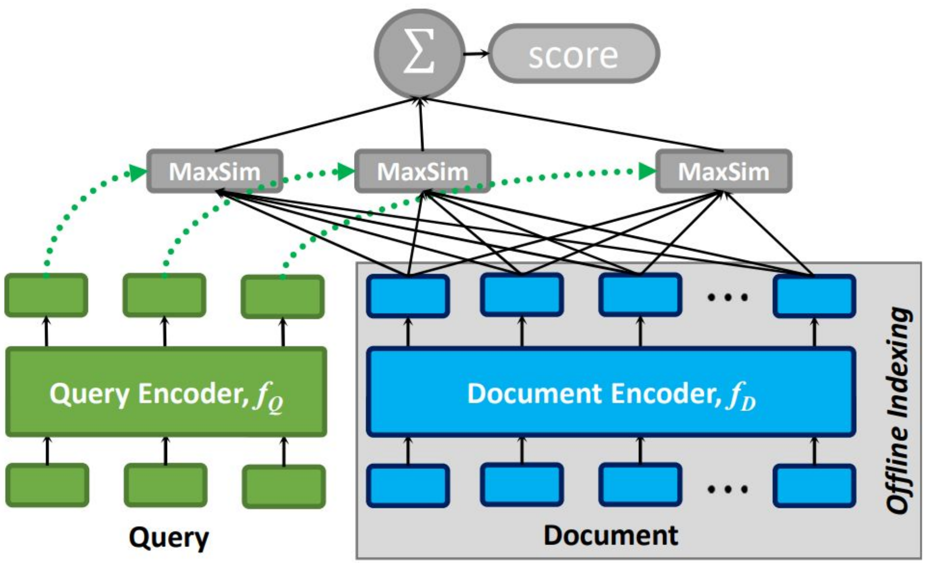
$$\mathcal{L}_{InB}(Q, P^+, P^-) = \frac{1}{2|Q|}\left(\sum_{i}^{|Q|}\sum_{p^-}^{P^-}\mathcal{L}_{Pair}(Q_i, P_i^+, p^-) + \sum_{i}^{|Q|}\sum_{p^+}^{P^+}\mathcal{L}_{Pair}(Q_i, P_i^+, p^+)\right)$$

# Modern (Neural) Search Systems
# Part 2: Late Interaction

# ColBERT (Late-Interaction) (Khattab et al. 2020)

## Mapping Individual tokens to fixed dimensional embeddings

- ColBERT model maps an individual token to a fixed dense embedding.
- ColBERT allows “token-level interactions” between queries and documents.




Sum of Maximum Similarity

$$S_{q,d} := \sum_{i \in [|E_q|]} \max_{j \in [|E_d|]} E_{q_i} \cdot E_{d_j}^T$$

Figure taken from ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT by Omar Khattab and Matei Zaharia.

# ColBERT (Late-Interaction) (Khattab et al. 2020)

## Inference Method of ColBERT model

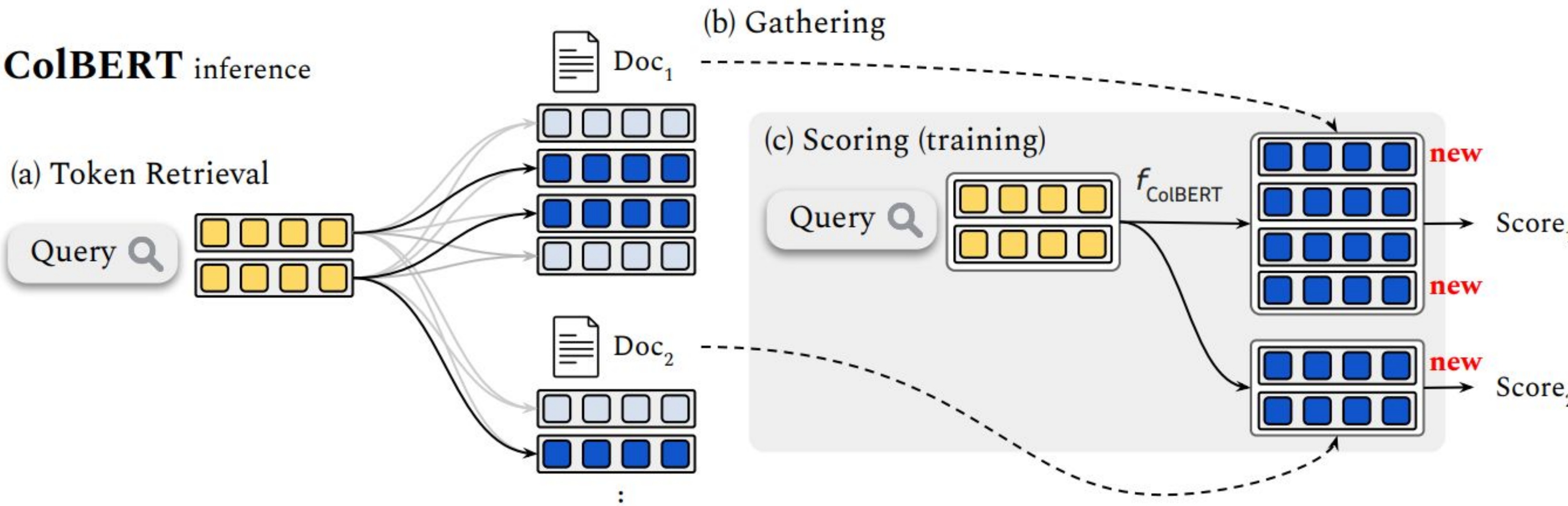



Figure taken from XTR: Rethinking the Role of Token Retrieval in Multi-Vector Retrieval by Jinhyuk Lee et. al.
.

**(a) Token Retrieval**
Query tokens used to search top-(k') doc tokens (among all tokens in corpus).

**(b) Gathering**
top-(k") tokens are mapped to the original doc-id.

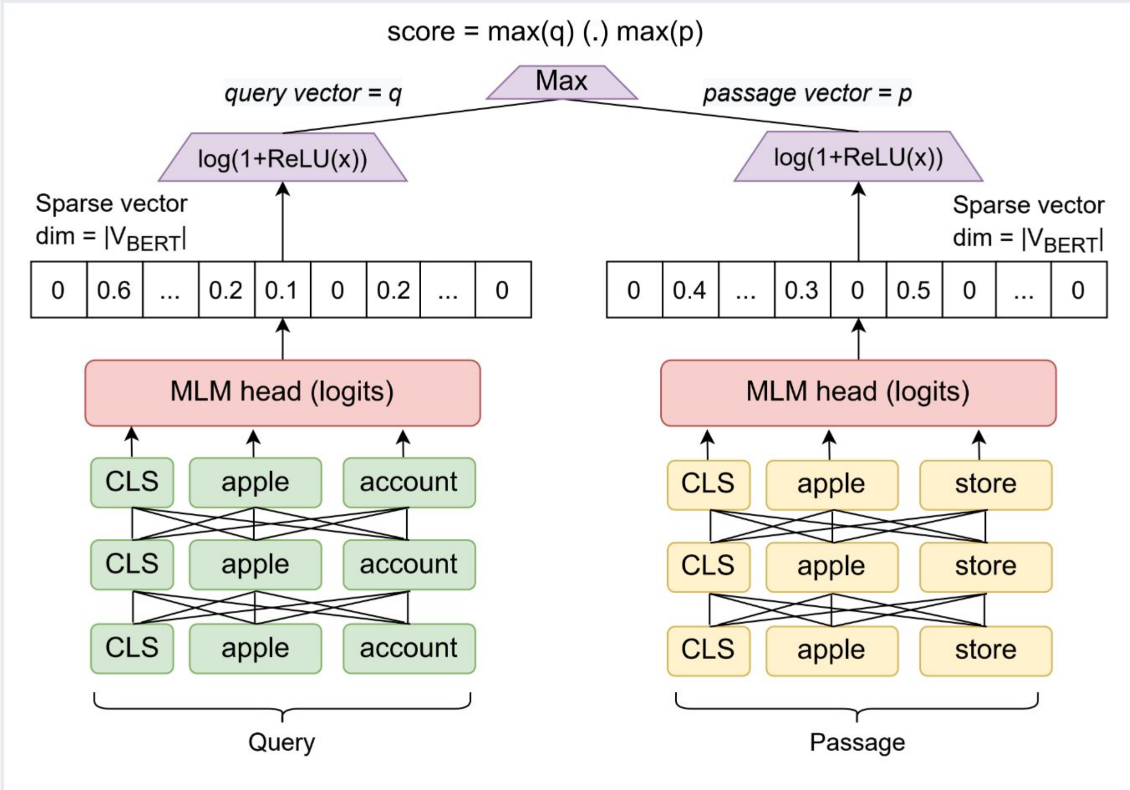
**(c) Scoring**
The unique documents are used to compute MaxSim and score.

