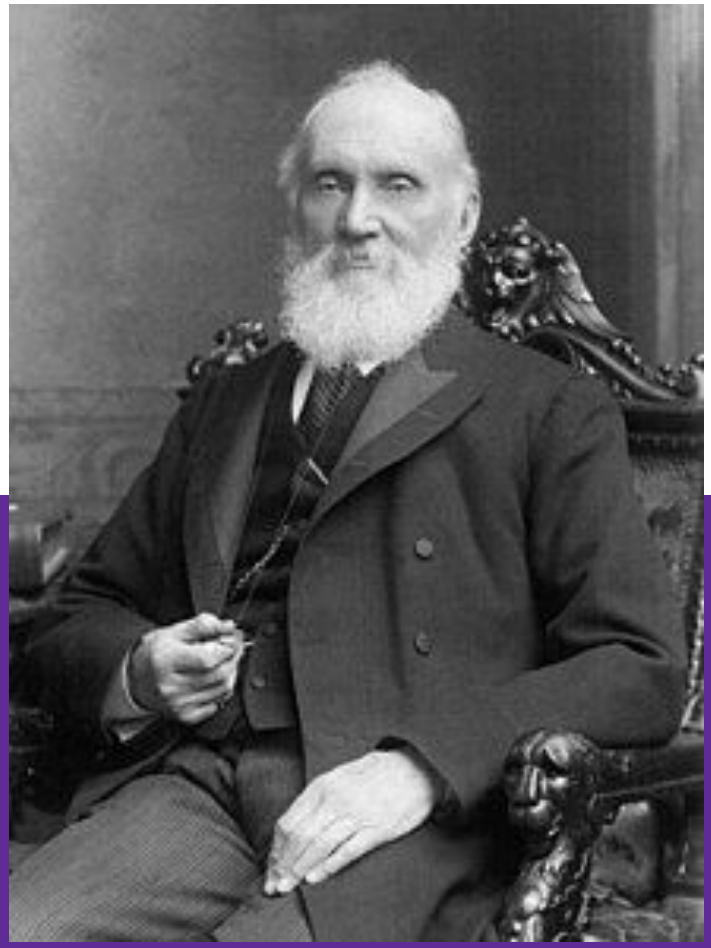
Heterogenous Benchmarking across Domains and Languages

The Key to Enable Meaningful Progress in IR Research

Nandan Thakur
Third-year Ph.D. Student
University of Waterloo

David R. Cheriton School of Computer Science
University of Waterloo



British mathematician, Lord Kelvin

"If you cannot measure it, you cannot improve it."

Today's agenda

(10 mins) **IR Fundamentals**

(5 mins) **Why Benchmarking?**

(10 mins) **The Zero-shot Paradigm: IR Generalization**

(10 mins) **Expanding the Horizon: Going Multilingual!**

(7 mins) **Data-Efficient Multilingual Retrieval**

(3 mins) **Open Research Problems & Conclusion**

UNIVERSITY OF WATERLOO

IR Fundamentals: A Refresher

What is 🔍 information retrieval?

🔍 **Which football club Lionel Messi plays for?**

natural language query

OR

🔍 **Messi football club**

keyword-based query

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 and captains both Major League Soccer club Inter Miami and the Argentina national team. Widely regarded as one of the greatest players of all time, Messi has won a record eight Ballon d'Or awards, a record six European Golden Shoes, and was named the world's best player by FIFA for a record eight times.

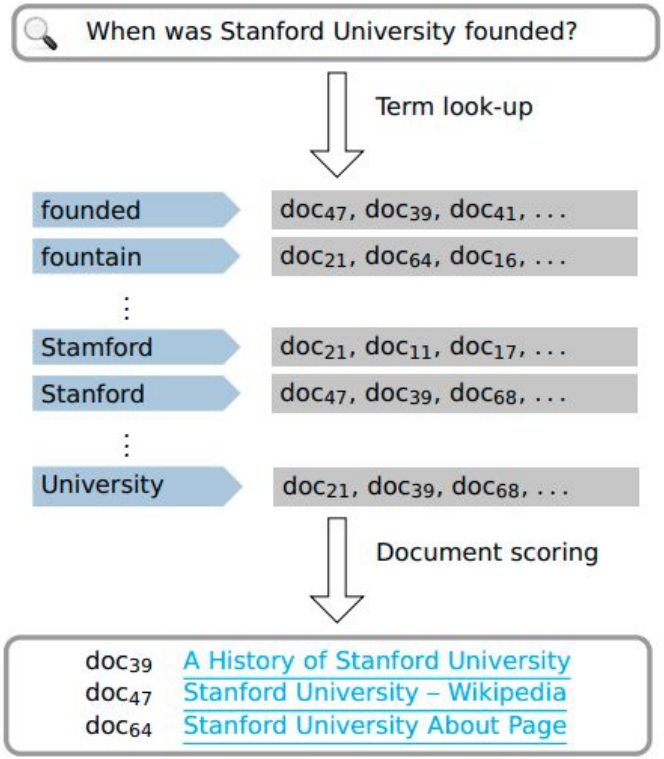
Information retrieval is omnipresent

Ubiquitous

present, appearing, or found everywhere.

Okapi BM25 (Best Match 25) [1]

Lexical-based search: bag of words/keyword matching



BM25 Intuition

- k_1 controls **query-term saturation**.
- **b** controls **document length normalization**.

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

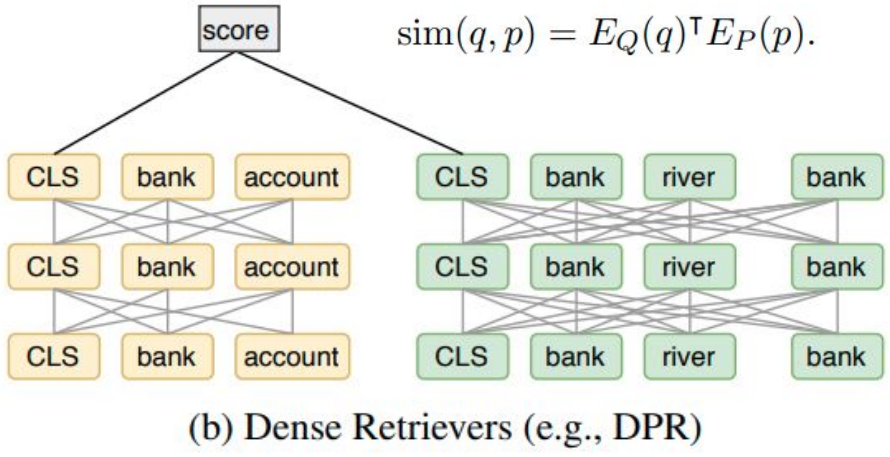
Ref: Christopher G Potts, ACL-IJCNLP 2021 keynote address:
https://web.stanford.edu/~cgpotts/talks/potts-acl2021-slides-handout.pdf

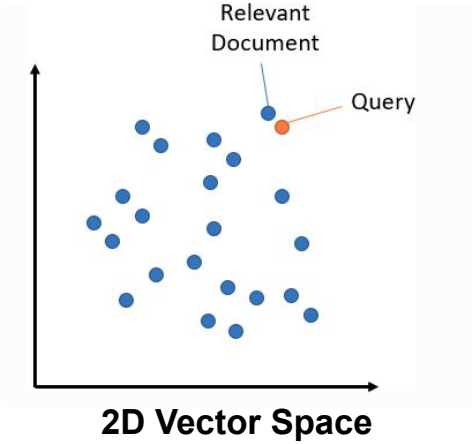
Dense retrieval with bi-encoders [2]

Mapping Individual Text to a fixed dimensional embedding!

Benefits

- Attention within query or passage tokens.
- **[CLS]** or **mean-pooling** is used for final representation.
- Fast and efficient at runtime.

