Stanford University, 17th November 2022

Heterogenous Benchmarking
The Key to Enable Meaningful Progress in IR Research

Nandan Thakur
PhD Student
Current: Student Researcher @ Google Research, MTV

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, MTV.
- **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

A Brief history of NLP/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

Popular Benchmarks in NLP and ML

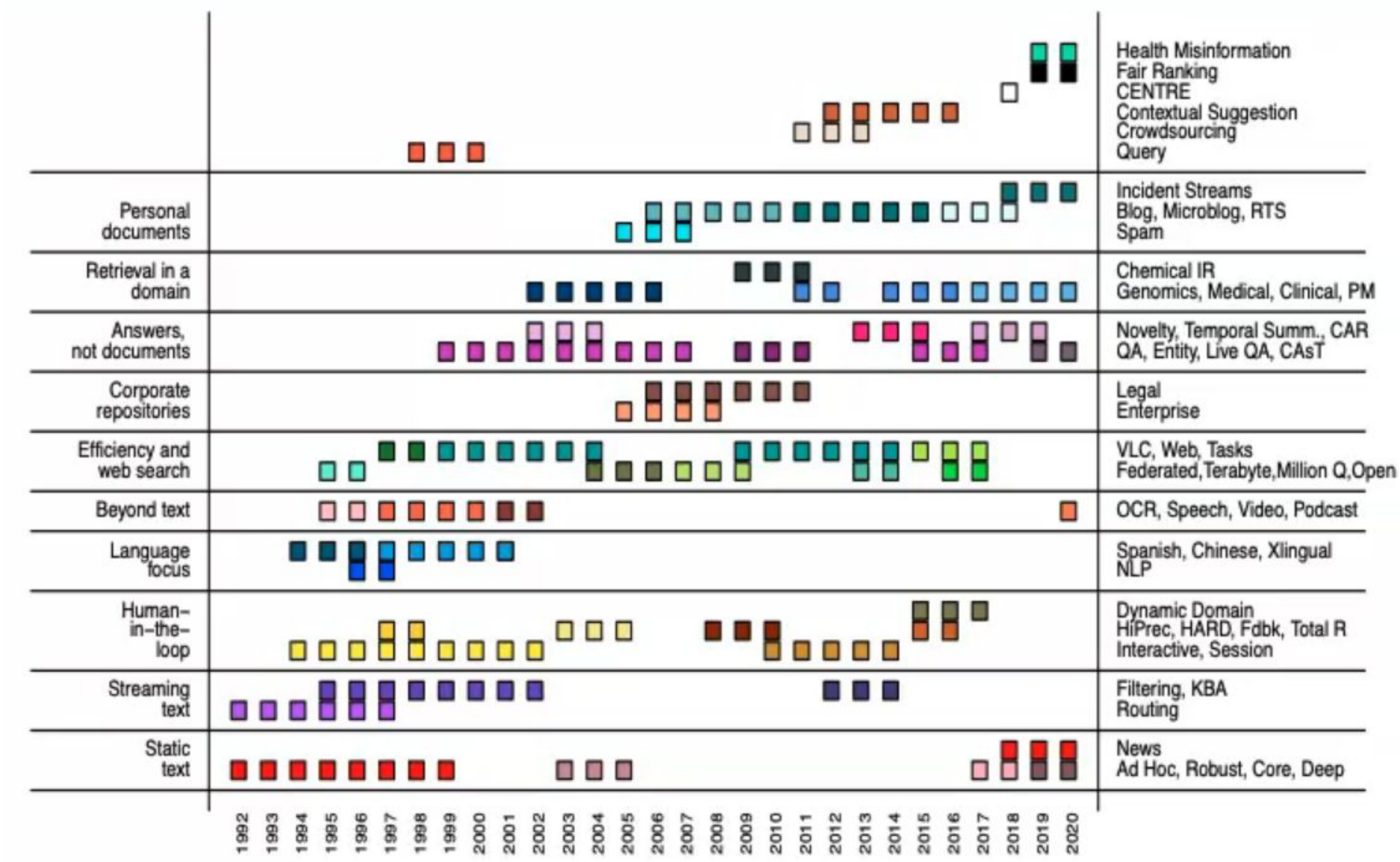
Corpus	|Train|	|Test|	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

GLUE

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task	Metric	Domain
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI	Acc.	Misc.
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase	Acc.	Wiki / Quora
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS	F1	Misc.
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER	F1	Wikipedia
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction	F1 / EM	Wikipedia
	MLQA			4,517–11,590	translations	7	Span extraction	F1 / EM	Wikipedia
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction	F1 / EM	Wikipedia
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval	F1	Wiki / news
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval	Acc.	misc.

XTREME

TREC Suite: History of IR Benchmarking



Information Retrieval (Recap)

What is Information Retrieval?

Which football club Lionel Messi plays for?

natural language query

OR

Messi football club

keyword-based query

WIKIPEDIA
The Free Encyclopedia

5.5M Articles

Lionel Messi

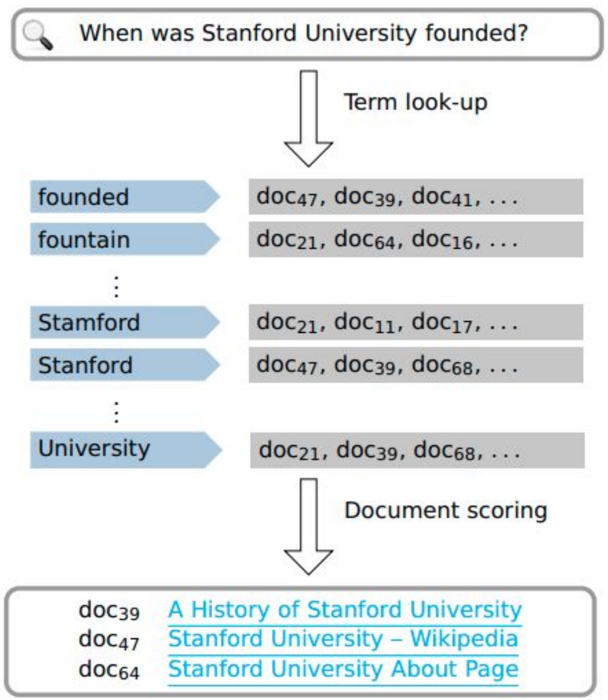
Lionel Andrés Messi (born 24 June 1987), also known as Leo Messi, is an Argentine professional footballer who plays as a forward for Ligue 1 club **Paris Saint-Germain** and captains the Argentina national team. Often considered the best player in the world and widely regarded as one of the greatest 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.

Information Retrieval is present everywhere!

Ubiquitous
present, appearing, or found everywhere.

BM25 (Bag of Words)

Keyword based Search: Exact Match of Words



$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

BM25 parameters

Elasticsearch: k1 = 1.2, b = 0.8

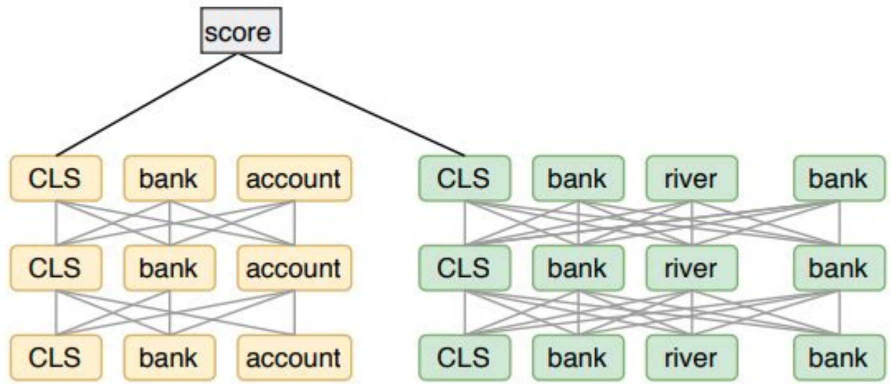
Anserini (Lucene): k1 = 0.9, b = 0.4

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

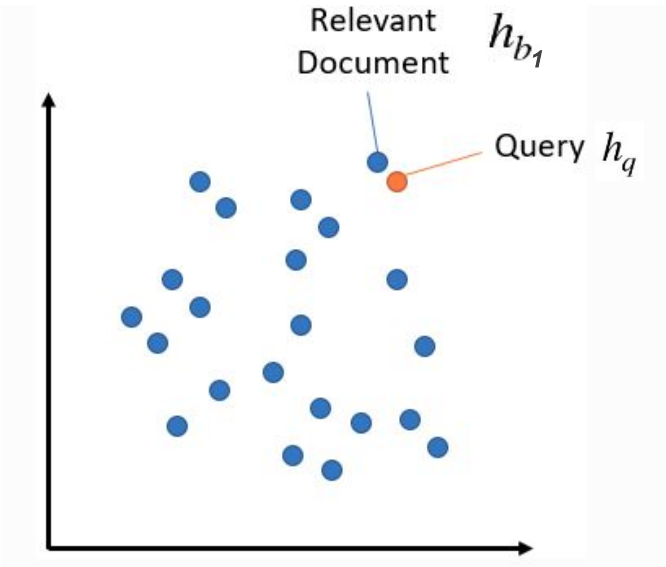
Dense Retrieval with Bi-Encoders

Mapping Individual Text to a fixed dimensional embedding!

$$\mathrm{sim}(q,p) = E_Q(q)^{\intercal} E_P(p).$$



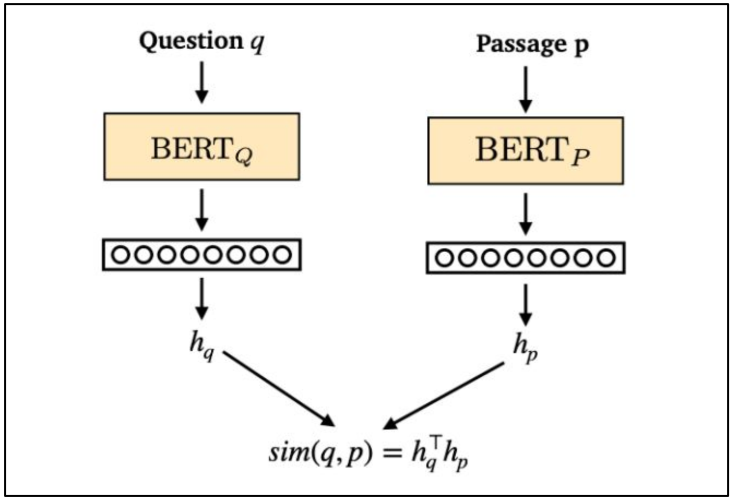
(b) Dense Retrievers (e.g., DPR)