# Modern (Neural) Search Systems
# Part 3: Sparse Retrieval

# SPLADE (Sparse Retrieval) (Formal et al. 2020)

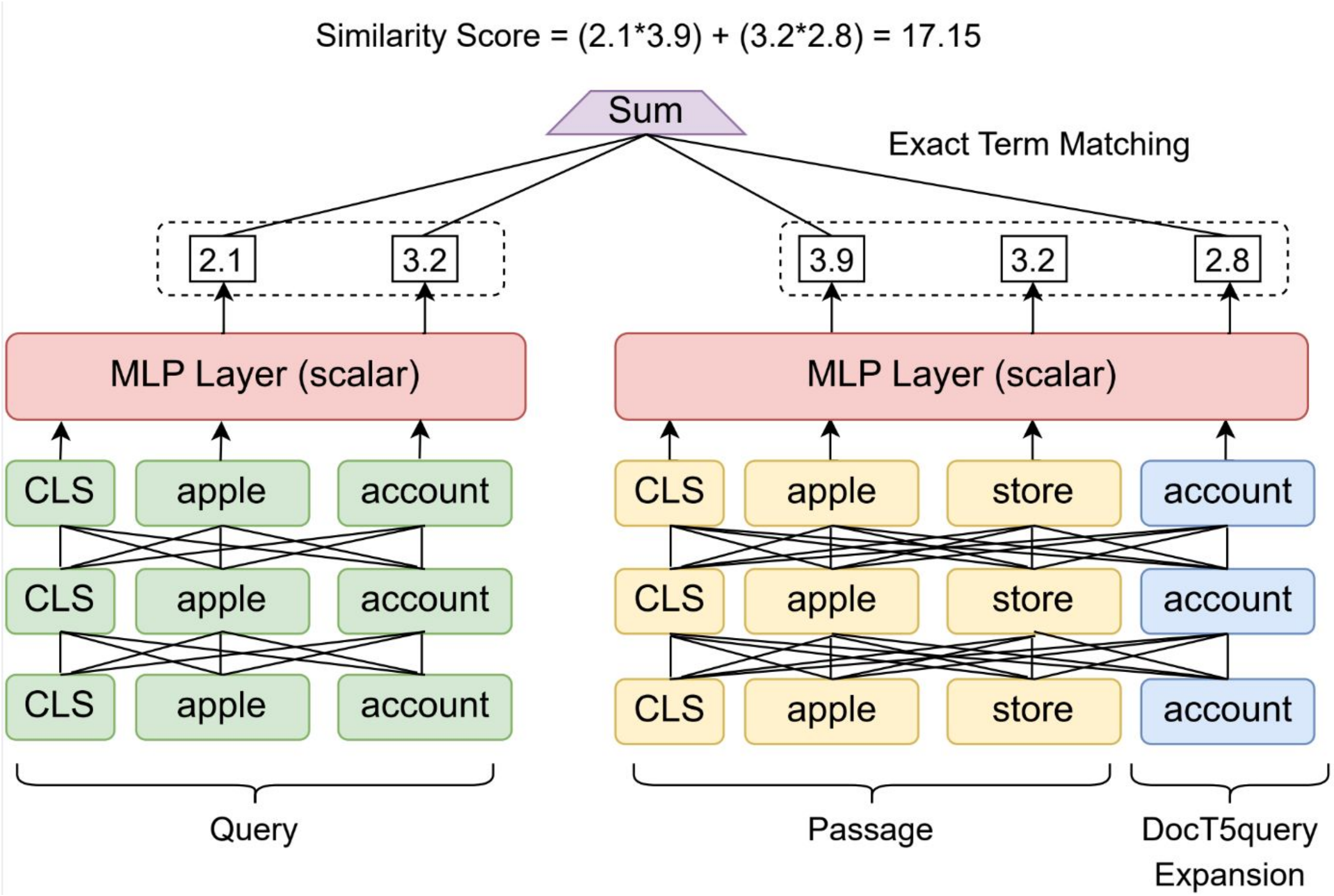
## Mapping scalar weights across whole BERT Vocabulary

- SPLADE model produces weights for a 30k long sparse vector.
- Score can be efficiently computed using an inverted index algorithm.



# uniCOIL (Sparse Retrieval) (Lin et al. 2021)

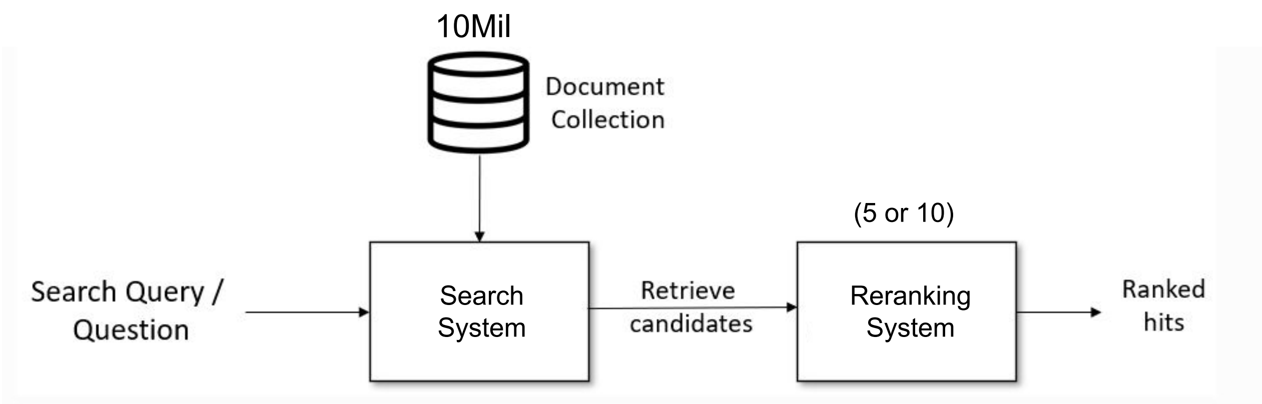
Mapping scalar weights across for words in input paragraph



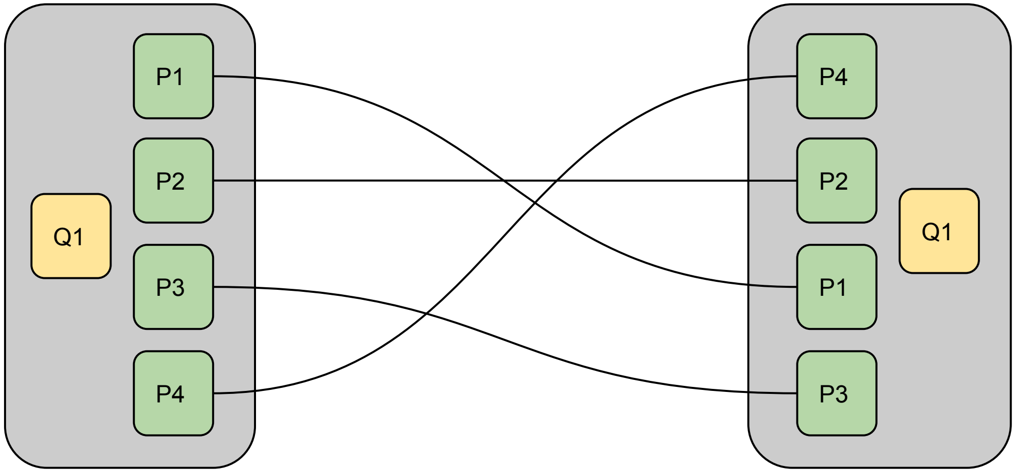
# Modern (Neural) Search Systems
# Part 4: Cross-Encoder Reranker

# Retrieve and Rerank

Change the order of docs and bring best documents on top.

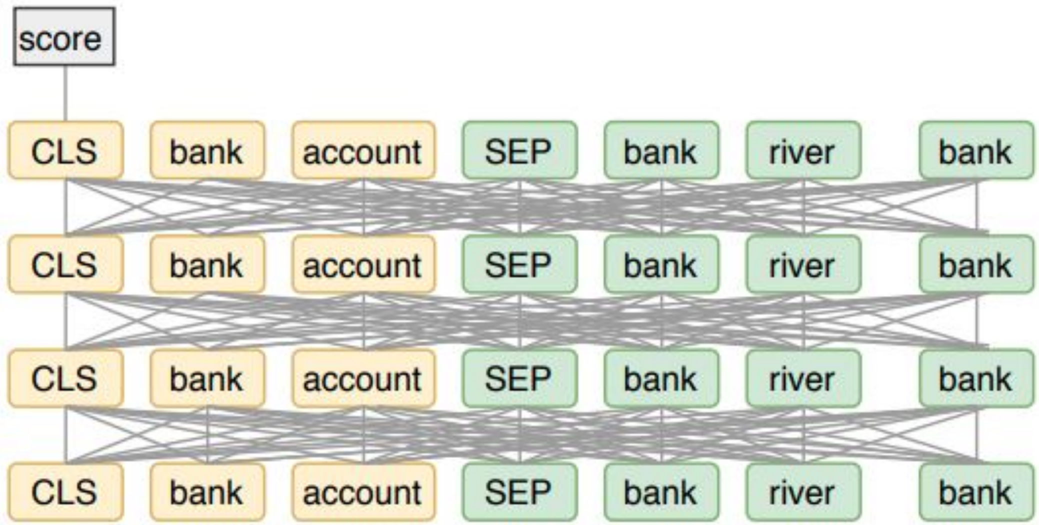



**Reranking Algorithm Intuition**



# Reranking with Cross-Encoders

Concatenate Query and Document together. No Embedding!