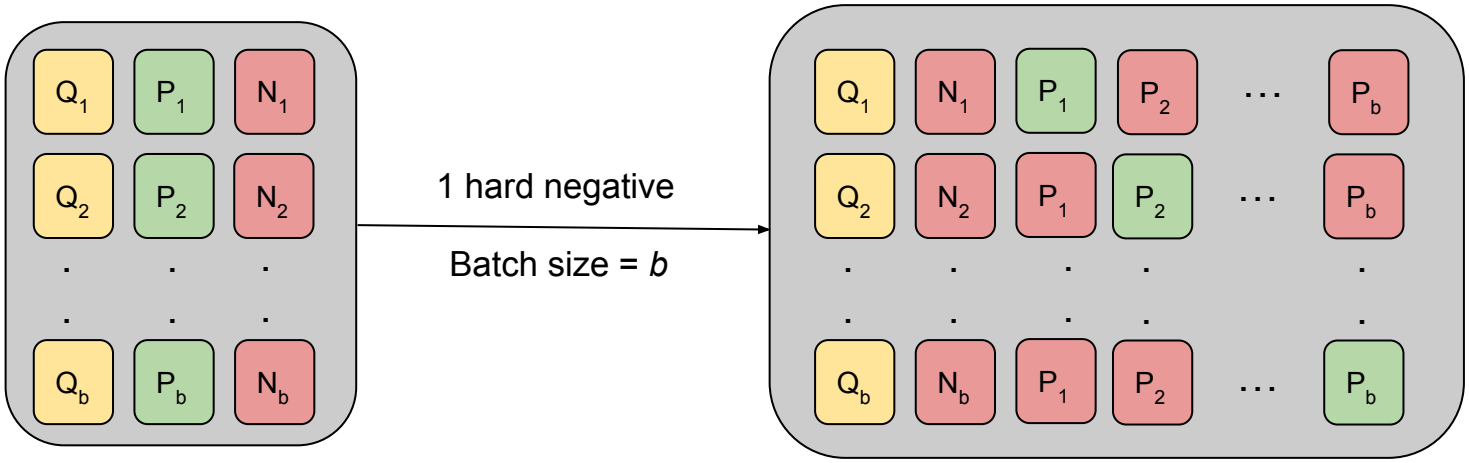




Taken from COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List.
https://aclanthology.org/2021.naacl-main.241/

How to fine-tune a bi-encoder model?

Using In-batch fine-tuning with hard negatives (N_1, .. , N_b)



Cross-entropy loss function

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

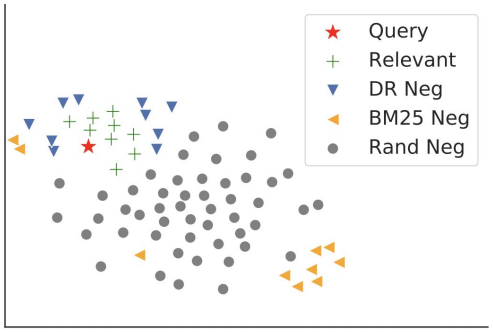


Figure taken from Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. https://arxiv.org/abs/2007.00808

Late-interaction model: ColBERT [3]

Mapping Individual tokens to fixed dimensional embeddings

(a) Token Retrieval

Query tokens used to search **top-k_1** doc tokens (among all tokens in corpus).

(b) Gathering

Top-k_2 tokens are mapped to the original document id.

(c) Scoring

All tokens in document are used to compute Maximum Similarity (*MaxSim*).

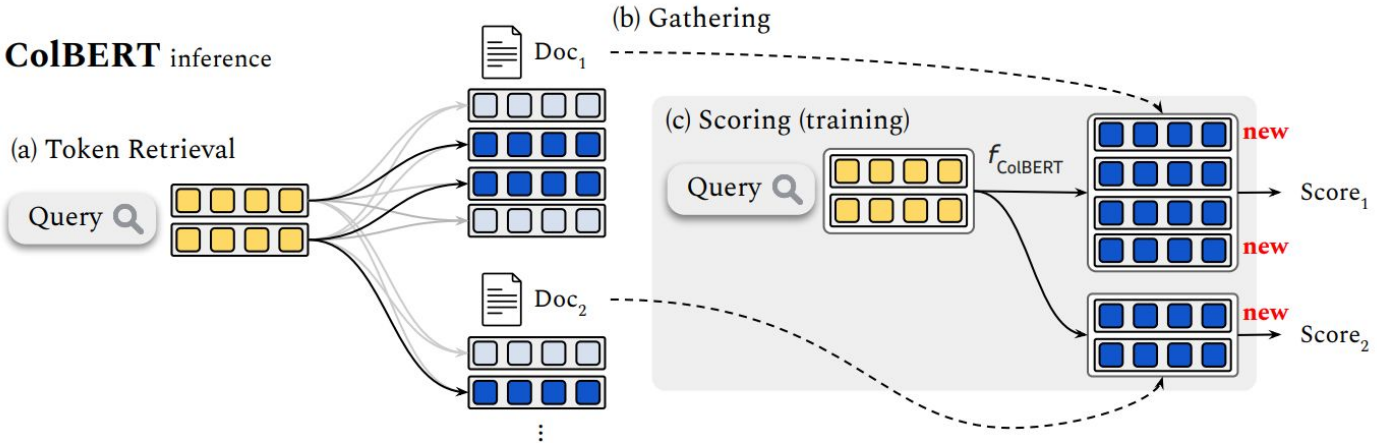


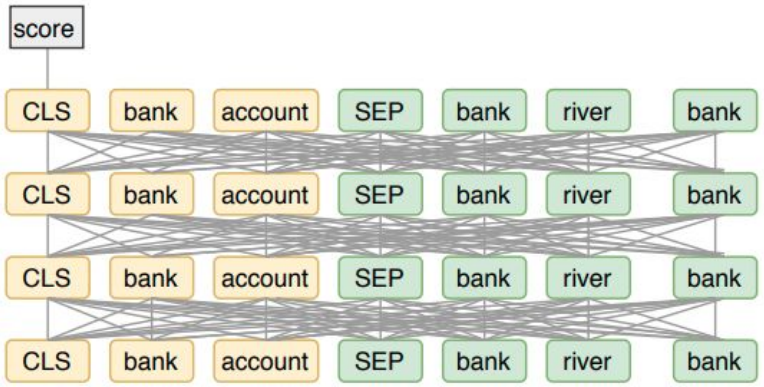
Figure taken from XTR: Rethinking the Role of Token Retrieval in Multi-Vector Retrieval.
https://openreview.net/forum?id=ZQzm0Z47jz

Reranking with Cross-Encoders [4]

Concatenate query and document together. No embedding!

Benefits

- Stronger but inefficient reranking performance as it trains a **classification objective**.
- **Cross-attention** between query and passage tokens.



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

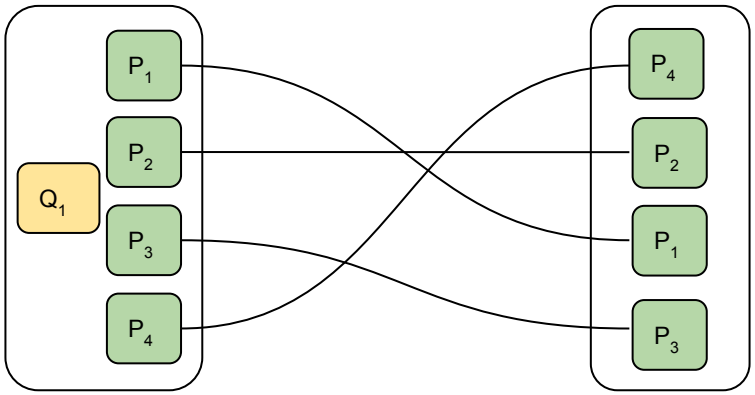


Illustration of the Reranking stage

Taken from COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. (https://aclanthology.org/2021.naacl-main.241/)

UNIVERSITY OF WATERLOO

Information Retrieval Benchmarking

What is benchmarking? Why is it useful?