2D Vector Space

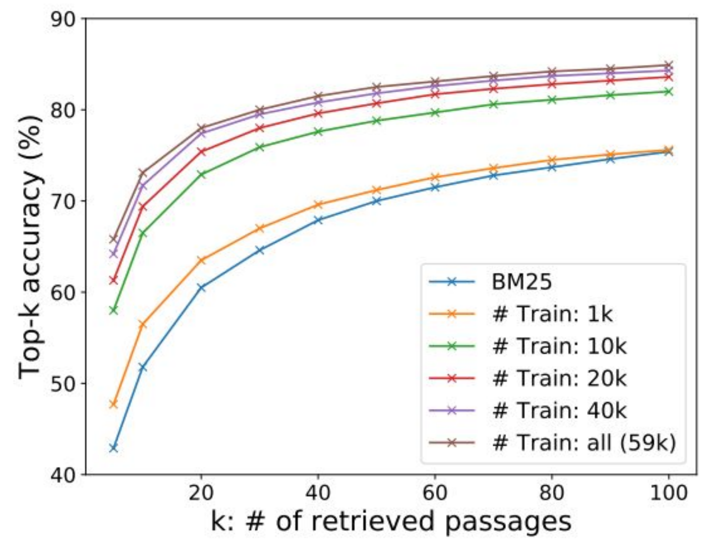
- Passage Embeddings can be precomputed using BERT and stored!
- Fast and efficient at runtime, ideal for a practical system!

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



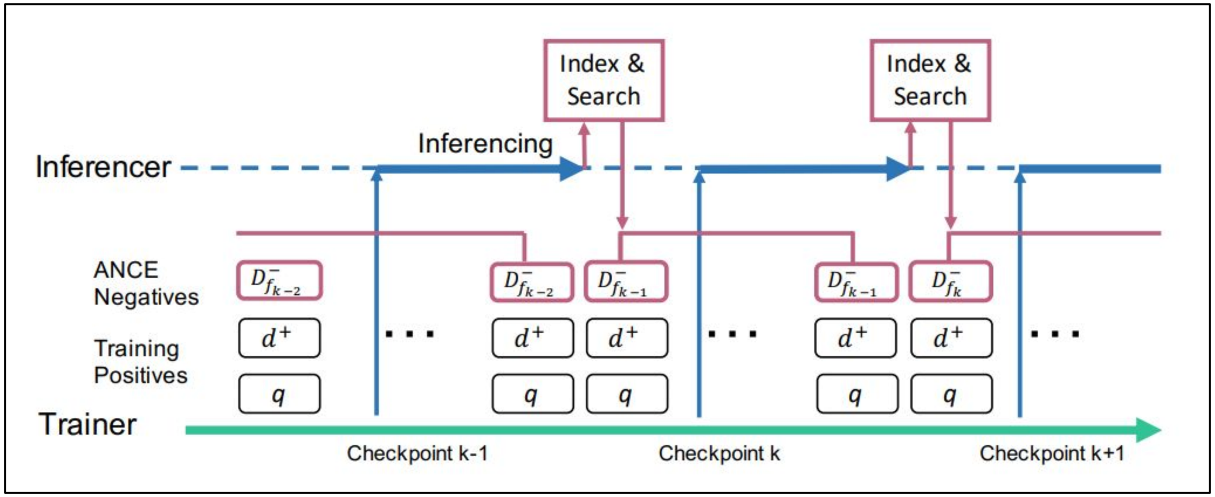
$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}$$

Natural Questions (Kwiatkowski et al., 2019)

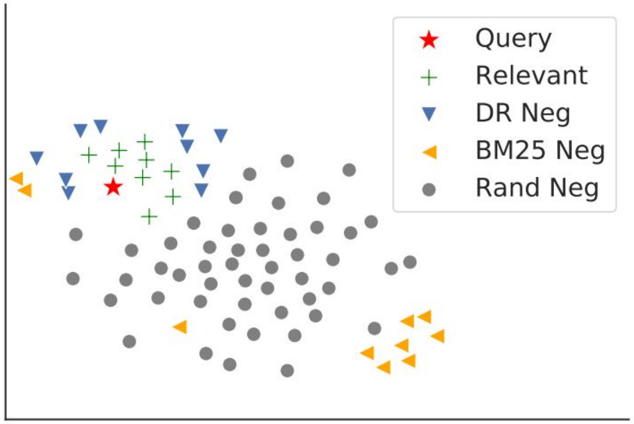


DPR can outperform a traditional IR system (such as BM25) using ~1k train examples.

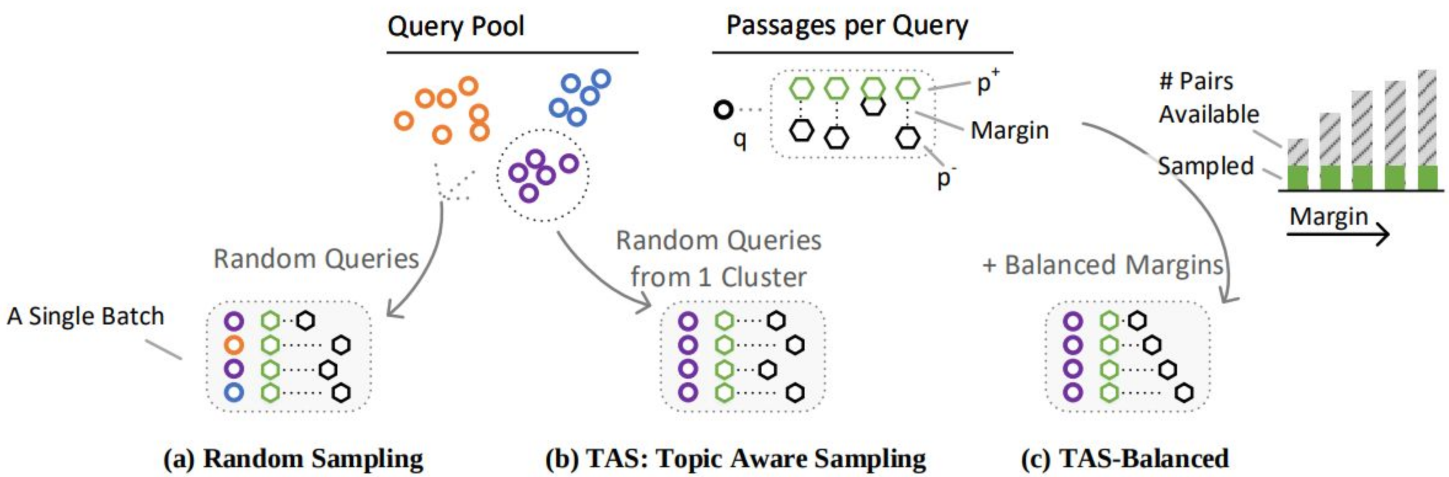
ANCE: Approximate Nearest Neighbor Negative Contrastive Learning (Xiong et al. 2021)

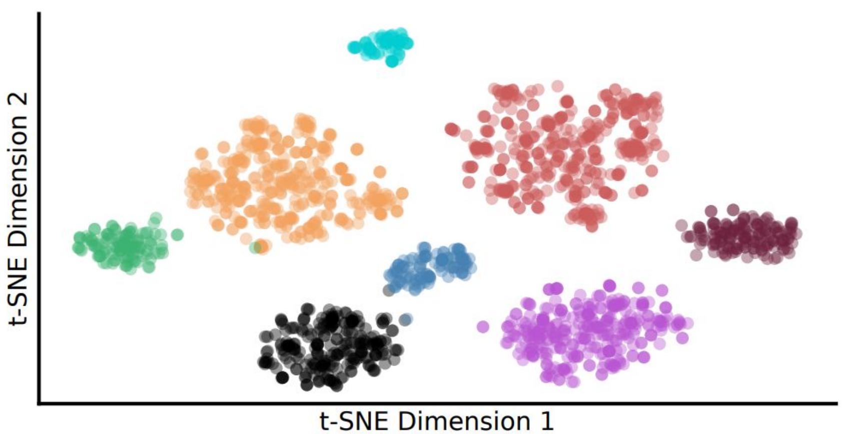


$$\theta^* = \text{argmin}_\theta \sum_q \sum_{d^+ \in D^+} \sum_{d^- \in D^-_{\text{ANCE}}} l(f(q, d^+), f(q, d^-)),$$