(a) Cross-Attention Model (e.g., BERT reranker)

- Inefficient, as scoring millions of (query, doc)-pairs is slow!
- Best performance, due to cross-attention across query and doc.

# Traditional IR Benchmarking

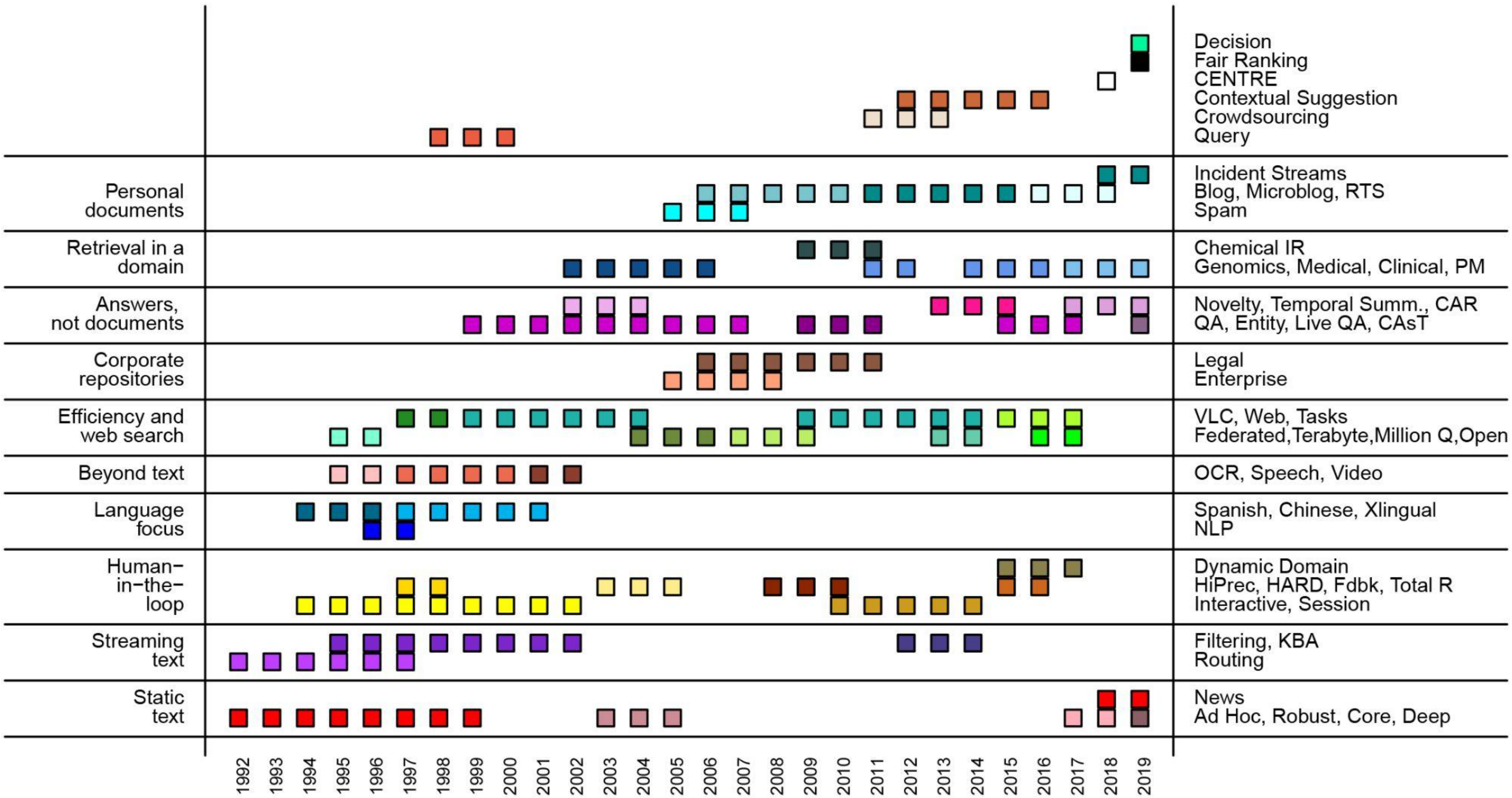
# What is Benchmarking? Why is it Useful?

Benchmarks in **NLP/IR** has three components: (1) it consists of one or multiple datasets, (2) one or multiple associated metrics, and (3) a way to aggregate performance.

**Advantages of Benchmarking**

- Helps provide a **unified platform** utilized for comparing our ML model performances
- Leads to a way of **discovering** what is state-of-the-art (SoTA) being achieved
- Useful in understanding fundamental **gaps** in existing evaluated models
- Benchmarks help to point out difference to **human level** performances
- Sets a **standard** for assessing the performance of different systems in the community

# TREC Suite: History of IR Benchmarking




The TREC tasks. A box represents the corresponding task occurring in the given year. The far right column lists the names of the TREC track that included the task, and the far left column provides a short gloss of the research focus of the task. Differing colors within a row show the evolution of the task in different tracks.

# How to build a TREC test Collection?

1. Build a corpus *C* using a set of documents and queries (also called topics in TREC)
   a. For e.g., Corpus with Law articles

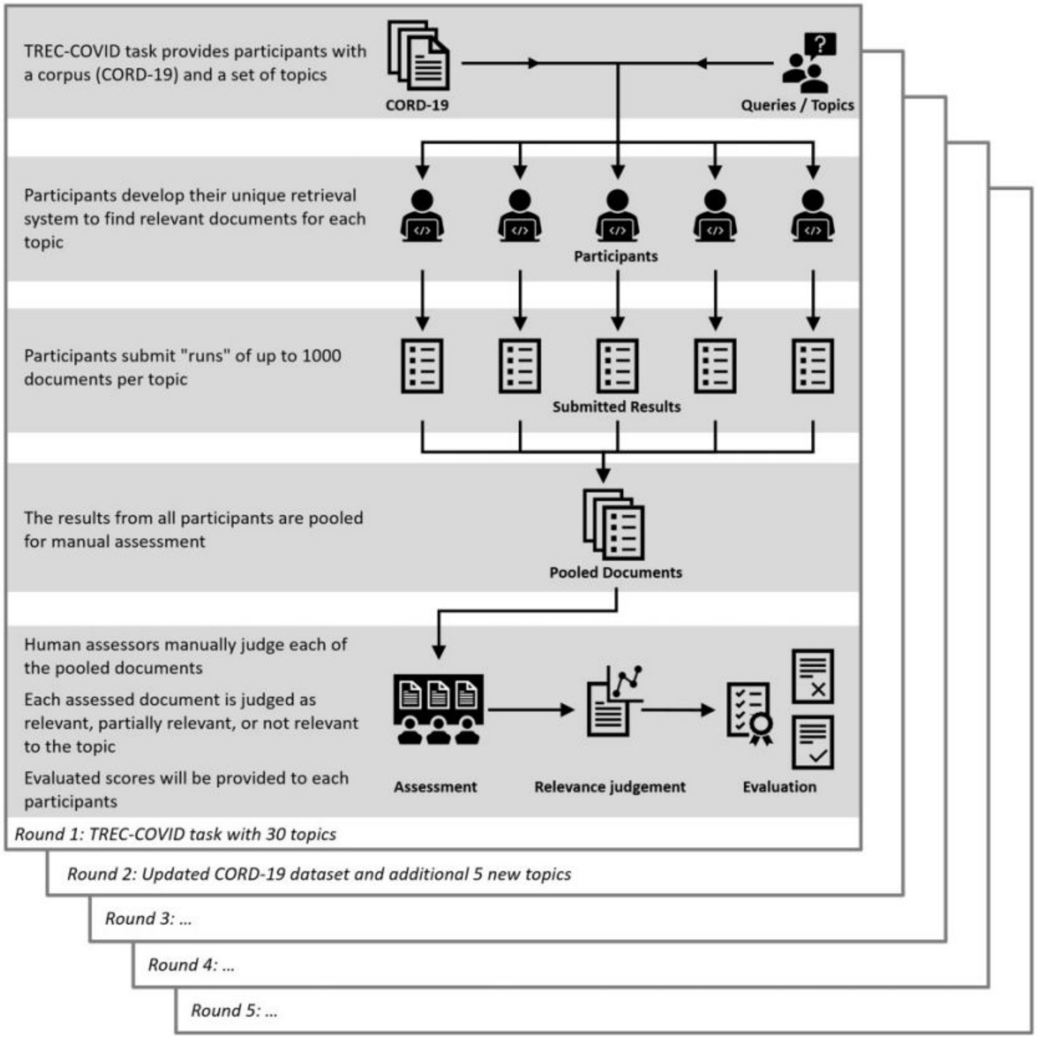
2. The initial participants in the TREC competition runs the queries against documents
   a. Returns the top documents per query
   b. Participants can develop any system for retrieval
   c. Coopetition = Cooperation + Competition

3. Evaluation pool is formed and then judged by relevance assessors
   a. Evaluated using relevance judgements (binary or multiple levels)

4. Results then are returned to participants who participated in the competition.

5. Relevance Judgements turn the documents and topics into test collection.

# Example: The TREC-COVID Test Collection




**Advantages:**

Pooling ensures diversity among the judged annotations.