Benchmarks has three crucial components: (1) one or multiple datasets, (2) one or multiple associated metrics, and (3) a way to aggregate performance.

Advantages of benchmarking

- Provides a **unified platform** and useful in understanding **fundamental gaps**.
- **Discover** SoTA model performances and compare difference w.r.t. **human performance**.
- Sets a **standard** for assessing the model performance to accelerate research.

Benchmarking in IR is over ~30 years old!

TREC is one of the prominent figures towards benchmarking in IR!

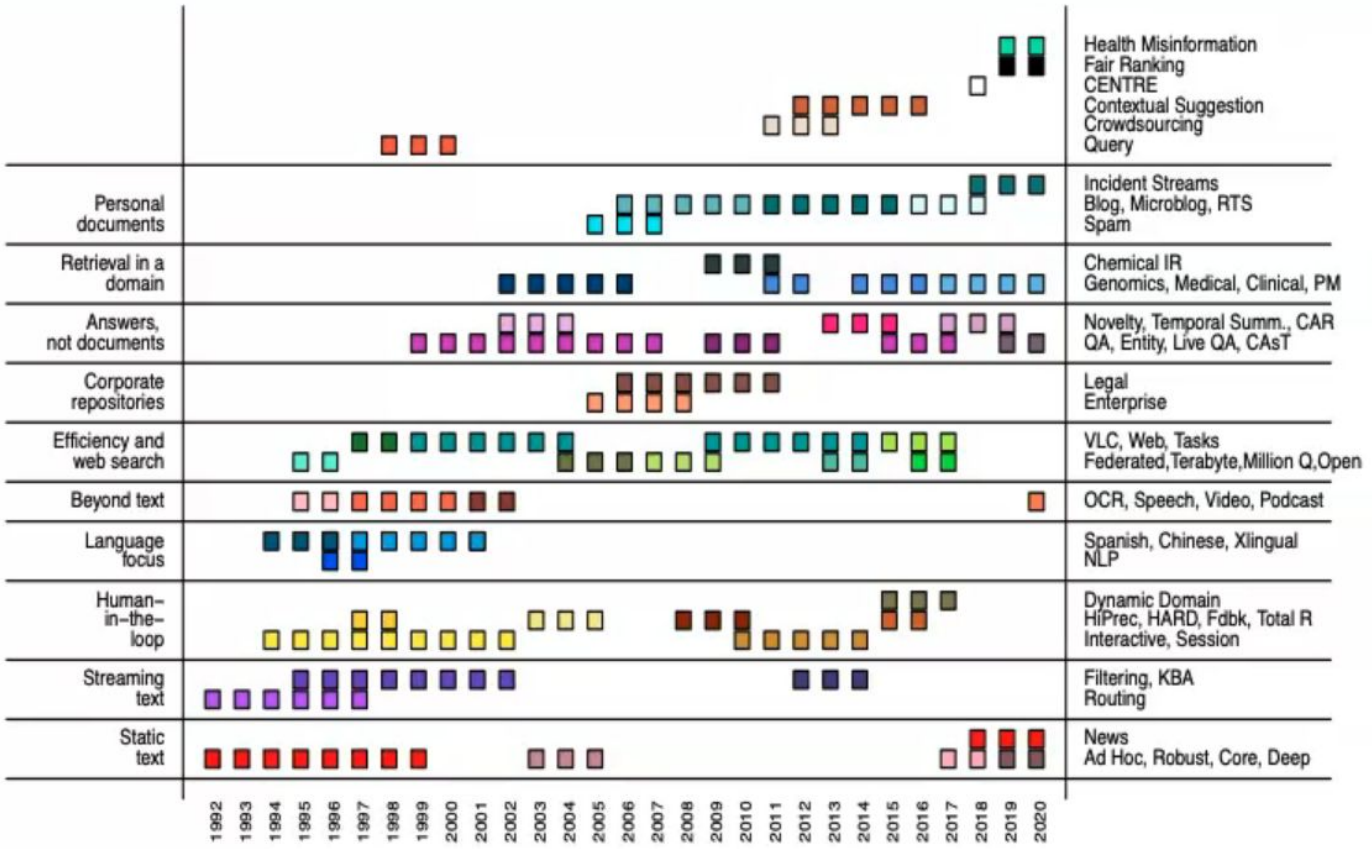


Figure: Overview of TREC tasks starting from 1992 onwards until present (table shows until 2020). Taken from NIST (https://www.nist.gov/image/tracksjpg)

Performance on standard IR benchmarks

Bi-encoders outperforms BM25 across all evaluated datasets!

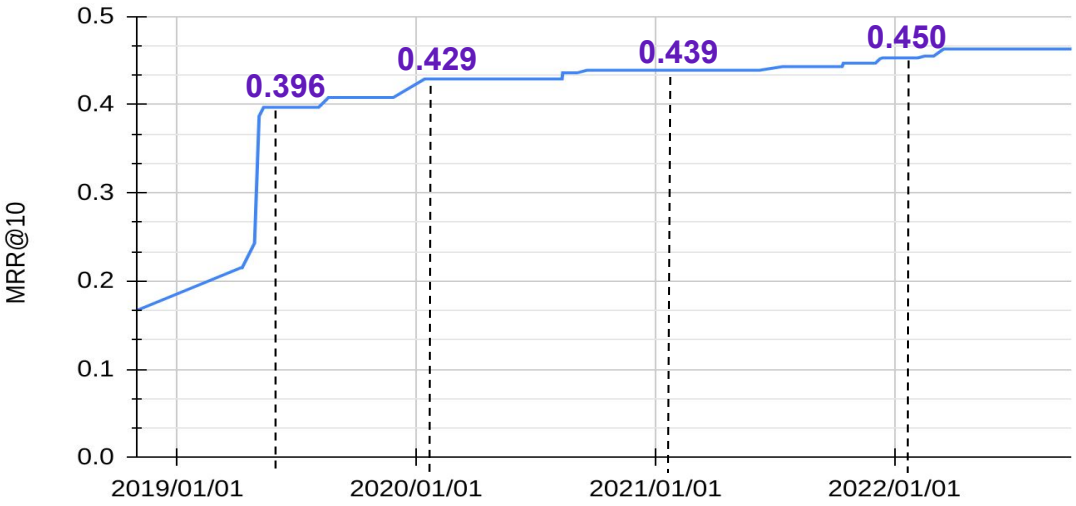
Dense Model	Baseline	Dataset	Performance Improvement
DPR [1]	**BM25**	NQ Retrieval	↑ **20.3** points (Top-20 Recall)
ANCE [5]	**BM25**	MSMARCO NQ Retrieval	↑ **9.0** points (MRR@10) **23.8** points (Top-20 Recall)
TAS-B [6]	**BM25**	MSMARCO	↑ **14.9** points (MRR@10)

Caveats

- Datasets such as MSMARCO and NQ contain over **100K+ human-labeled training data**.
- Does not reveal insights about model generalization.
- Unknown whether improvements are due to **overfitting to training dataset artifacts.**

Has MSMARCO already saturated?

- **Minimal performance improvement** in MSMARCO dev dataset (left).
- Assessors preferred a **majority of passages** provided by the neural ranker (right).