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





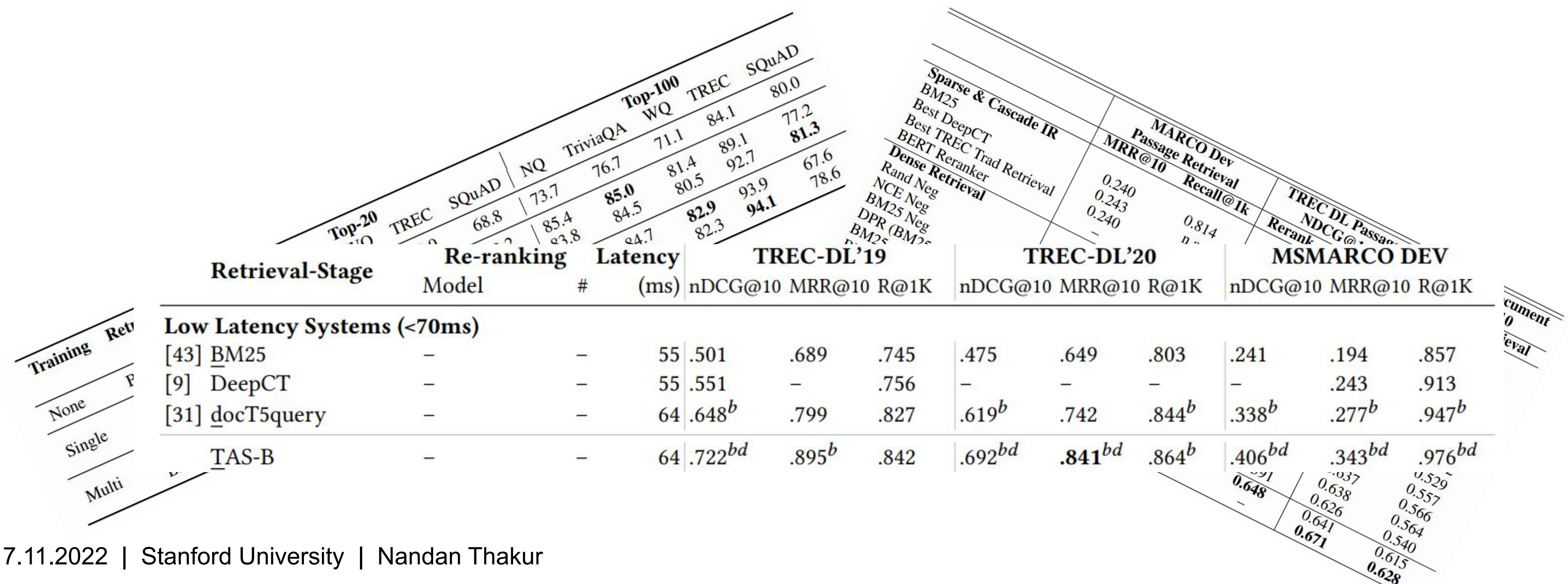
$$\mathcal{L}_{Pair}(Q, P^+, P^-) = \mathrm{MSE}(M_s(Q, P^+) - M_s(Q, P^-), M_t(Q, P^+) - M_t(Q, P^-))$$

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

How do Bi-Encoders Perform on Retrieval?

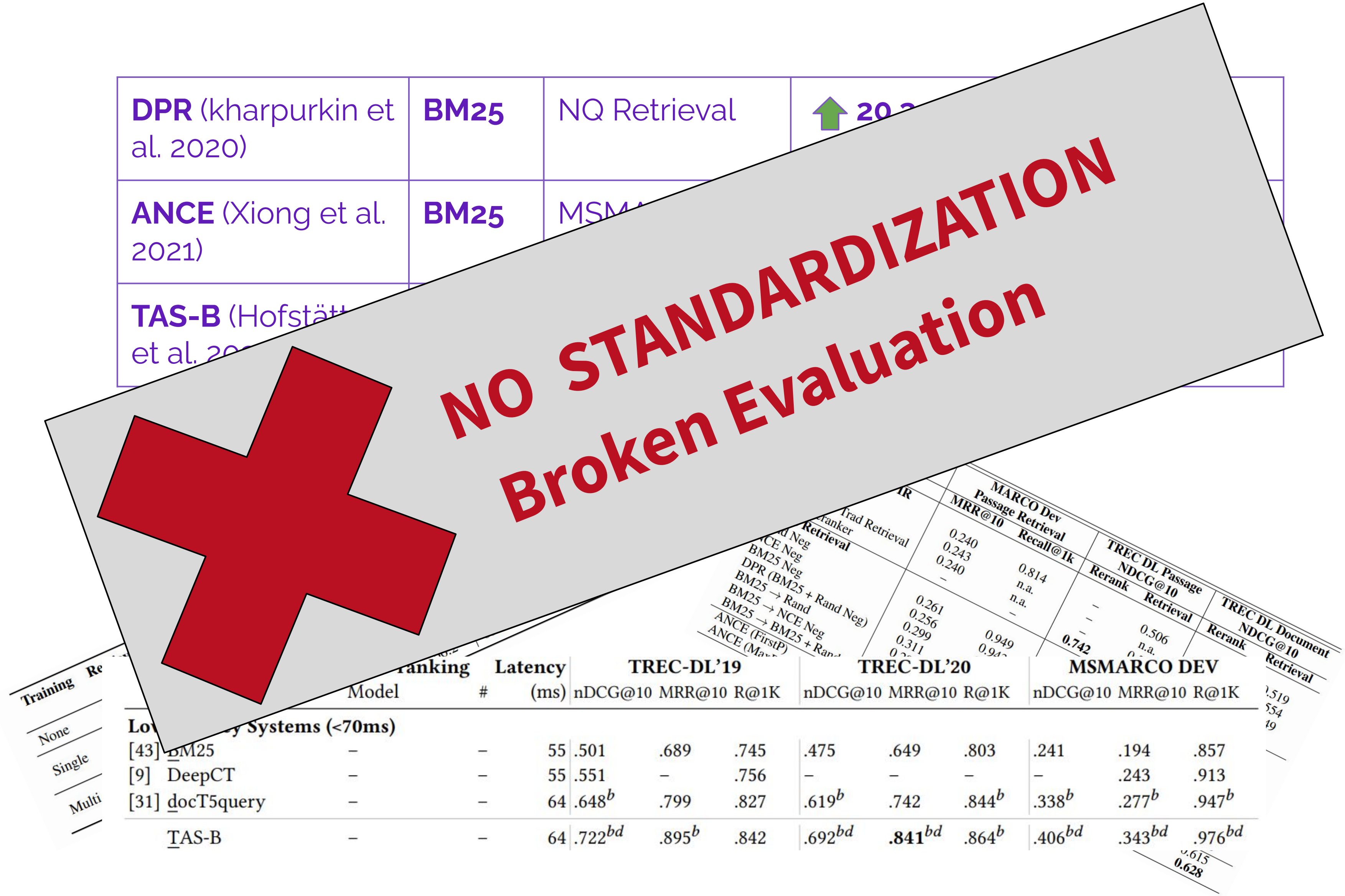
Bi-Encoders outperform BM25 across the datasets!

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

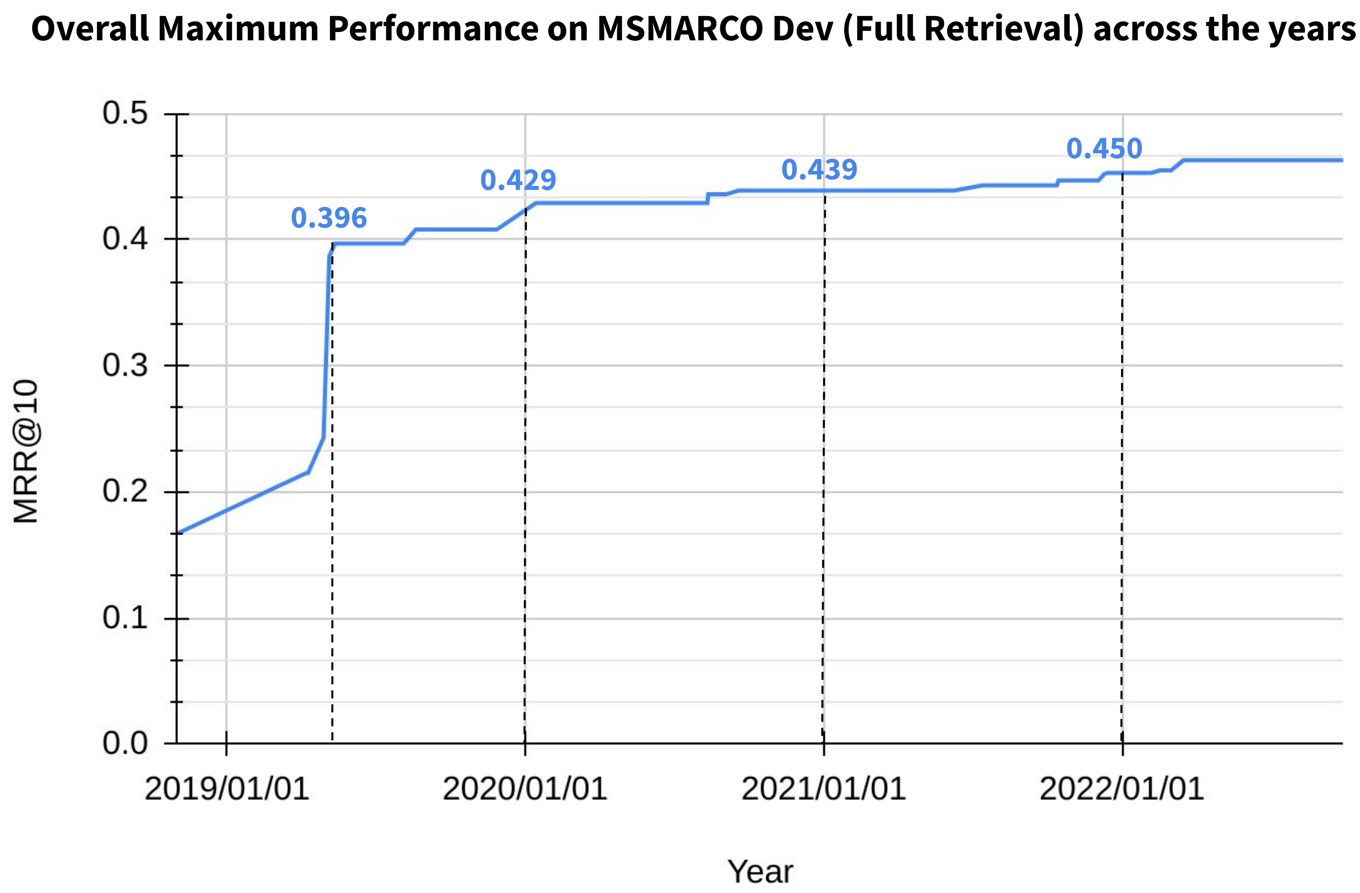


Retrieval-Stage	Re-ranking Model	#	Latency (ms)	TREC-DL'19 nDCG@10	TREC-DL'19 MRR@10	TREC-DL'19 R@1K	TREC-DL'20 nDCG@10	TREC-DL'20 MRR@10	TREC-DL'20 R@1K	MSMARCO DEV nDCG@10	MSMARCO DEV MRR@10	MSMARCO DEV R@1K
Low Latency Systems (<70ms)												
[43] BM25	–	–	55	.501	.689	.745	.475	.649	.803	.241	.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	$.692^{bd}$	$\mathbf{.841}^{bd}$	$.864^{b}$	$.406^{bd}$	$.343^{bd}$	$.976^{bd}$