Encourages audience to participate in lieu of their model retrieved results will get judged by annotators

Gradually keep on adding topics, and updating the dataset every year.

# IR Evaluation Metrics

# Common IR Evaluation Metrics

**Precision (position unaware)**: fraction of retrieved docs that are relevant = P(relevant|retrieved)

**Recall (position unaware)**: fraction of relevant docs that are retrieved = P(retrieved|relevant)

**MRR (position aware):** position of the first relevant doc which is retrieved = 1 / rank(i)

# Evaluation Metric: NDCG@10

Zero-shot setting, i.e. Model trained on (A), evaluated on (B).

NDCG is then *the ratio of DCG of recommended order to DCG of ideal order*.

$$NDCG = \frac{DCG}{iDCG}$$

$$Recommendations\ Order = [2,\ 3,\ 3,\ 1,\ 2] \qquad Ideal\ Order = [3,\ 3,\ 2,\ 2,\ 1]$$

$$DCG = \frac{2}{log_2(1+1)} + \frac{3}{log_2(2+1)} + \frac{3}{log_2(3+1)} + \frac{1}{log_2(4+1)} + \frac{2}{log_2(5+1)} \approx 6.64$$

$$iDCG = \frac{3}{log_2(1+1)} + \frac{3}{log_2(2+1)} + \frac{2}{log_2(3+1)} + \frac{2}{log_2(4+1)} + \frac{1}{log_2(5+1)} \approx 7.14$$

Thus, the NDCG for this recommendation set will be:

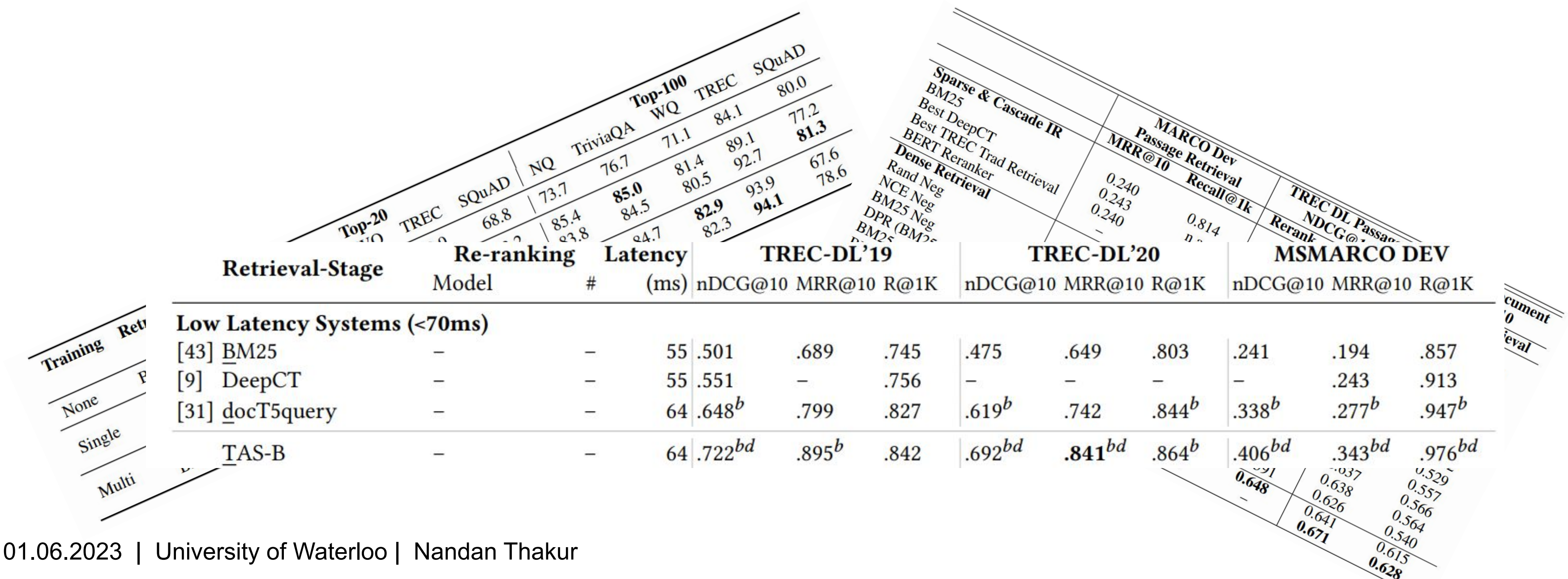
$$NDCG = \frac{DCG}{iDCG} = \frac{6.64}{7.14} \approx 0.93$$

# Retrieval System Evaluation

# How well do Dense Retrievers Perform?

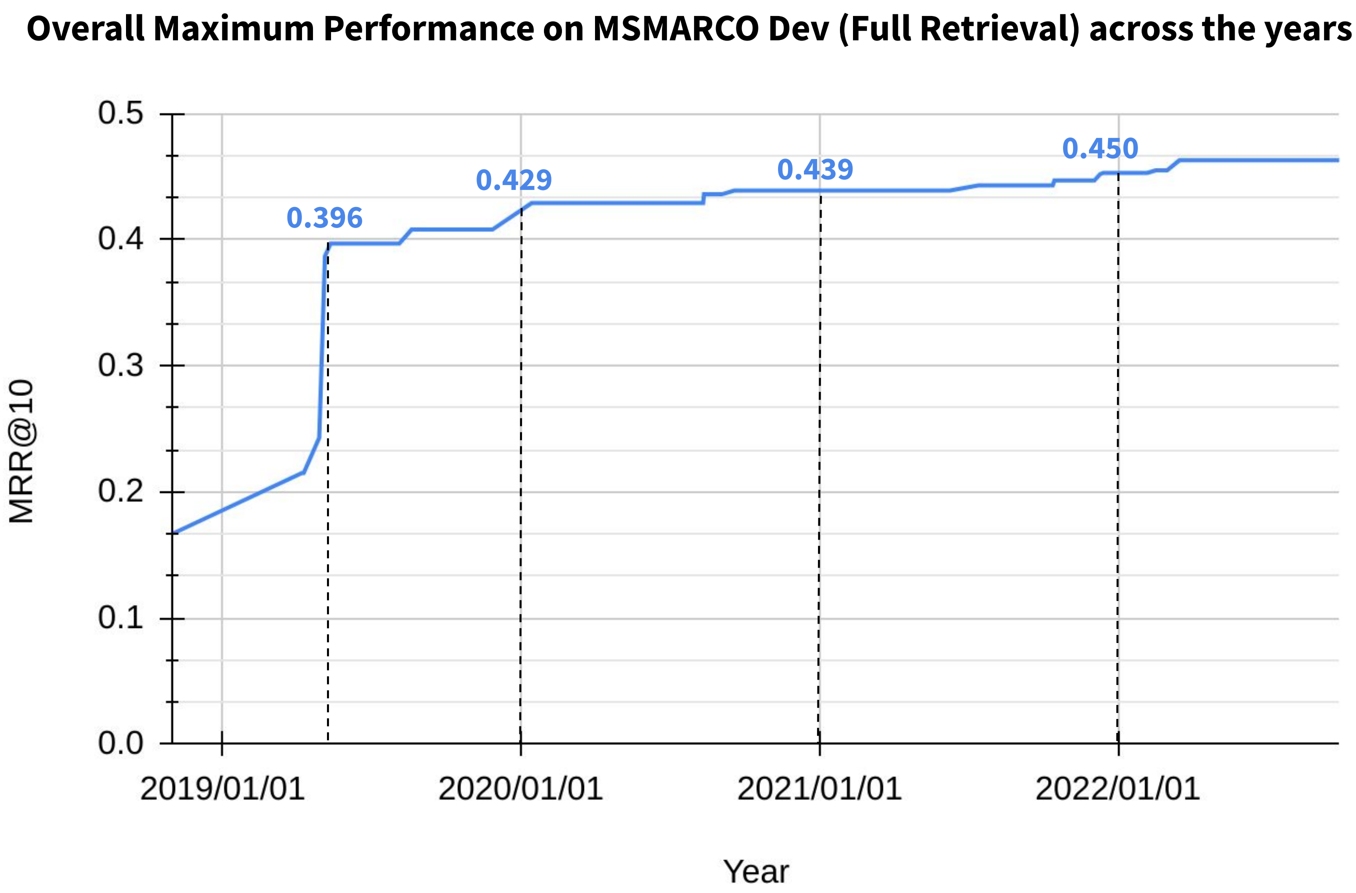
Dense Retrievers outperform BM25 on datasets with large training sizes!