Maximum Performance on MSMARCO Dev (Full Retrieval) from 2019 onwards.
Numbers taken from official leaderboard (https://microsoft.github.io/msmarco/)

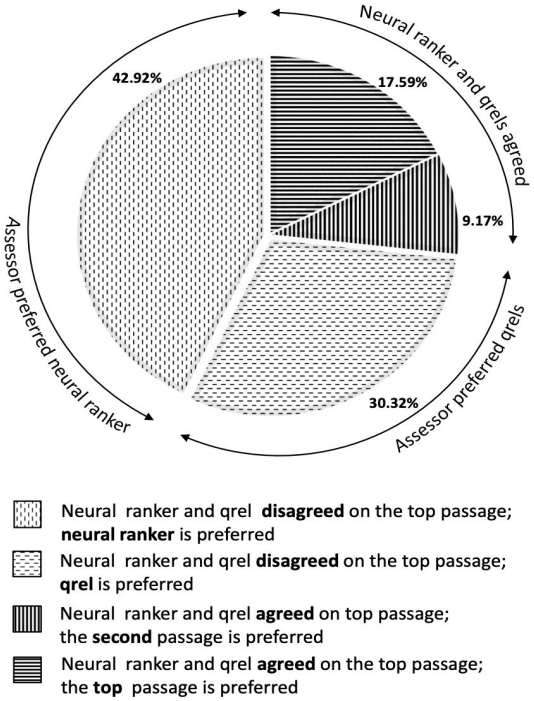
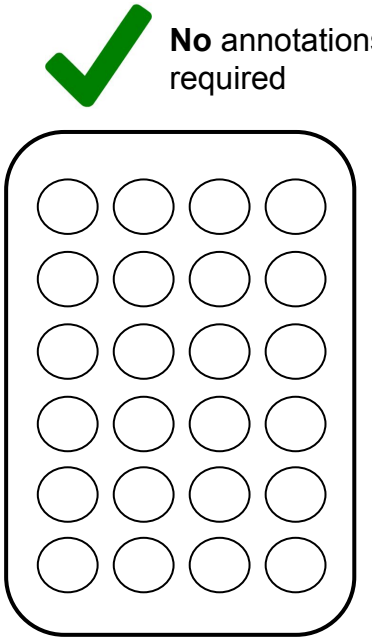


Diagram (right) taken from shallow pooling for sparse labels (https://arxiv.org/abs/2109.00062)

Why zero-shot evaluation in IR is necessary?

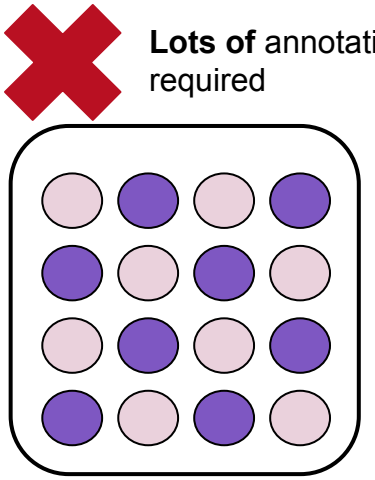
High-quality labeled training data is scarcely available!



No annotations required

Unlabeled Data

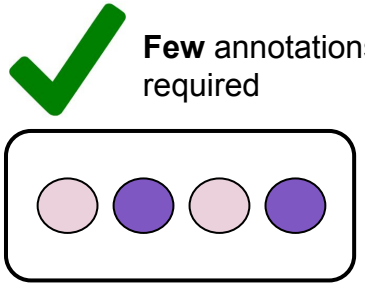
Typically in **~millions**



Lots of annotations required

Labeled Training Data

Typically **~100k** pairs



Few annotations required

Labeled Test Data

Typically **~100** pairs

Zero-shot Paradigm: IR Generalization

RQ: Can modern search systems generalize?

Will neural models perform well out-of-box (w/o) fine-tuning?

Bi-encoder /
Late-interaction /
Cross-encoder

MS MARCO

WIKIPEDIA
The Free Encyclopedia

Fine-tuned

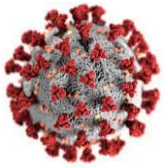
Evaluated on

Domains with training data

WIKIPEDIA
The Free Encyclopedia

I can answer any question you have!

Domains without training data



Evaluated on

twitter

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

Beir
Benchmarking IR

The BEIR Benchmark [7]

Zero-shot IR heterogeneous benchmark with 18 datasets!

- BEIR contains 18 datasets across **diverse retrieval based tasks and domains.**
- BEIR datasets do not contain a training split, enforcing a **zero-shot evaluation**.

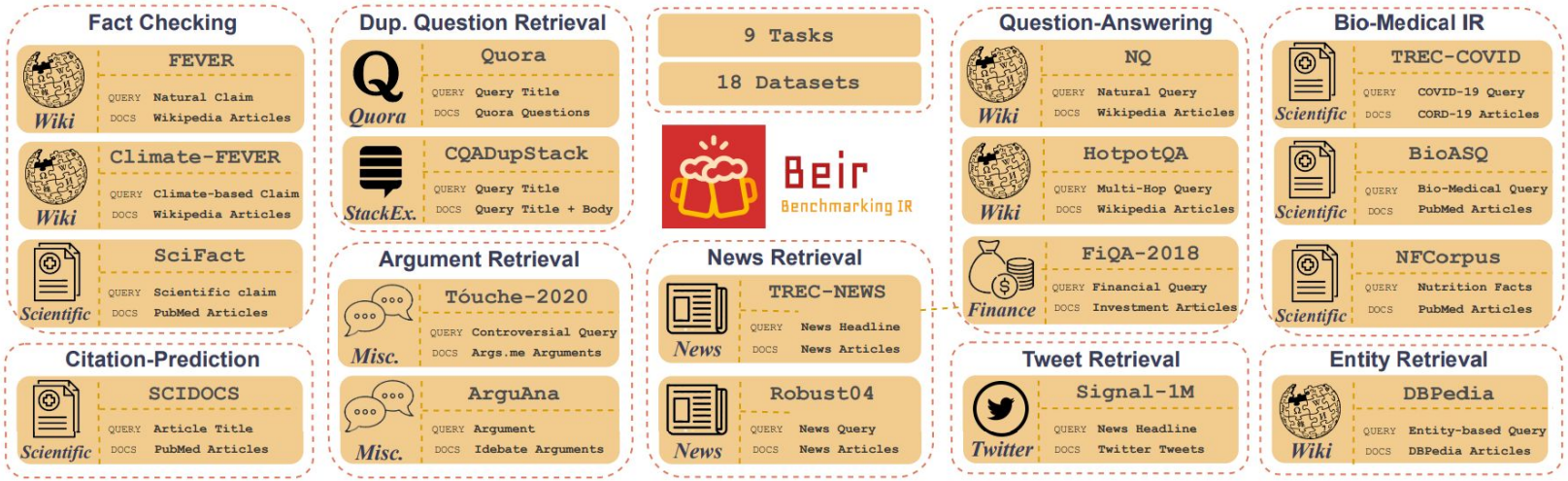


Figure taken from BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. (https://openreview.net/forum?id=wCu6T5xFjeJ)

How well do bi-encoders generalize on BEIR?

Bi-encoders surprisingly struggle to outperform BM25!

- **Bi-encoders struggle to beat BM25** on a majority of datasets.
- **ColBERT and cross-encoders** (+ BM25) are better at generalization.
- Recent 2023 models such as **ColBERTv2 and DRAGON+** outperform BM25.