Performance of Bi-Encoders >> BM25



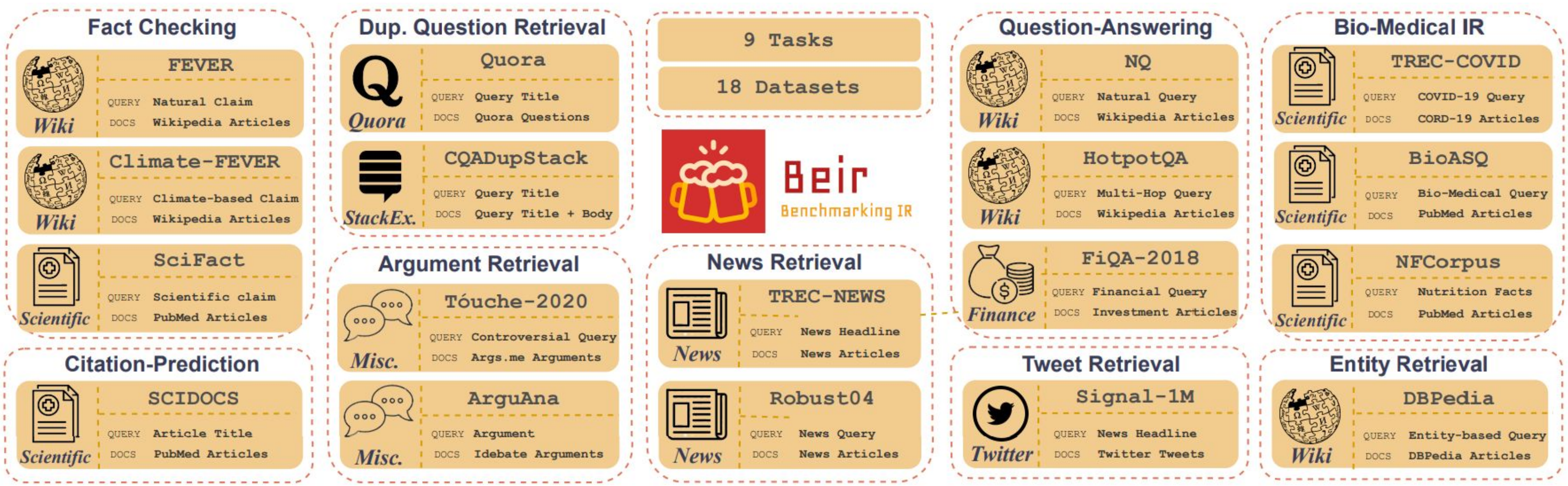
MS MARCO is Saturated: Too Old too Soon!



Solution: The BEIR Benchmark (Thakur et al. 2021)

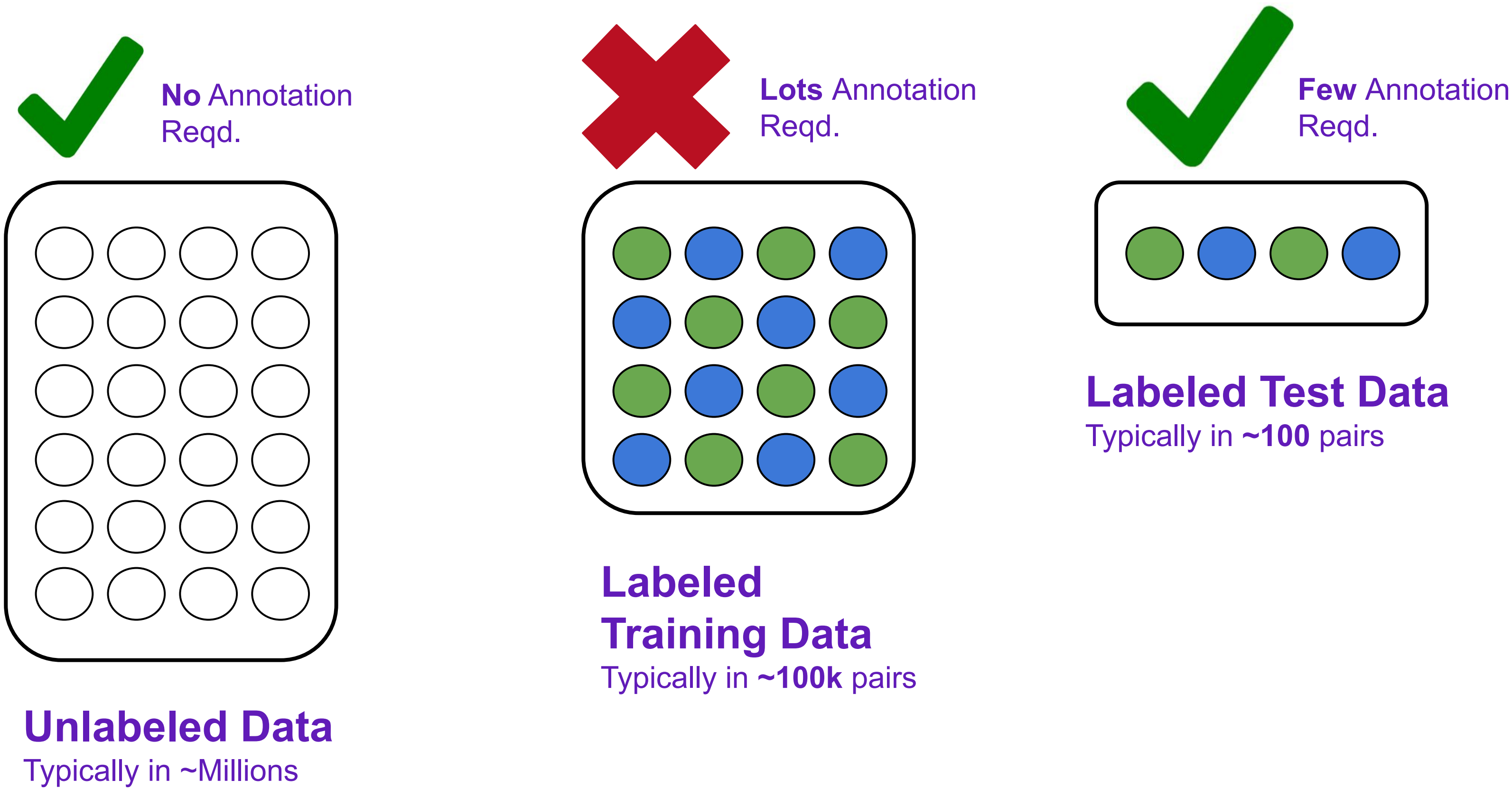
Diverse, Zero-shot retrieval benchmark with 18 datasets and tasks!

- 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.



Why Zero-Shot Evaluation in IR is Necessary?

Generating High-Quality Labeled Training Data is cumbersome!



How Well do Bi-Encoders Generalize?

Within the same domain, Bi-Encoders outperform BM25!

In-Domain Evaluation

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

Overall Dense Retriever performances >> BM25

How Well do Bi-Encoders Generalize?

On zero-shot evaluation, BM25 still a strong benchmark!

Zero-Shot Evaluation on BEIR Benchmark

DPR (kharpurkin et al. 2020)	**BM25**	BEIR (18 Datasets Avg.)	↓ **18.6** points (NDCG@10)
ANCE (Xiong et al. 2021)	**BM25**	BEIR (18 Datasets Avg.)	↓ **3.4** points (NDCG@10)
TAS-B (Hofstätter et al. 2021)	**BM25**	BEIR (18 Datasets Avg.)	↓ **0.8** points (NDCG@10)

Overall **BM25 >> Zero-shot Dense Retriever**

I.e., BM25 is still an effective and a strong out-of-domain baseline for zero-shot evaluation.

Why do Bi-Encoders Suffer from Zero-shot Generalization?

Curse of the Unknowns

- How does Bi-Encoders handle **unknown words?**
  - Not Seen during fine-tuning
  - Not seen during pre-training

- Where to put **new words** in the vector space?
  - XLNet
  - ColBERT
  - BEIR

- How to learn semantic **word relationships** with unknown words?
  - Coronavirus ⇐⇒ COVID-19 ⇐⇒ SARS-Cov-2
  - DPR ⇐⇒ ANCE ⇐⇒ TAS-B

How to Improve Bi-Encoder Generalization?

Scaling Law: LLM based Retrievers are better generalizers!

Scaling Law

- The larger the LLM Retriever, The better the model generalizes for Bi-Encoder.
- Recent works in **GTR** (Ni et al., 2021), **SGPT** (Muennighoff et al., 2022) and **CPT-Text** (Neelakantan et al., 2022) shown general improvement versus BM25 in zero-shot BEIR generalization.

CPT-text (XL) (Neelakantan et al. 2020)	**175B**	**BM25**	BEIR (11 Datasets Avg.)	⬆ **5.2** points (NDCG@10)
SGPT-5.8B (Muennighoff et al. 2021)	**5.8B**	**BM25**	BEIR (18 Datasets Avg.)	⬆ **6.2** points (NDCG@10)
GTR-XXL (Ni et al. 2021)	**4.8B**	**BM25**	BEIR (18 Datasets Avg.)	⬆ **3.5** points (NDCG@10)

How to Improve Bi-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.