| | | | |
|---|---|---|---|
| **DPR** (kharpurkin et al. 2020) | **BM25** | NQ Retrieval | **20.3** points (Top-20 Recall) |
| **ANCE** (Xiong et al. 2021) | **BM25** | MSMARCO<br>NQ Retrieval | **9.0** points (MRR@10)<br>**23.8** points (Top-20 Recall) |
| **TAS-B** (Hofstätter et al. 2021) | **BM25** | MSMARCO | **14.9** points (MRR@10) |



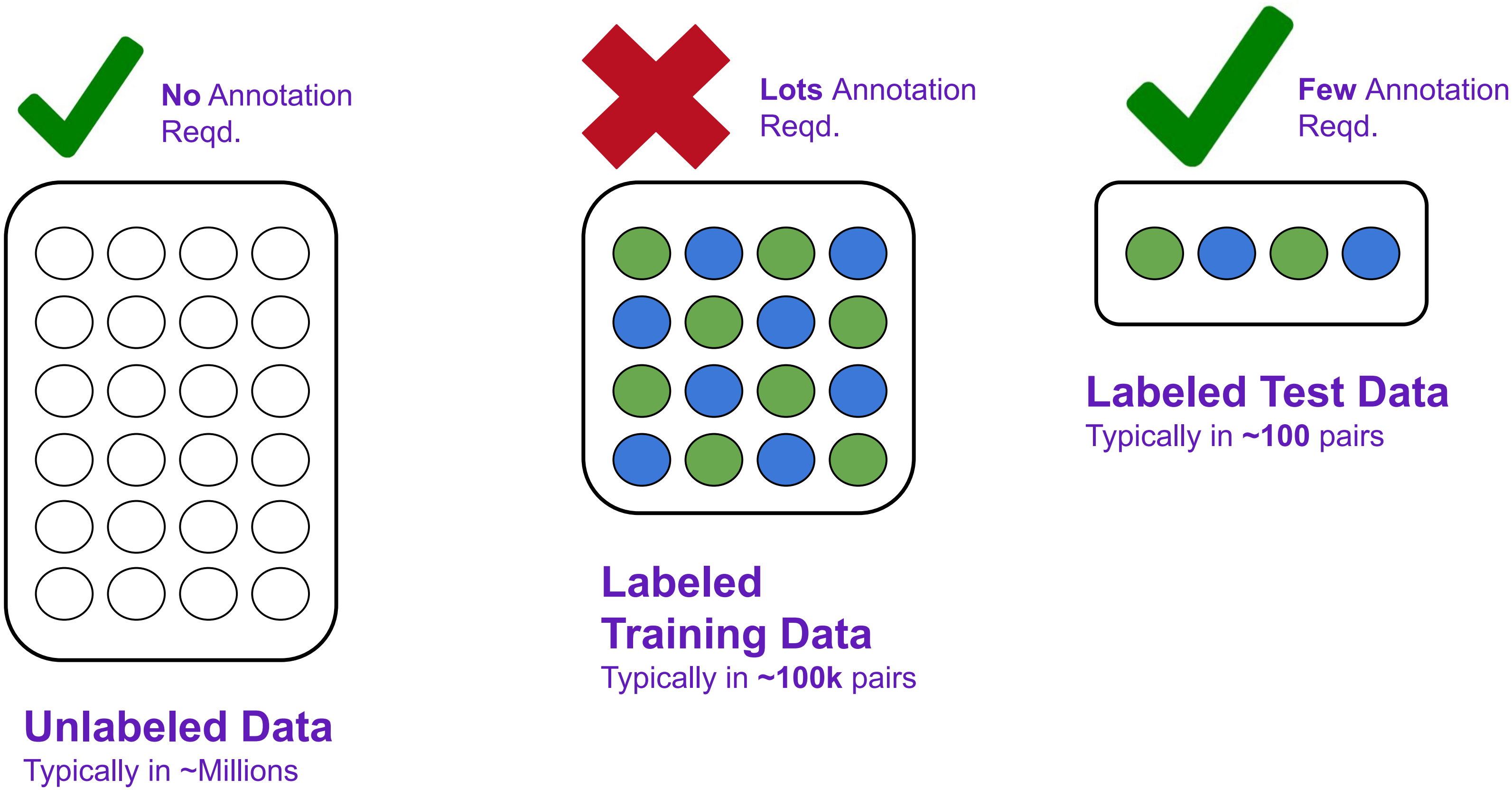
| Retrieval-Stage | Re-ranking Model | # | Latency (ms) | TREC-DL'19 nDCG@10 | TREC-DL'19 MRR@10 | TREC-DL'19 R@1K | TREC-DL'20 nDCG@10 | TREC-DL'20 MRR@10 | TREC-DL'20 R@1K | MSMARCO DEV nDCG@10 | MSMARCO DEV MRR@10 | MSMARCO DEV R@1K |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Low Latency Systems (<70ms)** | | | | | | | | | | | | |
| [43] BM25 | – | – | 55 | .501 | .689 | .745 | .475 | .649 | .803 | .241 | .194 | .857 |
| [9] DeepCT | – | – | 55 | .551 | – | .756 | – | – | – | – | .243 | .913 |
| [31] docT5query | – | – | 64 | .648$^{b}$ | .799 | .827 | .619$^{b}$ | .742 | .844$^{b}$ | .338$^{b}$ | .277$^{b}$ | .947$^{b}$ |
| TAS-B | – | – | 64 | .722$^{bd}$ | .895$^{b}$ | .842 | .692$^{bd}$ | **.841**$^{bd}$ | .864$^{b}$ | .406$^{bd}$ | .343$^{bd}$ | .976$^{bd}$ |

# MS MARCO is Saturated: Too Old too Soon!



# Why Zero-Shot Evaluation is Important?

Generating High-Quality Labeled Training Data is cumbersome!



# RQ: Can Modern Search Systems Generalize?

## Will these neural models perform well out-of-box (w/o) training?

Bi-Encoders / Cross-Encoders / Late-Interaction / Sparse Retrievers

**Trained on**

Wikipedia
The Free Encyclopedia

Natural Questions

Beir
Benchmarking IR

Evaluated on

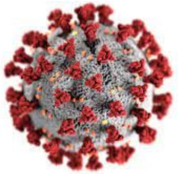
<u>Domains with Training Data</u>

Wikipedia
The Free Encyclopedia

I can answer any question you have!

Evaluated on

<u>Domains without Training Data</u>



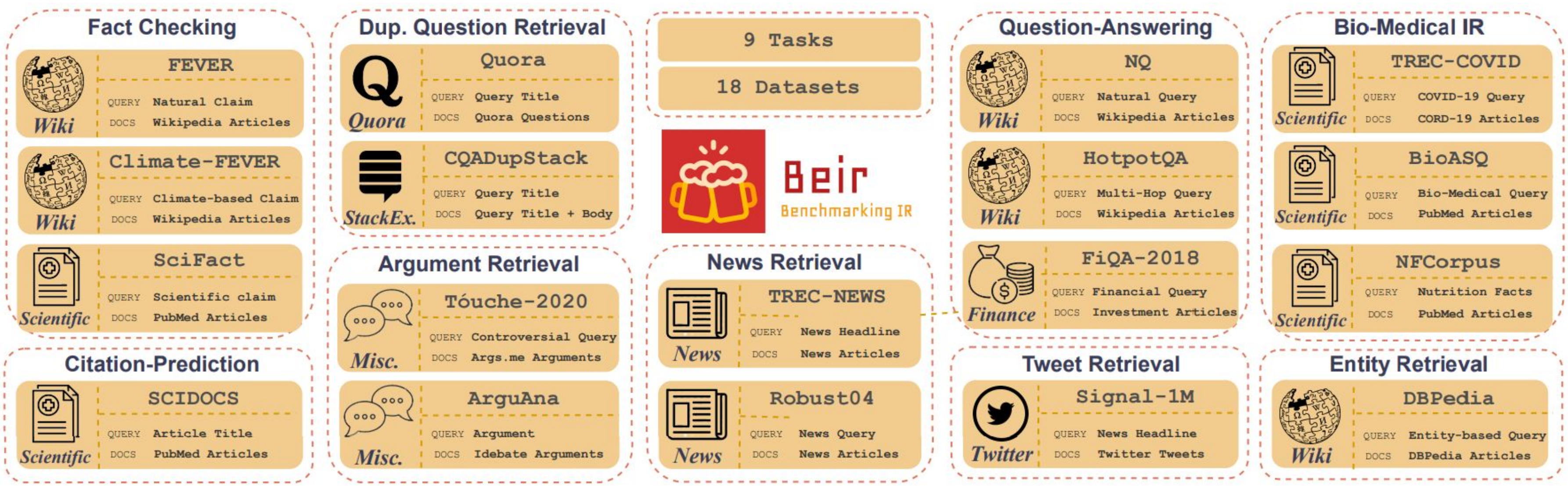
twitter

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

# The BEIR Benchmark (Thakur et al. 2021)

## Diverse, Zero-shot retrieval benchmark with 18 datasets and tasks!

- BEIR provides a **standardized benchmark** for comparison of zero-shot IR-based systems
- BEIR contains 18 **broad** datasets across **diverse** retrieval based tasks and domains
- BEIR contains evaluation datasets created using diverse annotation strategies.