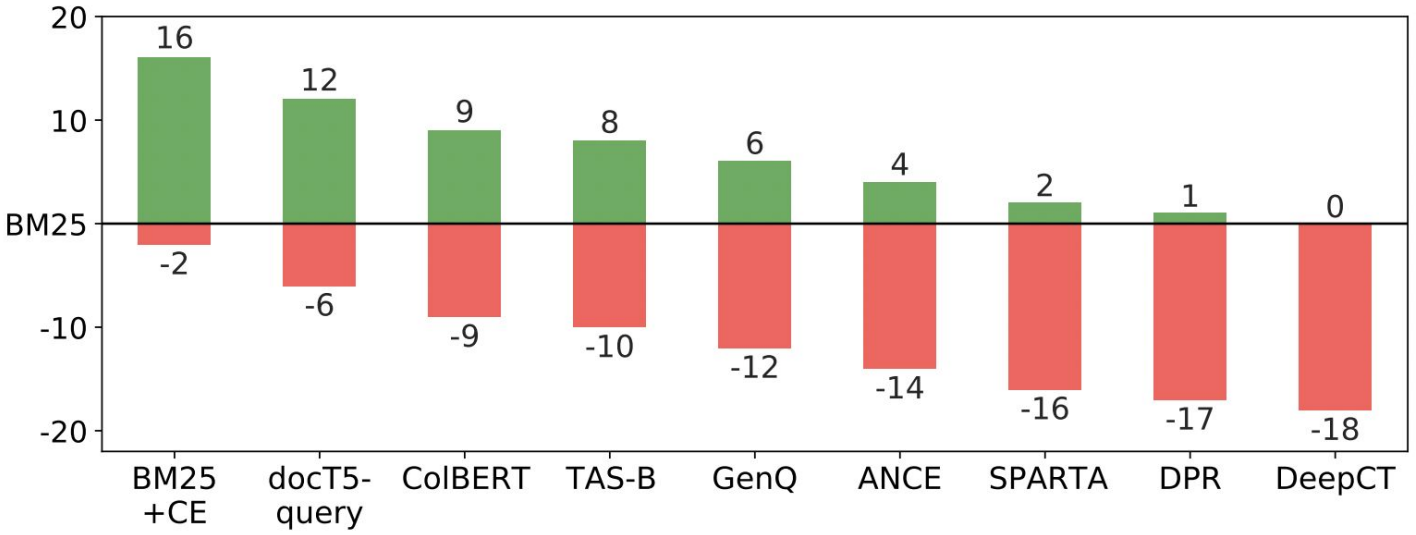


Figure taken from BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models in 2021. (https://openreview.net/forum?id=wCu6T5xFjeJ)

Why do Bi-encoders suffer in generalization?

Domain-specific vocabulary not known by the model

- How does bi-encoders handle **unknown words** unseen during fine-tuning?
- Where to put **new words and concepts** in the vector space of dense retriever?
- How to learn semantic **word relationships** with unknown words?

GPL – Generative Pseudo Labeling [8]

Three techniques in GPL are prominent in **unsupervised domain adaptation**.

(1) Query Generation: Synthetic queries for each passage in your corpora.

(2) Negative mining: Hard-negative passage mining using multiple models.

(3) Pseudo Labeling: Distill knowledge from cross-encoder using Margin MSE loss function.

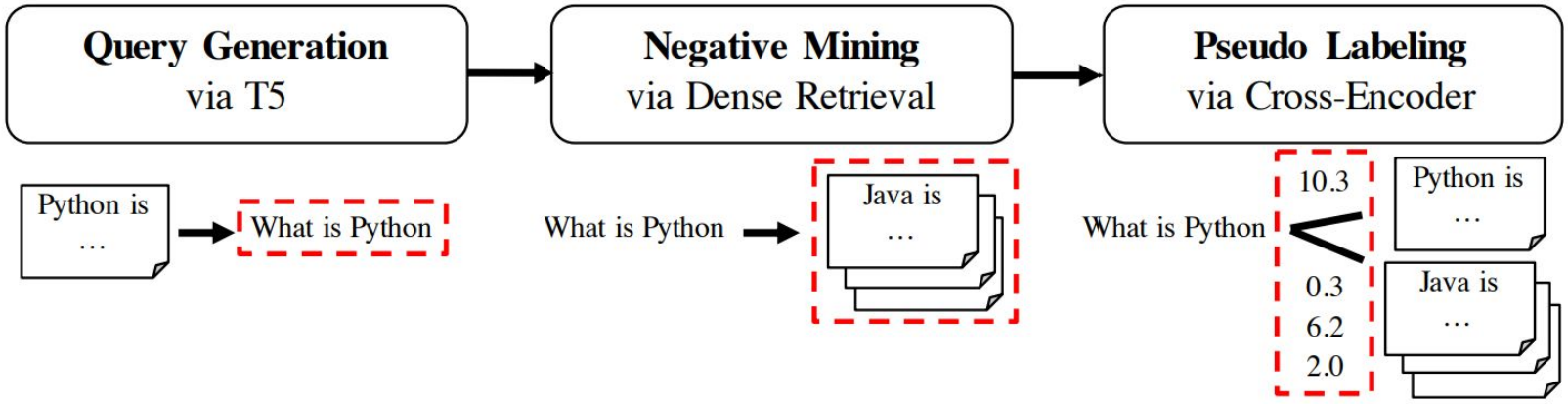


Figure taken from GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. (https://aclanthology.org/2022.naacl-main.168/)

GPL Results on the BEIR Benchmark

Models	BEIR (6 Datasets Avg.)
Zero-shot (TAS-B)	45.2
BM25 (Anserini)	48.5
Unsupervised Domain Adaptation	
QGen	48.8
TSDAE+QGen	50.3
Generative Pseudo Labeling	
GPL	51.5
TSDAE+GPL	**52.9**

Performance starts to saturate after 120K training steps.

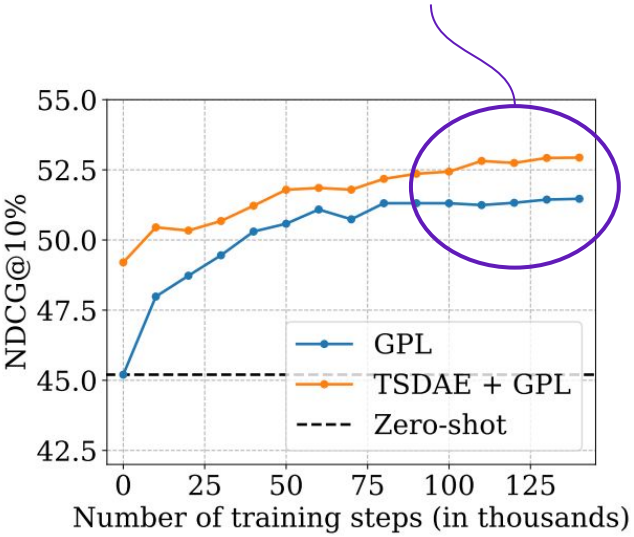


Table taken from GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. (https://aclanthology.org/2022.naacl-main.168/)

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

QGen (Cross-entropy loss function)

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

GPL (Margin-MSE loss function)

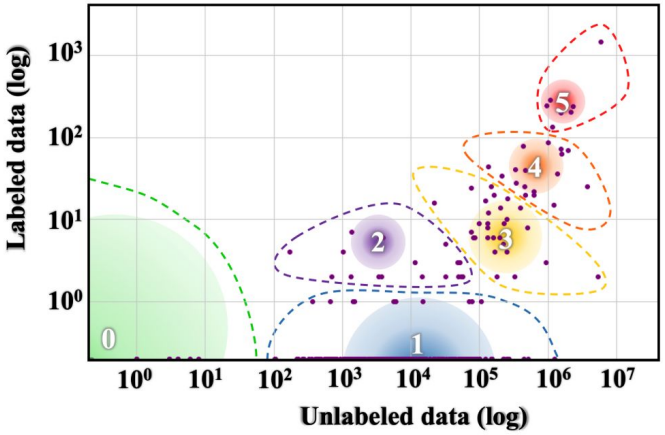
$$L_{\text{MarginMSE}}(\theta) = -\frac{1}{M}\sum_{i=0}^{M-1}|\hat{\delta}_i - \delta_i|^2$$

Table taken from GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. (https://aclanthology.org/2022.naacl-main.168/)

UNIVERSITY OF WATERLOO

Expanding the Horizon: Going Multilingual!