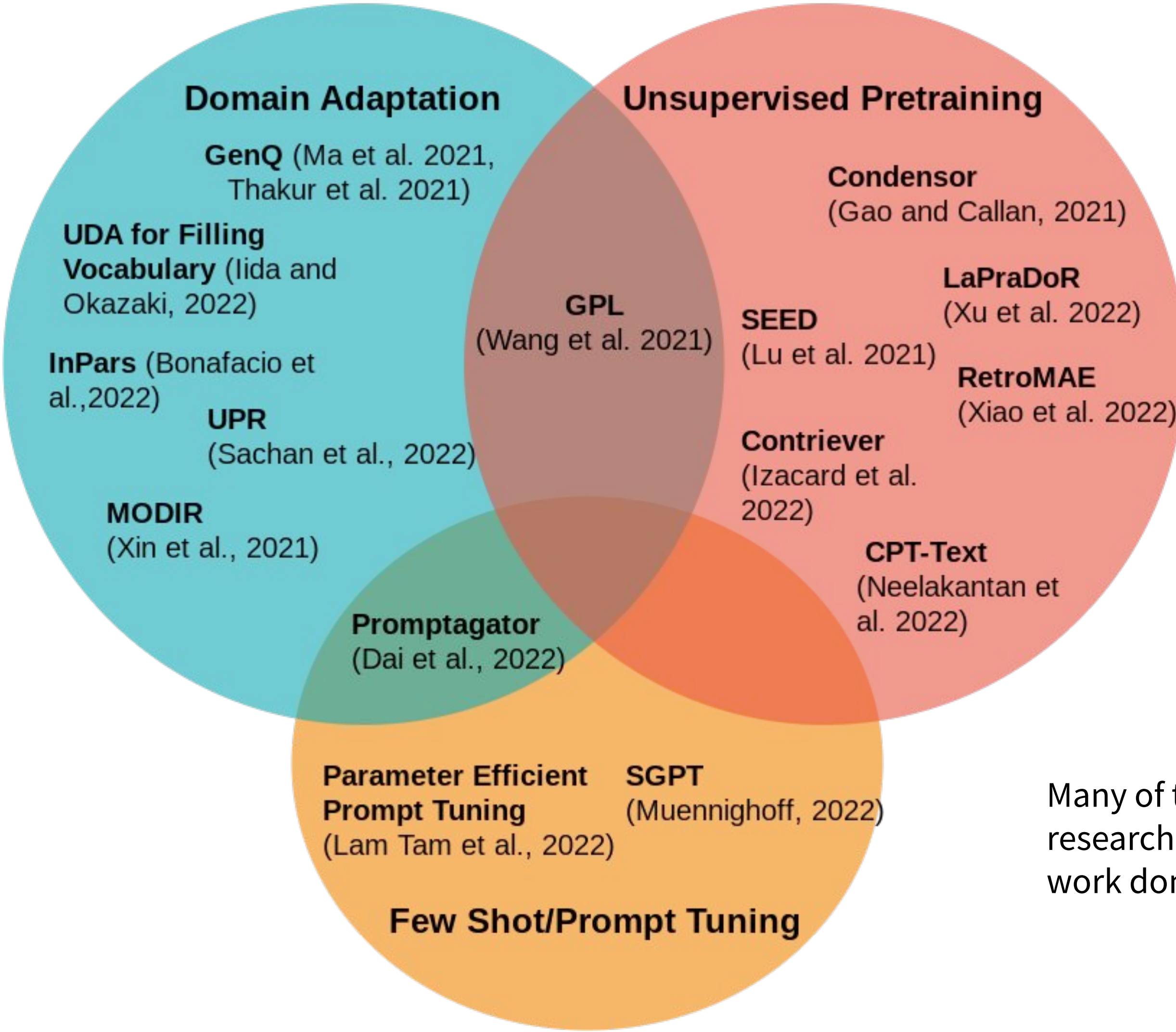
Unsupervised Pre-training

- 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 Bi-Encoder Generalization

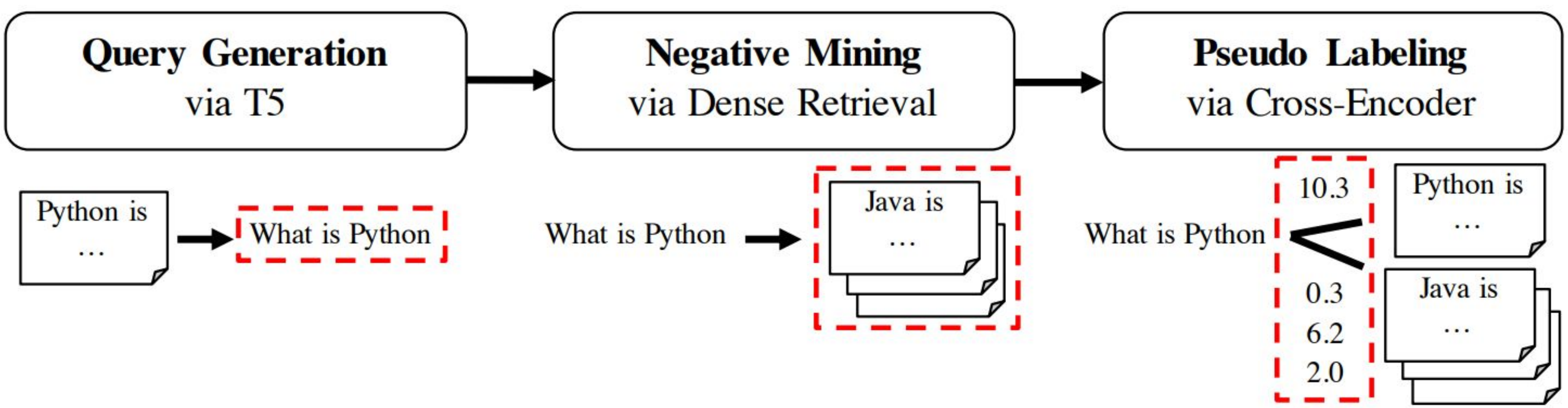


Many of these ideas (by other researchers) got inspired by work done in BEIR :)

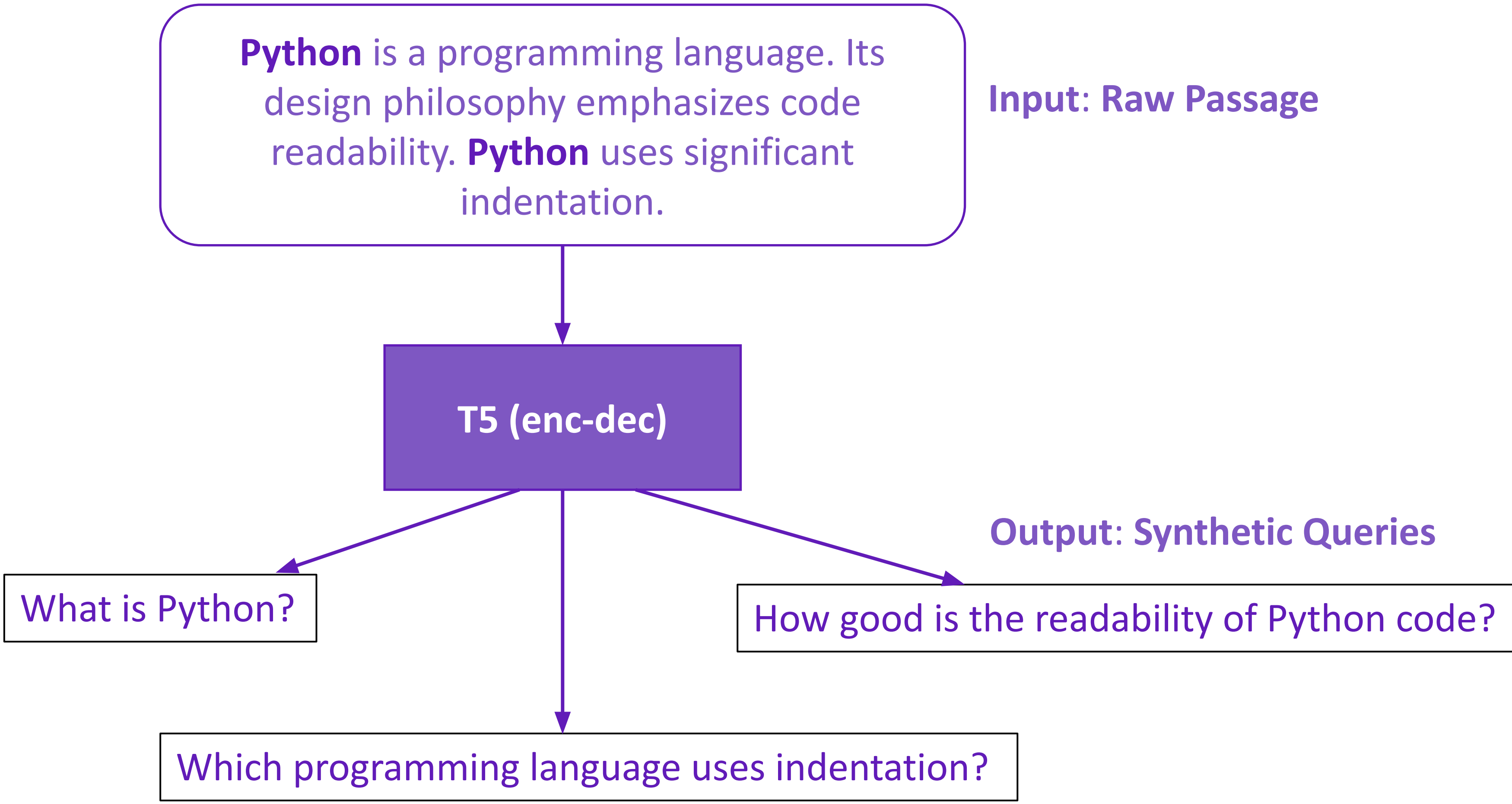
GPL – Generative Pseudo Labeling (Wang et al. 2021)