# Zero-shot Retrieval Results on BEIR

| **Model (→)** | Lexical | Sparse | | | Dense | | | | Late-Interaction | Re-ranking |
|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset (↓)** | **BM25** | **DeepCT** | **SPARTA** | **docT5query** | **DPR** | **ANCE** | **TAS-B** | **GenQ** | **ColBERT** | **BM25+CE** |
| MS MARCO | 0.228 | 0.296‡ | 0.351‡ | 0.338‡ | 0.177 | 0.388‡ | 0.408‡ | 0.408‡ | **0.425‡** | 0.413‡ |
| TREC-COVID | 0.656 | 0.406 | 0.538 | 0.713 | 0.332 | 0.654 | 0.481 | 0.619 | 0.677 | **0.757** |
| BioASQ | 0.465 | 0.407 | 0.351 | 0.431 | 0.127 | 0.306 | 0.383 | 0.398 | 0.474 | **0.523** |
| NFCorpus | 0.325 | 0.283 | 0.301 | 0.328 | 0.189 | 0.237 | 0.319 | 0.319 | 0.305 | **0.350** |
| NQ | 0.329 | 0.188 | 0.398 | 0.399 | 0.474‡ | 0.446 | 0.463 | 0.358 | 0.524 | **0.533** |
| HotpotQA | 0.603 | 0.503 | 0.492 | 0.580 | 0.391 | 0.456 | 0.584 | 0.534 | 0.593 | **0.707** |
| FiQA-2018 | 0.236 | 0.191 | 0.198 | 0.291 | 0.112 | 0.295 | 0.300 | 0.308 | 0.317 | **0.347** |
| Signal-1M (RT) | 0.330 | 0.269 | 0.252 | 0.307 | 0.155 | 0.249 | 0.289 | 0.281 | 0.274 | **0.338** |
| TREC-NEWS | 0.398 | 0.220 | 0.258 | 0.420 | 0.161 | 0.382 | 0.377 | 0.396 | 0.393 | **0.431** |
| Robust04 | 0.408 | 0.287 | 0.276 | 0.437 | 0.252 | 0.392 | 0.427 | 0.362 | 0.391 | **0.475** |
| ArguAna | 0.315 | 0.309 | 0.279 | 0.349 | 0.175 | 0.415 | 0.429 | **0.493** | 0.233 | 0.311 |
| Touché-2020 | **0.367** | 0.156 | 0.175 | 0.347 | 0.131 | 0.240 | 0.162 | 0.182 | 0.202 | 0.271 |
| CQADupStack | 0.299 | 0.268 | 0.257 | 0.325 | 0.153 | 0.296 | 0.314 | 0.347 | 0.350 | **0.370** |
| Quora | 0.789 | 0.691 | 0.630 | 0.802 | 0.248 | 0.852 | 0.835 | 0.830 | **0.854** | 0.825 |
| DBPedia | 0.313 | 0.177 | 0.314 | 0.331 | 0.263 | 0.281 | 0.384 | 0.328 | 0.392 | **0.409** |
| SCIDOCS | 0.158 | 0.124 | 0.126 | 0.162 | 0.077 | 0.122 | 0.149 | 0.143 | 0.145 | **0.166** |
| FEVER | 0.753 | 0.353 | 0.596 | 0.714 | 0.562 | 0.669 | 0.700 | 0.669 | 0.771 | **0.819** |
| Climate-FEVER | 0.213 | 0.066 | 0.082 | 0.201 | 0.148 | 0.198 | 0.228 | 0.175 | 0.184 | **0.253** |
| SciFact | 0.665 | 0.630 | 0.582 | 0.675 | 0.318 | 0.507 | 0.643 | 0.644 | 0.671 | **0.688** |
| Avg. Performance vs. BM25 | | **- 27.9%** | **- 20.3%** | **+ 1.6%** | **- 47.7%** | **- 7.4%** | **- 2.8%** | **- 3.6%** | **+ 2.5%** | **+ 11%** |

**BM25 (Lexical)**

BM25 is an overall strong system. It doesn't require to be trained.

**Cross-Encoders (Rerank)**

Reranking Models generalize best. They outperform BM25 on **11/18** retrieval datasets.

**Bi-Encoders (Dense)**

Dense models suffer from generalization. They outperform BM25 on **7/18** datasets.

# Zero-shot Retrieval Results on BEIR

| Dataset | BM25† | BM25 | DocT5 | Baselines SPLADEv2-distil |
|---|---|---|---|---|
| arguana | 42.25 | 41.42 | 46.90 | 47.91 |
| bioasq | 47.67 | 46.46 | 43.10 | 50.80 |
| climate-fever | 21.32 | 21.29 | 20.10 | 23.53 |
| cqadupstack | 28.53 | 29.87 | 32.50 | **35.01** |
| dbpedia-entity | 32.26 | 31.28 | 33.10 | **43.50** |
| fever | 74.35 | 75.31 | 71.40 | 78.62 |
| fiqa | 24.30 | 23.61 | 29.10 | **33.61** |
| hotpotqa | 60.13 | 60.28 | 58.00 | 68.44 |
| nfcorpus | 32.67 | 32.55 | 32.80 | 33.43 |
| nq | 32.87 | 32.86 | 39.90 | **52.08** |
| quora | 74.71 | 78.86 | 80.20 | 83.76 |
| robust04 | 41.91 | 40.84 | 43.70 | 46.75 |
| scidocs | 15.83 | 15.81 | 16.20 | 15.79 |
| scifact | 66.28 | 66.47 | 67.50 | 69.25 |
| signal1m | 32.69 | 33.05 | 30.70 | 26.56 |
| trec-covid | 71.23 | 65.59 | 71.30 | 71.04 |
| trec-news | 40.33 | 39.77 | 42.00 | 39.18 |
| webis-touche2020 | 35.40 | **36.73** | 34.70 | 27.18 |
| Average | 43.04 | 42.89 | 44.07 | 47.02 |
| Best on | 0 | 1 | 0 | 4 |

| Corpus | Models without Distillation | | | | Models with Distillation | | | |
|---|---|---|---|---|---|---|---|---|
| | ColBERT | DPR-M | ANCE | MoDIR | TAS-B | RocketQAv2 | SPLADEv2 | ColBERTv2 |
| BEIR Search Tasks (nDCG@10) | | | | | | | | |
| **DBPedia** | 39.2 | 23.6 | 28.1 | 28.4 | 38.4 | 35.6 | 43.5 | **44.6** |
| **FiQA** | 31.7 | 27.5 | 29.5 | 29.6 | 30.0 | 30.2 | 33.6 | **35.6** |
| **NQ** | 52.4 | 39.8 | 44.6 | 44.2 | 46.3 | 50.5 | 52.1 | **56.2** |
| **HotpotQA** | 59.3 | 37.1 | 45.6 | 46.2 | 58.4 | 53.3 | **68.4** | 66.7 |
| **NFCorpus** | 30.5 | 20.8 | 23.7 | 24.4 | 31.9 | 29.3 | 33.4 | **33.8** |
| **T-COVID** | 67.7 | 56.1 | 65.4 | 67.6 | 48.1 | 67.5 | 71.0 | **73.8** |
| **Touché (v2)** | - | - | - | - | - | 24.7 | **27.2** | 26.3 |
| BEIR Semantic Relatedness Tasks (nDCG@10) | | | | | | | | |
| **ArguAna** | 23.3 | 41.4 | 41.5 | 41.8 | 42.7 | 45.1 | **47.9** | 46.3 |
| **C-FEVER** | 18.4 | 17.6 | 19.8 | 20.6 | 22.8 | 18.0 | **23.5** | 17.6 |
| **FEVER** | 77.1 | 58.9 | 66.9 | 68.0 | 70.0 | 67.6 | **78.6** | **78.5** |
| **Quora** | 85.4 | 84.2 | 85.2 | **85.6** | 83.5 | 74.9 | 83.8 | 85.2 |
| **SCIDOCS** | 14.5 | 10.8 | 12.2 | 12.4 | 14.9 | 13.1 | **15.8** | 15.4 |
| **SciFact** | 67.1 | 47.8 | 50.7 | 50.2 | 64.3 | 56.8 | **69.3** | **69.3** |

**Sparse Retrieval (SPLADEv2)**

Sparse models are able to "generalise". They outperform BM25 on **12/18** datasets.

**Late Interaction (ColBERTv2)**

Late Interaction also "generalizes" well and outperforms BM25 on **11/13** datasets evaluated.

# Efficiency and Memory Comparison on BEIR

Retrieval Latency (in ms) and Index Sizes (in GB)

| DBPedia [19] (1 Million) | | | Retrieval Latency | | Index |
|---|---|---|---|---|---|
| Rank | Model | Dim. | GPU | CPU | Size |
| (1) | BM25+CE | – | 450ms | 6100ms | 0.4GB |
| (2) | ColBERT | 128 | 350ms | – | 20GB |
| (3) | docT5query | – | – | 30ms | 0.4GB |
| (4) | BM25 | – | – | 20ms | 0.4GB |
| (5) | TAS-B | 768 | 14ms | 125ms | 3GB |
| (6) | GenQ | 768 | 14ms | 125ms | 3GB |
| (7) | ANCE | 768 | 20ms | 275ms | 3GB |
| (8) | SPARTA | 2000 | – | 20ms | 12GB |
| (9) | DeepCT | – | – | 25ms | 0.4GB |
| (10) | DPR | 768 | 19ms | 230ms | 3GB |

How to see the table:
Smaller the better!