Information access is a right for everyone!

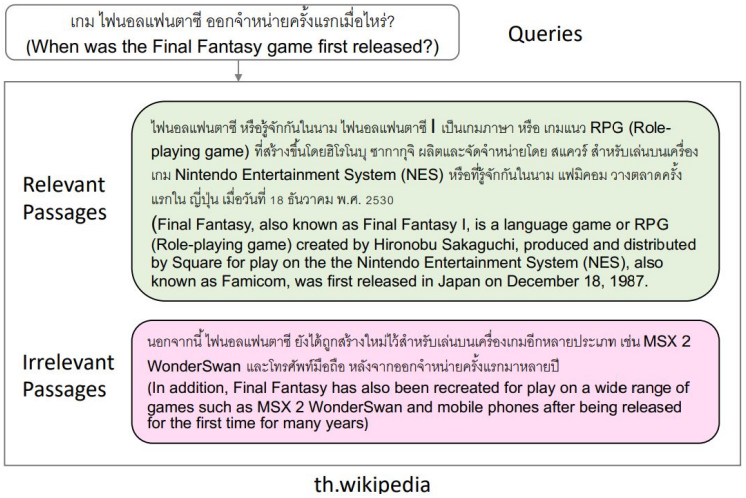


- Over **two-three billion** non-English native speakers.
- Languages have **diverse typologies**, originate from different families and have varying amounts of **available resources**.

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

Figure taken from The State and Fate of Linguistic Diversity and Inclusion in the NLP World. (https://aclanthology.org/2020.acl-main.560/)

Towards better multilingual IR benchmarking [9]



MIRACL

Multilingual Information Retrieval Across a Continuum of Languages

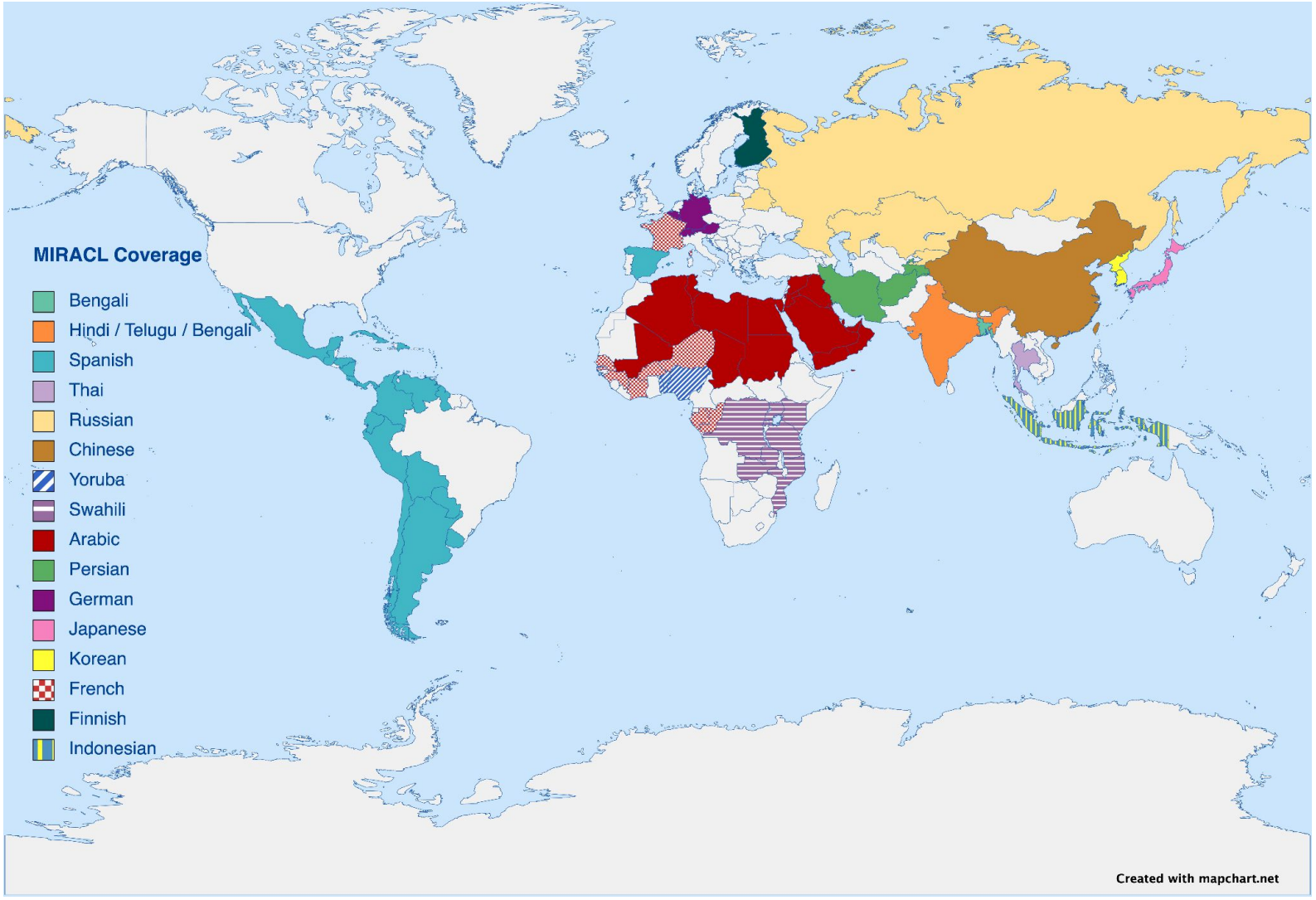
A large amount of relevance judgements across 40+ native annotators

training pairs available

Dataset Name	Natural Queries	Natural Passages	Human Labels	# Lang	Avg # Q	Avg # Labels/Q	Total # Labels	Training?
NeuCLIR (Lawrie et al., 2023)	✓	✓	✓	3	160	32.74	5.2k	×
MKQA (Longpre et al., 2021)	×	✓	✓	26	10k	1.35	14k	×
mMARCO (Bonifacio et al., 2021)	×	×	✓	13	808k	0.66	533k	✓
CLIRMatrix (Sun and Duh, 2020)	×	✓	×	139	352k	693	34B	✓
Mr. TyDi (Zhang et al., 2021)	✓	✓	✓	11	6.3k	1.02	71k	✓
MIRACL (our work)	✓	✓	✓	18	4.3k	9.23	726k	✓

Table and figure taken from MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. (https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00595/117438/MIRACL-A-Multilingual-Retrieval-Dataset-Covering)