Fine-tuned model (e.g. on MSMARCO) → **GPL**: Unsupervised domain adaptation

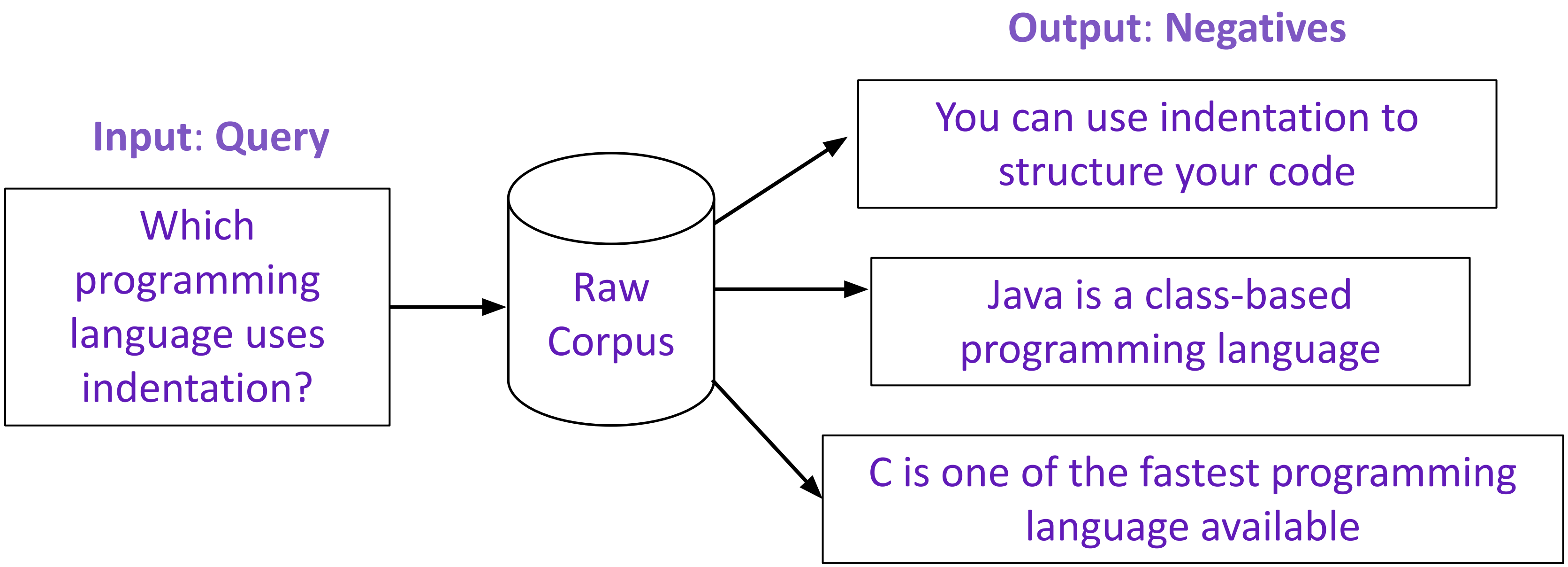
GPL:



GPL Step 1: Generate Queries



GPL Step 2: Mine Negatives

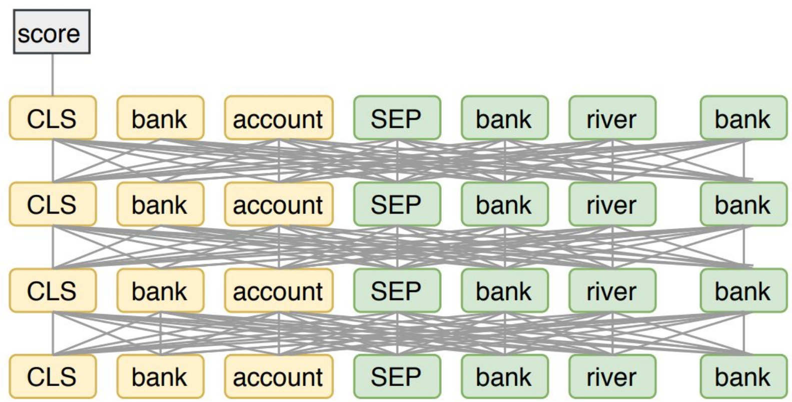


GPL Step 3: Label using Cross-Encoder

(Query, Doc1)

(Query, Doc2)

(Query, Doc3)



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

Cross-Encoder

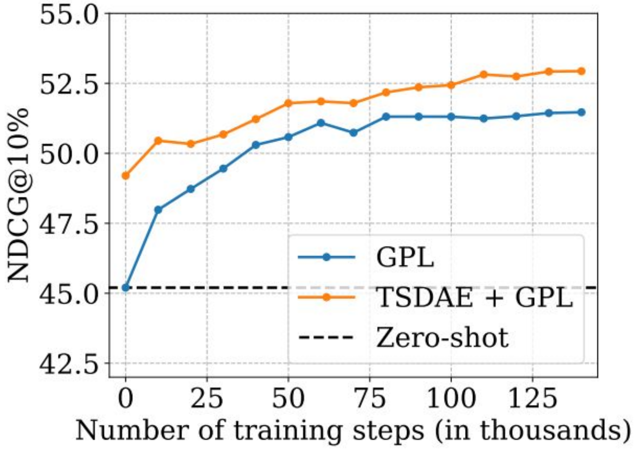
5.2

0.1

-3.8

GPL Results on BEIR Benchmark

Models	BEIR (6 Datasets Avg.)
Zero-shot (TAS-B)	45.2
Target -> Source	
TSDAE	49.2
MLM	46.7
Generative Pseudo Labeling	
GPL	51.5
TSDAE+GPL	**52.9**



GPL Success: Fine-grained Relevance Scores

Item	Text	GPL	QGen
Query	what is **futures contract**	–	–
Positive	**Futures contracts** are a member of a larger class of financial assets called derivatives ...	10.3	1
Negative 1	... Anyway in this one example the s&p 500 **futures contract** has an "initial margin" of $19,250, meaning ...	2.0	0
Negative 2	... but the moment you exercise you must have $5,940 in a margin account to actually use the **futures contract** ...	0.3	0
Negative 3	... a **futures contract** is simply a contract that requires party A to buy a given amount of a commodity from party B at a specified price...	8.2	0
Negative 4	... A **futures contract** commits two parties to a buy/sell of the underlying securities, but ...	6.9	0

GPL (Margin-MSE Loss)

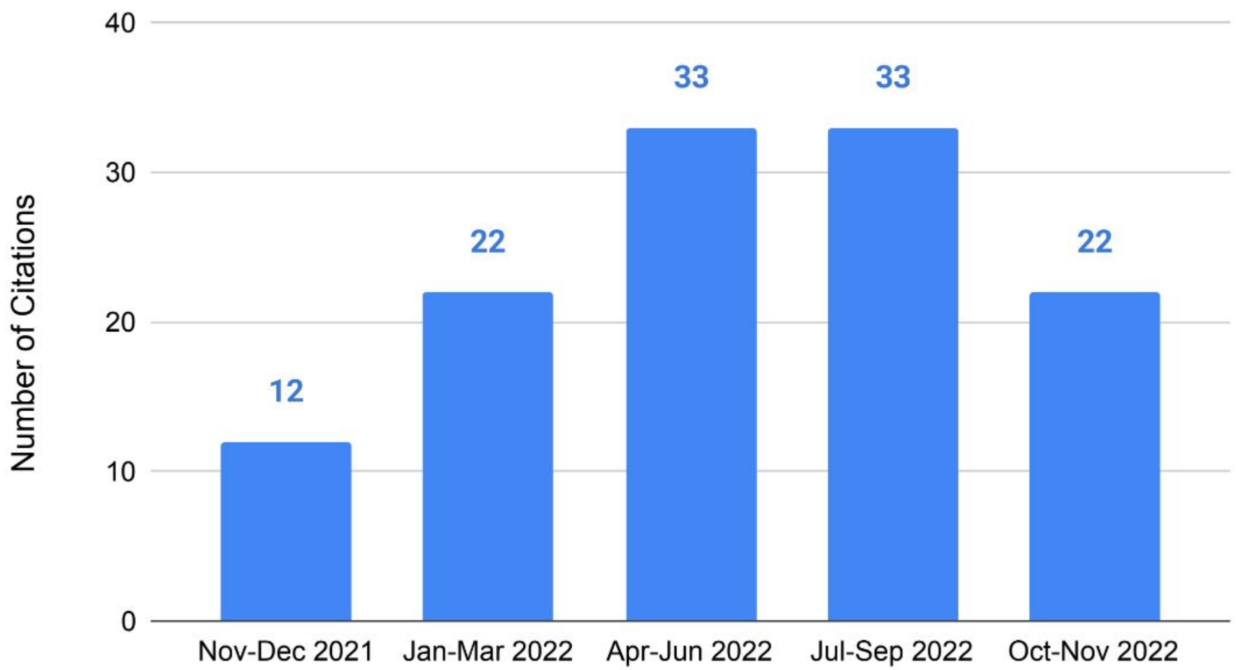
$$L_{\text{MarginMSE}}(\theta) = -\frac{1}{M}\sum_{i=0}^{M-1}|\hat{\delta}_i - \delta_i|^2$$

QGen (Cross-Entropy Loss)

$$L_{\text{MNRL}}(\theta) = -\frac{1}{M}\sum_{i=0}^{M-1}\log\frac{\exp\left(\tau\cdot\sigma(f_\theta(Q_i), f_\theta(P_i))\right)}{\sum_{j=0}^{M-1}\exp\left(\tau\cdot\sigma(f_\theta(Q_i), f_\theta(P_j))\right)}$$

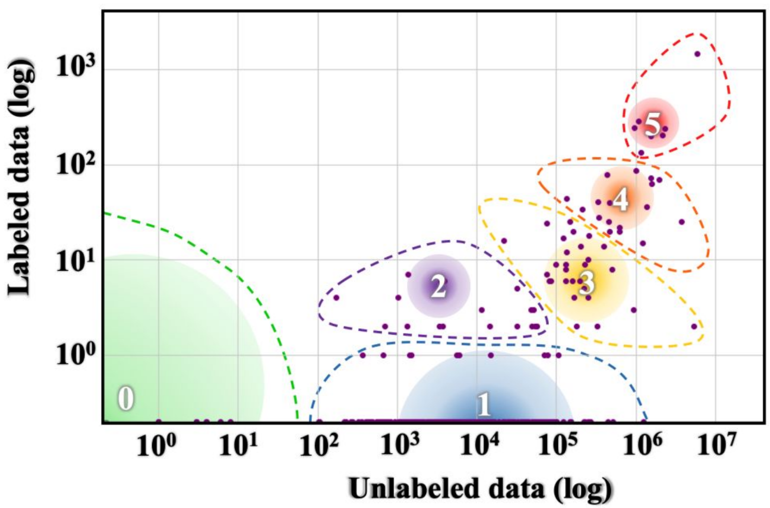
Expanding the Horizon: Going Multilingual!