**BM25 (Lexical)**

BM25 is overall **fast** and **efficient**. They require small indexes.

**Cross-Encoders (Rerank)**

Rerankers are **slow** at retrieval. They can also produce **bulky** indexes for retrieval.

**Bi-Encoders (Dense)**

Dense retrievers are **fast** and **efficient**. They consume less memory with **small** indexes.

Ref: Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeurIPS 2021 Dataset and Benchmark Track.

# Interesting Future Directions in IR

# (1) How to Improve Dual Encoder Generalization?

As training data is scarce, focus is on unsupervised techniques!

## Unsupervised Domain Adaptation

- Generate synthetic queries and use query-passage pairs across each domain.
- Trains a model separately across each domain/dataset.

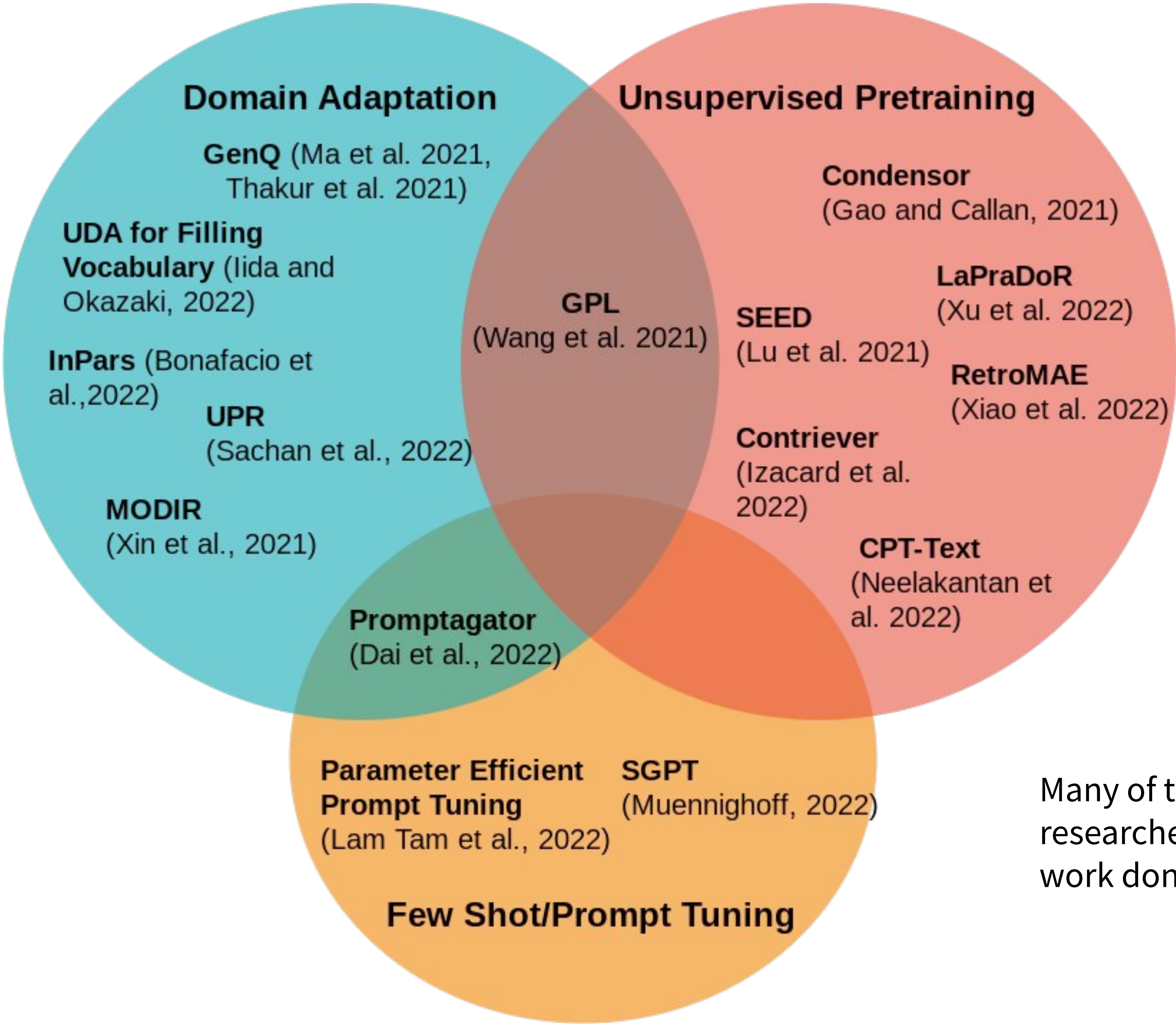
## Unsupervised Pre-training

- Pretrains Bi-Encoder usually in a self-supervised fashion across (a lot) of raw data.
- Few techniques also involve a light decoder setup, training in an autoencoder setup.

## Few-shot Training/Prompt Tuning

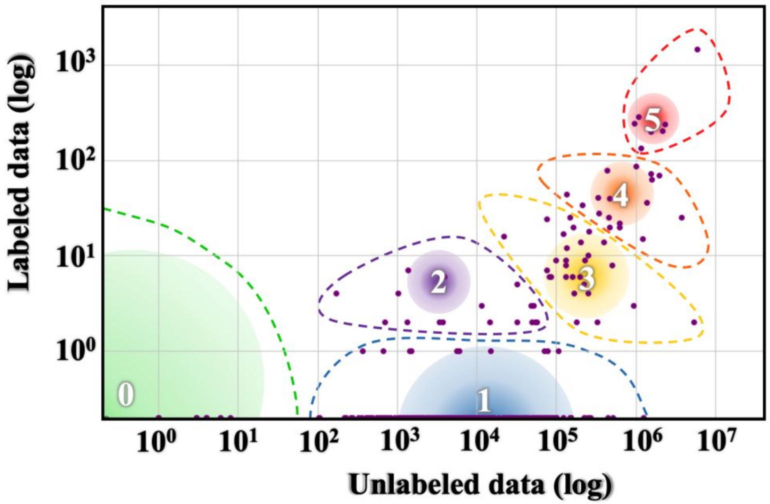
- Few-shot training involves training Bi-Encoder with only a handful of training examples.
- Prompt-Tuning involves changing weights of prompt layers and keeping the LM unchanged.

# Summary of Recent Works to Improve Dual Encoder Generalization




Many of these ideas (by other researchers) got inspired by work done in BEIR :)

# (2) Multilingual IR: Providing Information Access to Everyone!




- Prior research in IR is heavily focused across a single language: **English**.
- There are collectively over **two-three billion** native speakers around the world who speak non-English languages.
- These languages have **diverse typologies**, originate from many different language families, and often contain varying amounts of available resources.

| Class | 5 Example Languages | #Langs | #Speakers | % of Total Langs |
|---|---|---|---|---|
| 0 | Dahalo, Warlpiri, Popoloca, Wallisian, Bora | 2191 | 1.0B | 88.17% |
| 1 | Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo | 222 | 1.0B | 8.93% |
| 2 | Zulu, Konkani, Lao, Maltese, Irish | 19 | 300M | 0.76% |
| 3 | Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew | 28 | 1.1B | 1.13% |
| 4 | Russian, Hungarian, Vietnamese, Dutch, Korean | 18 | 1.6B | 0.72% |
| 5 | English, Spanish, German, Japanese, French | 7 | 2.5B | 0.28% |

# What is Challenging in Multilingual Retrieval?

## Information Scarcity

Information, i.e. documents available in non-English languages, are less than English.

ডেট্রয়েট ইন্সটিটিউট অফ আর্ট এর প্রতিষ্ঠাতা কে ?
(Who is the founder of Detroit Institute of Art?)

**William Reinhold Valentiner** (May 2, 1880 – September 6, 1958) was a German-American art historian ... founded Detroit Museum of Art in 1885

William Reinhold Valentiner (en.wiki)

デトロイト美術館は1885年に開館されたアメリカ合衆国ミシガン州デトロイトにある美術館。

デトロイト美術館 (Detroit Institute of Arts) (ja.wiki)

## Information Asymmetry

Queries can be about culturally specific topics (e.g., *Maacher Jhol* in Bengali)

速水堅曹はどこで製糸技術を学んだ？ (Where did Kenso Hayami learn silk-reeling technique?)