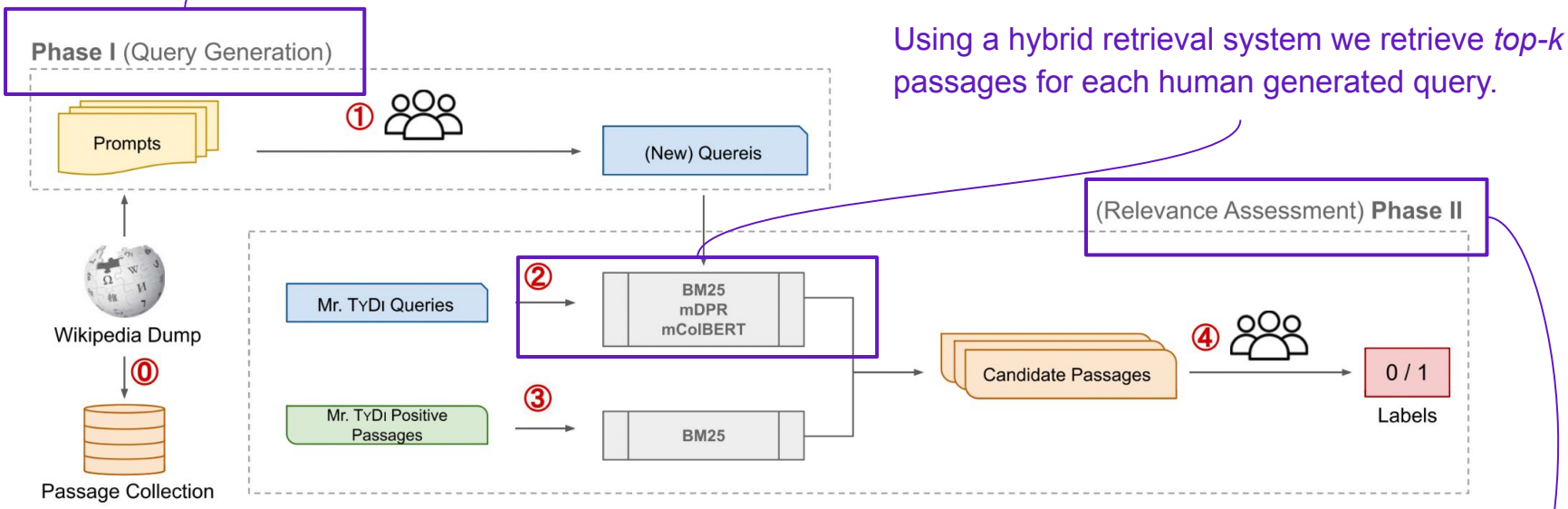
Language coverage in MIRACL



How did we construct MIRACL?

Annotators generate related questions for a short text snippet shown (prompt).

Using a hybrid retrieval system we retrieve *top-k* passages for each human generated query.



Annotators label the retrieved top-k passages for each query as relevant (label = 1) or non-relevant (label = 0).

Figure taken from MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. (https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00595/117438/MIRACL-A-Multilingual-Retrieval-Dataset-Covering)

Challenges faced during MIRACL construction

Questions generated in low-resource languages are unanswerable.

- In Yoruba only 5-10% of the generated queries contain at least a single relevant passage.
- What to do with queries with no-relevant passages?
- Lots of manual effort and $$ wasted ….

How to fill up holes and how many judgements per query?

- Which Retrievers to use to retrieve passages per query?
- Depth of human judgements. |k| = 5 or 10 or 100?

UNIVERSITY OF WATERLOO

Data-Efficient Multilingual Retrieval

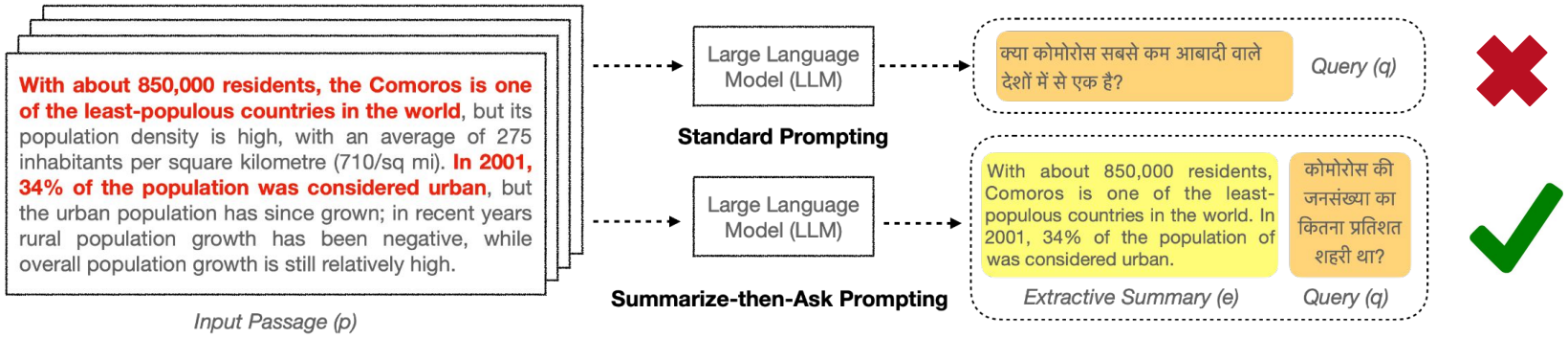
Generate synthetic data using LLM [10]

Standard prompting does not work well in multilingual settings

- Generated queries are rather **"uninformative"** or easily **answered**.

SAP Prompting with multilingual LLM

- Enhances LLM's ability to generate informative queries in various languages.
- Two steps: **summary extraction** and **query generation.**



SWIM-IR: Synthetic Multilingual IR Dataset

Advantages

- **Cost-effective alternative** to expensive human-labeled retrieval training data.
- Contains **language-specific topic diversity** over machine-translated (MT) English datasets.
- Covers **20 Indic languages**, e.g. Hindi, Tamil, Punjabi, Bengali, etc.

Dataset	Q Gen.	Cross.	Mono.	# L	Domain	# Train
NeuCLIR	Human	EN→L	L→L	3	News (hc4)	×
MKQA	Human	L→EN	×	26	Wikipedia	10K
mMARCO	Translate	×	L→L	13	MS MARCO	533K
Mr.TyDI	Human	×	L→L	11	Wikipedia	49K
MIRACL	Human	×	L→L	18	Wikipedia	726K
JH-POLO	GPT-3	EN→L	×	3	News (hc4)	78K
SWIM-IR	**PaLM 2**	**L→EN**	**L→L**	**33**	**Wikipedia**	**28M**

SWIM-IR contains about **37 times** more training pairs over existing multilingual datasets.

Figure showing comparison of SWIM-IR against other IR datasets.

Table taken from Leveraging LLMs for Synthesizing Training Data Across Many Languages in Multilingual Dense Retrieval. (https://arxiv.org/abs/2311.05800)

Example of a SWIM-IR training pair

(a) Cross-lingual Training Pair in SWIM-IR

Title: Menlo Park, New Jersey
Text: Menlo Park is an unincorporated community located within Edison Township in Middlesex County, New Jersey, United States. In 1876, Thomas Edison set up his home and research laboratory in Menlo Park, which at the time was the site of an unsuccessful real estate development named after the town of Menlo Park, California. While there, he earned the nickname "the Wizard of Menlo Park". The Menlo Park lab was significant in that it was one of the first laboratories to pursue practical commercial applications of research. It was in his Menlo Park laboratory that Thomas Edison invented the phonograph and developed it.

Passage (ID: 10770836) from English Wikipedia (*en*)

托马斯·爱迪生在哪里发明了留声机?

Translation: (Where did Thomas Edison invent the phonograph?)

LLM-generated Query in Chinese (*zh*)

(b) Monolingual Training Pair in SWIM-IR

Title: En la tierra del Guarán
Text: Es considerada una de las primeras realizaciones sonoras de la región y uno de los primeros antecedentes de cooperación entre dos países de la zona (Paraguay y Argentina) para la realización de un filme.

Translation: (*In the land of Guarán*: It is considered one of the first sound productions in the region and one of the first precedents of cooperation between two countries in the area (Paraguay and Argentina) for the making of a film.)

Passage (ID:spanish_5170543#3) from Spanish Wikipedia (*es*)

¿Qué película es una de las primeras realizaciones sonoras de la región?

Translation: (What film is one of the first sound films in the region?)