BEIR Benchmark Outreach on Zero-shot English IR



Arxiv	**60**	ECIR	**4**
SIGIR	**10**	ACL	**3**
CIKM	**7**	FINDINGS	**3**
NAACL	**6**	NAACL-HLT	**2**

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

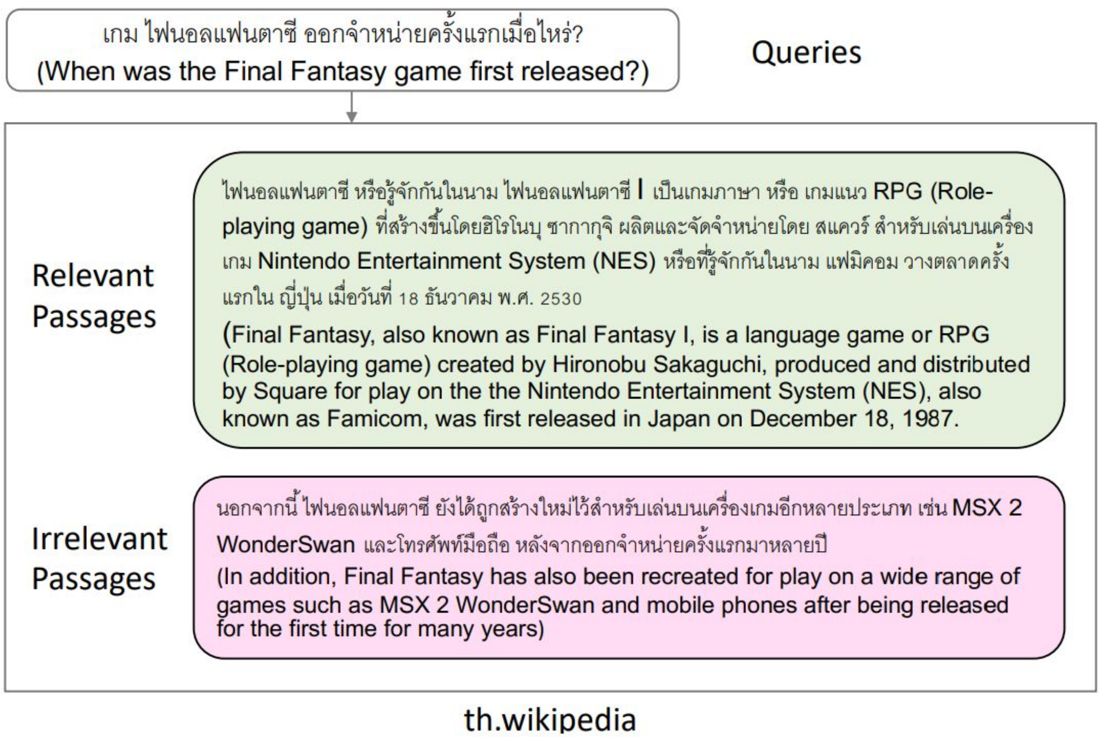
Queries can be about culturally specific topics (e.g., *Maacher Jhol* in Bengali)

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)

Push towards Multilingual IR Benchmarking



Got Selected at WSDM Cup'23

Competition and Leaderboard is public!

MIRACL Benchmark (in collaboration with Huawei)

Dataset Name	# Lang.	Avg # Q	Avg # Label / Q	# Human Labels	Training Data?	Not Translated?	Manual?
FIRE 2012	5	50	89	224k	×	✓	✓
MKQA	26	10k	1.35	14k	×	✓	✓
mMARCO	13	808k	0.66	533k	✓	×	✓
CLIR Matrix	139	352k	693	0	✓	✓	×
Mr. TyDi	11	6.3k	1.02	71k	✓	✓	✓
MIRACL (ours)	18	23.7k	10	434k	✓	✓	✓

- **Scarce** 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**).

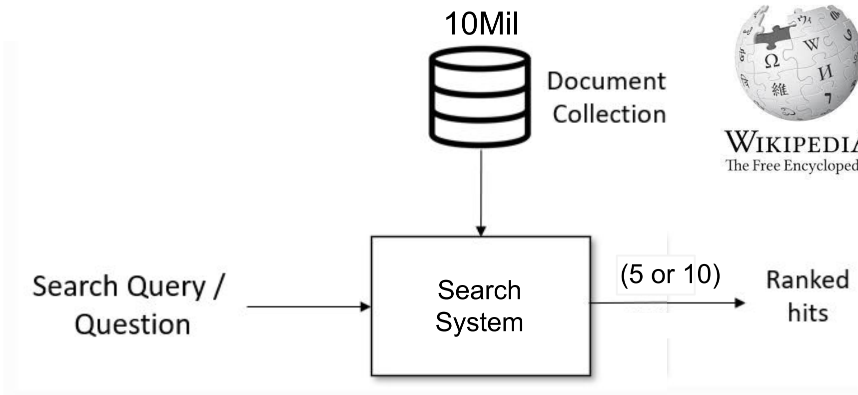
  - Better reflect speakers' **true interests** and **linguistic phenomena**
  - Hired over **40 native speakers** for the wide-scale annotation study
  - Performance will **lead to different insights** across languages, as each language has its own linguistic features.

Conclusions

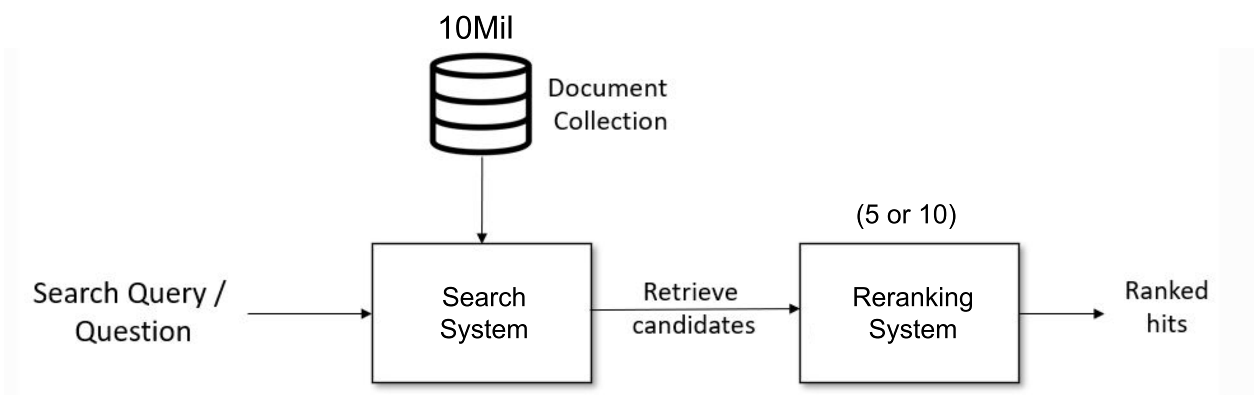
- Benchmarks are useful to measure progress in a meaningful way!
- Limitations seen in benchmarks help accelerate future research progress to eliminate them!
- Always evaluate your models across meaningful benchmarks containing diverse datasets!
- Do not always chase leaderboard (SoTA) improvement, especially on saturated leaderboards!

Thank you for listening!

Breaking down popular 🔍 IR Tasks



Retrieval



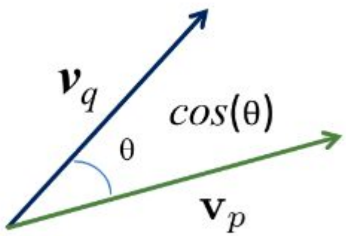
Retrieve and Rerank

Traditional BoW Search Systems

Vocabulary Mismatch (Cat vs. Kitty)

Limitations with Traditional Search Systems

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



$d_1 >> d_2$

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

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.*

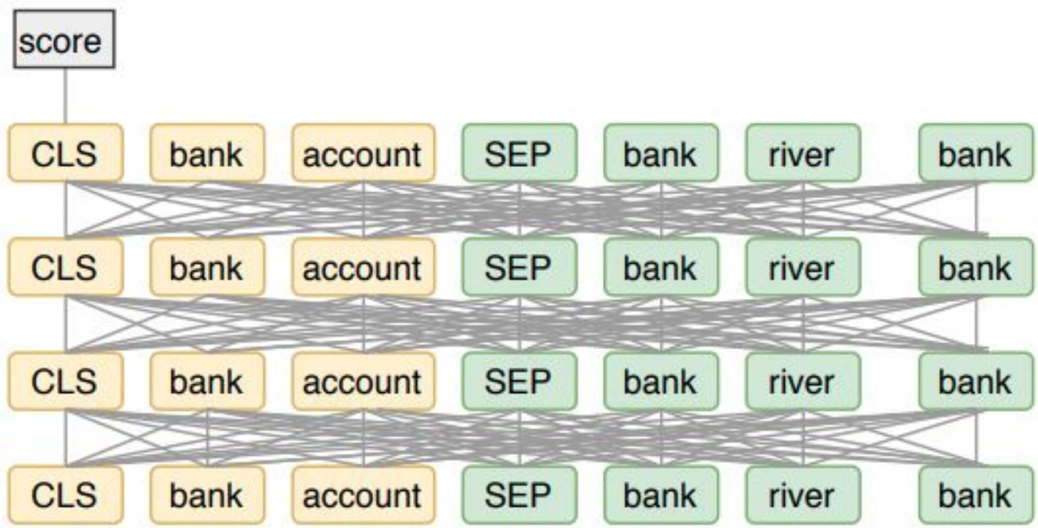
Ref: Danqi Chen, ACL 2020 OpenQA Tutorial
https://github.com/danqi/acl2020-openqa-tutorial/blob/master/slides/part5-dense-retriever-e2e-training.pdf

Modern (Neural) Search Systems

1. Retrieval: Bi-Encoders
2. Reranking: Cross-Encoders

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.

A Simple Illustration

Performance (Cross-Encoder > Bi-Encoder > BM25)

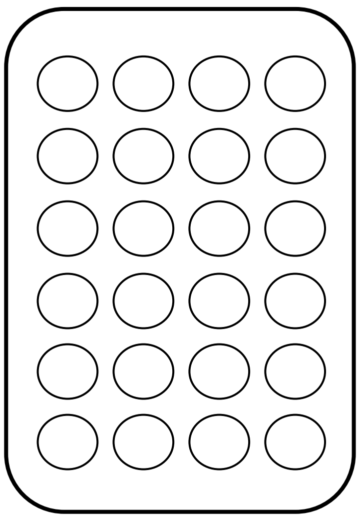
The Script uses the smaller Simple English Wikipedia as document collection. We test out sample user queries below and compare results:

https://colab.research.google.com/drive/1l6stpYdRMmeDBK_vw0L5NitdiAuhdsAr?usp=sharing

Why Zero-Shot Evaluation is Important?