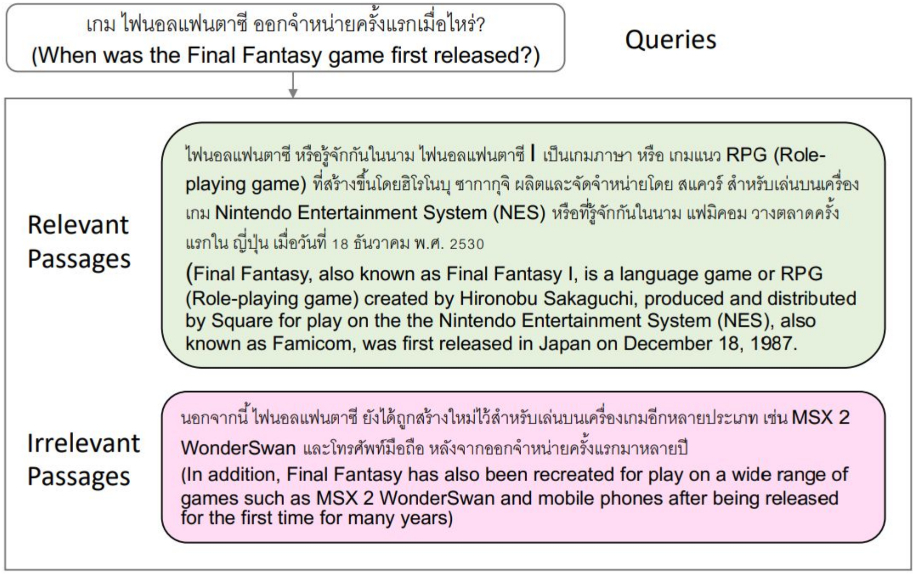
速水堅曹は藩営前橋製糸所を前橋に開設。カスパル・ミュラーから直接、器械製糸技術を学び
(Kenso Hayami founded Hanei Maebashi Silk Mill and learned instrumental silk reeling techniques directly from Caspal Müller)

速水堅曹 (Kenso Hayami) (ja.wiki)

# MIRACL Benchmark (in collaboration with Huawei)

MIRACL

Multilingual Information Retrieval Across a Continuum of Languages



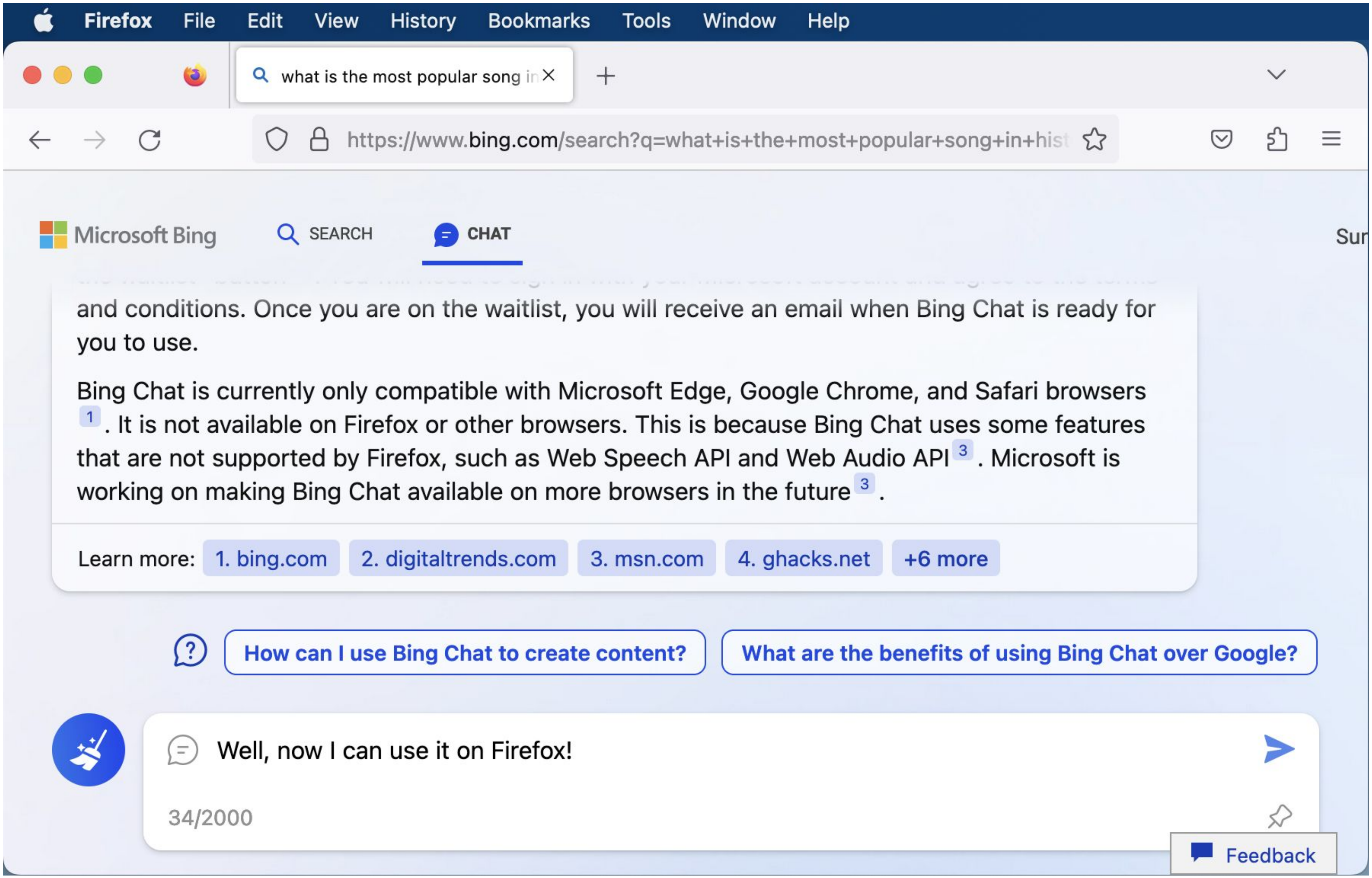

th.wikipedia

- **Scarcity** resources available for mono and cross-lingual retrieval evaluation.

- The community has progressed immensely on English, however lacks behind on the multilingual front due to lack of **training data** and **standard evaluation** benchmarks.

- For **MIRACL**, we annotated datasets in each language (e.g., **TyDi QA**).
  - Better reflect speakers' **true interests** and **linguistic phenomena**
  - Hired over **40 native speakers** for the wide-scale annotation study
  - Performance will **lead to different insights** across languages, as each language has its own linguistic features.

# (3) Generative Search and QA



# (3) Generative Search and QA

**Fusion-in-Decoder (FiD) Method by Izacard et al. 2021**

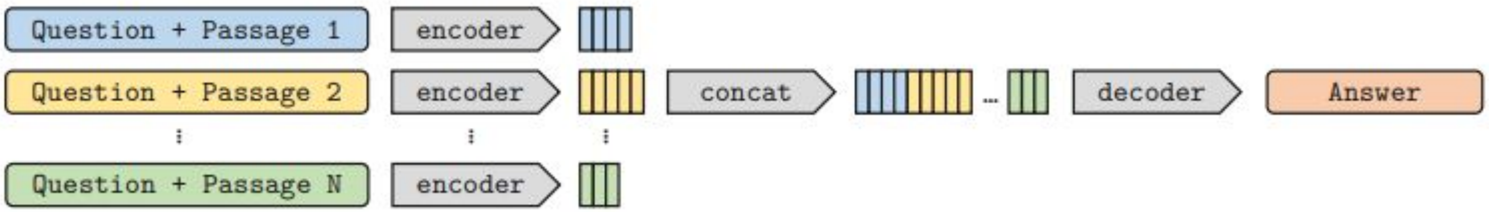
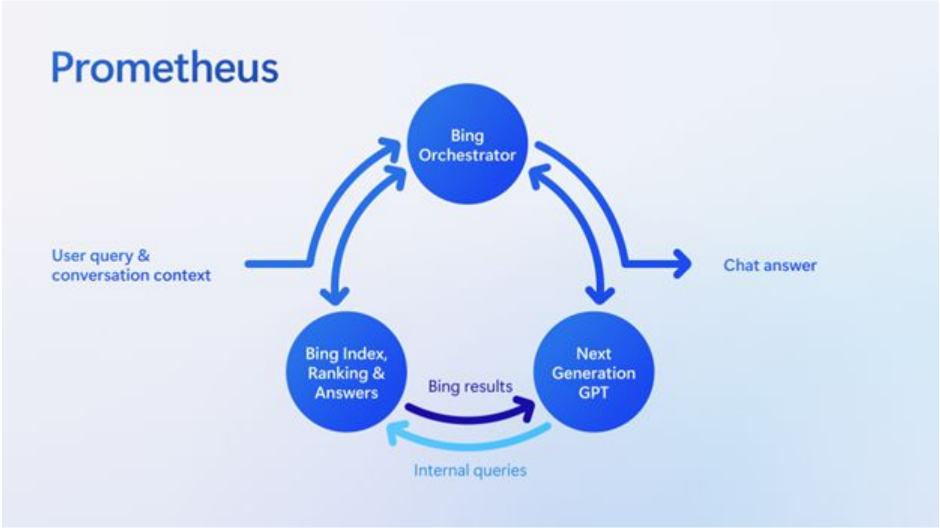



Figure 2: Architecture of the Fusion-in-Decoder method.

**Building the New Bing. Blogpost. Microsoft 2023.**



# Thank you for listening!