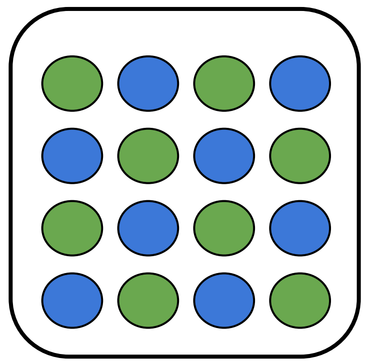
Generating High-Quality Labeled Training Data is cumbersome!

No Annotation Reqd.



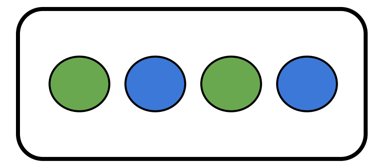
Unlabeled Data
Typically in ~Millions

Lots Annotation Reqd.



Labeled Training Data
Typically in **~100k** pairs

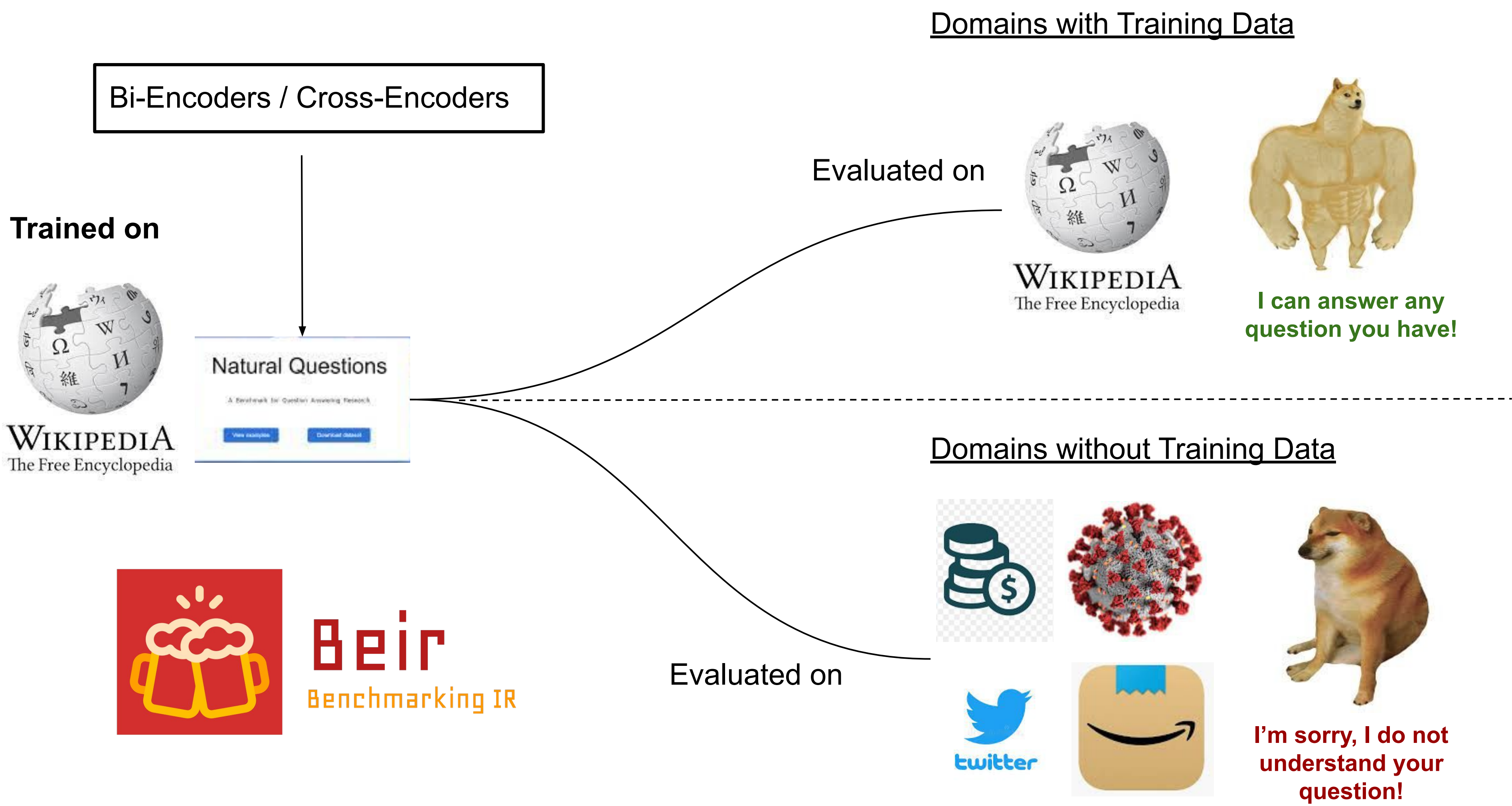
Few Annotation Reqd.



Labeled Test Data
Typically in **~100** pairs

RQ: Can Modern Search Systems Generalize?

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



BEIR: Evaluation Benchmark for IR Systems

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

Split (→)					Train	Dev	Test			Avg. Word Lengths	
Task (↓)	**Domain (↓)**	**Dataset (↓)**	**Title**	**Relevancy**	**#Pairs**	**#Query**	**#Query**	**#Corpus**	**Avg. D / Q**	**Query**	**Document**
Passage-Retrieval	Misc.	MS MARCO [42]	✗	Binary	532,761	—	6,980	8,841,823	1.1	5.96	55.98
Bio-Medical Information Retrieval (IR)	Bio-Medical	TREC-COVID [63]	✓	3-level	—	—	50	171,332	493.5	10.60	160.77
	Bio-Medical	NFCorpus [7]	✓	3-level	110,575	324	323	3,633	38.2	3.30	232.26
	Bio-Medical	BioASQ [59]	✓	Binary	32,916	—	500	14,914,602	4.7	8.05	202.61
Question Answering (QA)	Wikipedia	NQ [32]	✓	Binary	132,803	—	3,452	2,681,468	1.2	9.16	78.88
	Wikipedia	HotpotQA [74]	✓	Binary	170,000	5,447	7,405	5,233,329	2.0	17.61	46.30
	Finance	FiQA-2018 [41]	✗	Binary	14,166	500	648	57,638	2.6	10.77	132.32
Tweet-Retrieval	Twitter	Signal-1M (RT) [57]	✗	3-level	—	—	97	2,866,316	19.6	9.30	13.93
News Retrieval	News	TREC-NEWS [56]	✓	5-level	—	—	57	594,977	19.6	11.14	634.79
	News	Robust04 [62]	✗	3-level	—	—	249	528,155	69.9	15.27	466.40
Argument Retrieval	Misc.	ArguAna [65]	✓	Binary	—	—	1,406	8,674	1.0	192.98	166.80
	Misc.	Touché-2020 [6]	✓	3-level	—	—	49	382,545	19.0	6.55	292.37
Duplicate-Question Retrieval	StackEx.	CQADupStack [23]	✓	Binary	—	—	13,145	457,199	1.4	8.59	129.09
	Quora	Quora	✗	Binary	—	5,000	10,000	522,931	1.6	9.53	11.44
Entity-Retrieval	Wikipedia	DBPedia [19]	✓	3-level	—	67	400	4,635,922	38.2	5.39	49.68
Citation-Prediction	Scientific	SCIDOCS [9]	✓	Binary	—	—	1,000	25,657	4.9	9.38	176.19
Fact Checking	Wikipedia	FEVER [58]	✓	Binary	140,085	6,666	6,666	5,416,568	1.2	8.13	84.76
	Wikipedia	Climate-FEVER [13]	✓	Binary	—	—	1,535	5,416,593	3.0	20.13	84.76
	Scientific	SciFact [66]	✓	Binary	920	—	300	5,183	1.1	12.37	213.63

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:

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

Zero-shot 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 **11/17** retrieval datasets.

Bi-Encoders (Dense)

Dense models suffer from generalization. They outperform BM25 on **7/17** datasets.

Efficiency and Memory Comparison on BEIR

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

DBPedia (1 Million)			Retrieval Latency		Index
Rank	Model	Dim.	GPU	CPU	Size
(1)	Cross-Encoders	768	550ms	7100ms	0.4GB
(2)	Cross-Encoders	128	350ms	–	20GB
(3)	BM25	–	–	20ms	0.4GB
(4)	Bi-Encoders	768	14ms	125ms	3GB
(5)	Bi-Encoders	768	20ms	275ms	3GB
(6)	Bi-Encoders	768	14ms	125ms	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. arXiv preprint arXiv:2104.08663.

Conclusions (To Recap)

Traditional vs Modern Search Systems

1. Traditional Search Systems like BM25 use keyword based-search which miss out on Synonyms.

2. Bi-Encoders map query and document to a dense vector space, efficient and practical. However, they fail to perform well in zero-shot setting and are unable to generalize well!

3. Cross-Encoders take the query and document together, best performing on zero-shot. But quite impractical for real-world setting!

4. Generalization with models is quite a difficult task and there is no free lunch!

Thank You For Listening!
Any Questions?

Paper Link:
https://openreview.net/forum?id=wCu6T5xFjeJ

BEIR: A Heterogenous 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

Neural IR models have often been studied in homogeneous and narrow settings, which has considerably limited insights into their generalization capabilities. To address this, and to allow researchers to more broadly establish the effectiveness of their models, we introduce **BEIR** (***Be**nchmarking **IR***), a *heterogeneous benchmark* for information retrieval.

the keywords also present within the query. Further, queries and documents are treated in a bag-of-words manner which does not take word ordering into consideration.

Recently, deep learning and in particular pre-trained Transformer models like BERT (Devlin et al., 2018) have became popular in the information retrieval space (Lin et al., 2020). They overcome the lexical gap by mapping queries and

GitHub: https://github.com/UKPLab/beir

beir

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

Python ★ 213 25

Google colab

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