LLM-generated Query in Spanish (*es*)

Figure taken from Leveraging LLMs for Synthesizing Training Data Across Many Languages in Multilingual Dense Retrieval. . (https://arxiv.org/abs/2311.05800)

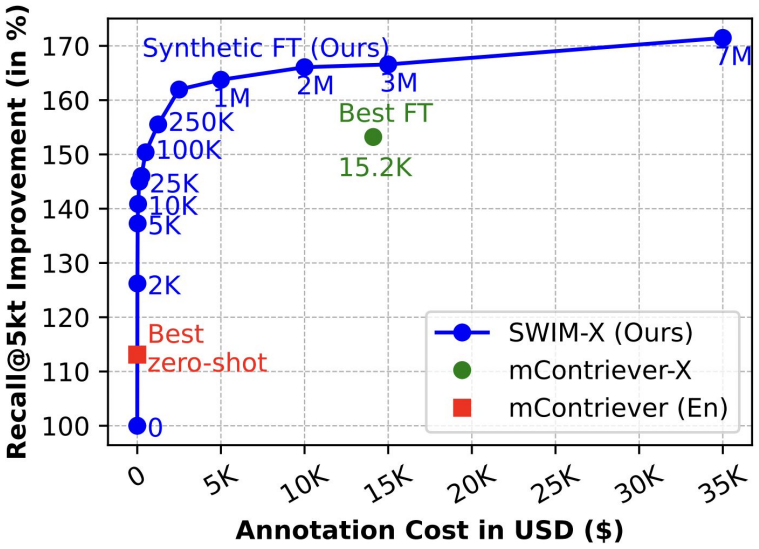
Fine-tuning on synthetic data is effective!

Model	Pretrain?	Fine-tune?	XOR-Retrieve (7 languages) Recall@5kt	MIRACL (18 languages) nDCG@10	XTREME-UP (20 languages) Avg. MRR@10
Zero-shot retrieval baselines (English Supervised data only)					
mContriever	mC4	-	38.9	-	-
mDPR (En)	-	MSMARCO	39.3	39.8	6.3
mContriever (En)	mC4	MSMARCO	44.0	37.8	7.9
Supervised FT baselines (Language-specific Supervised training data)					
mDPR-X	-	MSMARCO + FT	58.2	39.6	8.4
mContriever-X	mC4	MSMARCO + FT	59.6	**55.4**	12.4
Synthetic FT baselines (Language-specific Synthetic training data, i.e. no human-labeling)					
SWIM-X	mC4	SWIM-IR	**66.7**	46.4	**26.1**

Cost comparison and Indic language transfer

Our summarized findings

- SWIM-IR is 14 times cheaper and can outperform best mContriever with 250K synthetic pairs.
- Languages such as Hindi (hn), Kannada (kn) transfer well to other Indic languages.
- Very-low resource languages such as Boro (brx) or Manipuri (mni) cannot generate synthetic pairs.



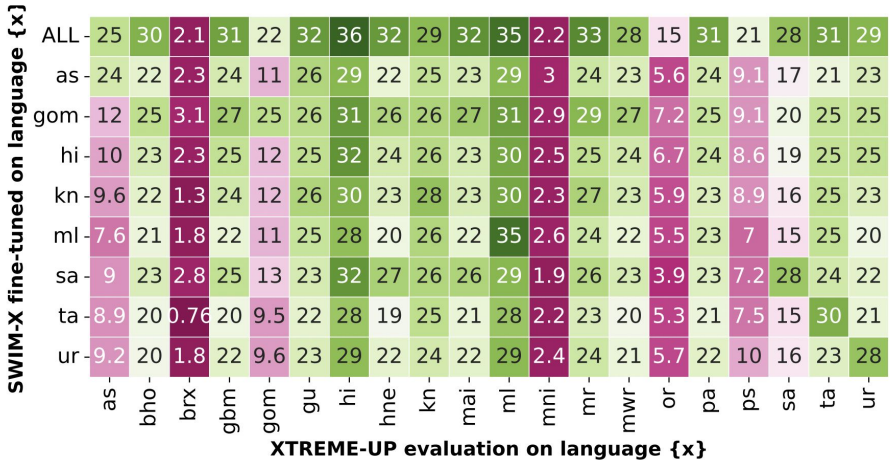


Figure taken from Leveraging LLMs for Synthesizing Training Data Across Many Languages in Multilingual Dense Retrieval. . (https://arxiv.org/abs/2311.05800)

UNIVERSITY OF WATERLOO

Open Research Problems and Conclusion

Open research problems

RQ: Are retrieval models overfitting on the BEIR benchmark?

- **E5-small/base/large** [11] all pre-train on Wikipedia, S2ORC, News, StackExchange datasets. These are all available in BEIR.
- Pre-trained E5 models on BEIR **perform worse** zero-shot on BEIR (few datasets) after fine-tuning on MSMARCO. Similar trends in others such as BGE [12].

RQ: LLM hallucinations in Retrieval-Augmented Generation (RAG) setup.

- **NoMIRACL** [16] is a multilingual dataset to evaluate **LLM hallucinations** in retrieval settings.
- To better evaluate RAG systems we are conducting a **TREC RAG challenge** in 2024!

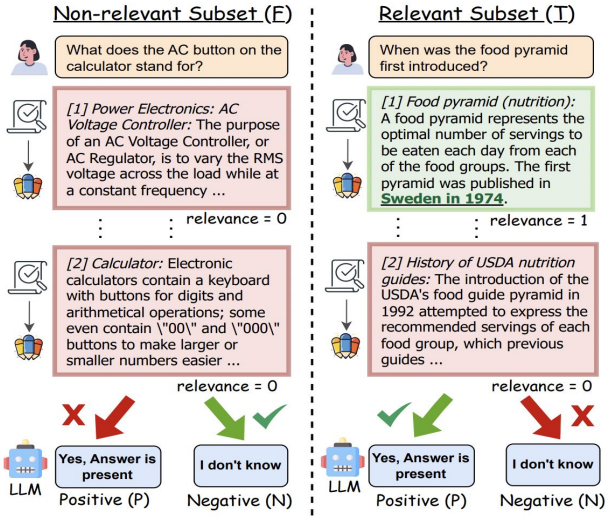
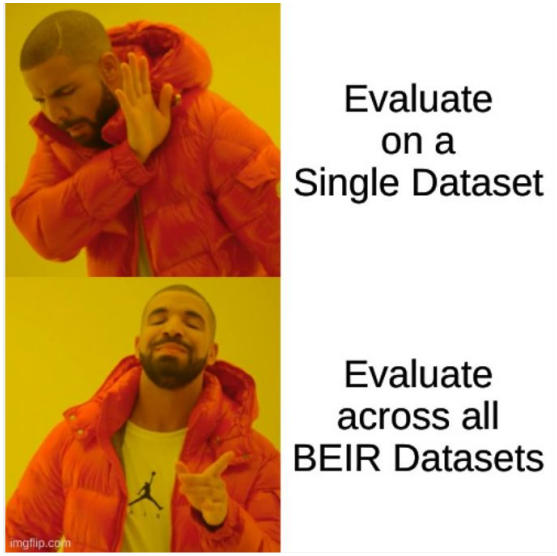


Figure taken from NoMIRACL: Knowing When You Don't Know for Robust Multilingual Retrieval-Augmented Generation. (https://arxiv.org/abs/2312.11361)

Conclusion

- **Limitations** found in models **during evaluation** on benchmarks help accelerate better model and progress to eliminate them!
- Always **evaluate your models** from a practical point of view, include challenging zero-shot tasks, domains and benchmarks!
- Do not **always chase** leaderboard (SoTA) improvement, especially on saturated leaderboards!
- There is no **free-lunch** to achieve good retrieval model generalizability across domains or languages.

Thank you for listening!



Open for Questions/Discussion

UNIVERSITY OF WATERLOO

References used in my presentation

References

[1] Stephen E. Robertson, Hugo Zaragoza: The Probabilistic Relevance Framework: BM25 and Beyond. Found. Trends Inf. Retr. 3(4): 333-389 (2009).

[2] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih: Dense Passage Retrieval for Open-Domain Question Answering. EMNLP (1) 2020: 6769-6781.

[3] Omar Khattab, Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. SIGIR 2020: 39-48.

[4] Rodrigo Frassetto Nogueira, Kyunghyun Cho. Passage Re-ranking with BERT. CoRR abs/1901.04085 (2019).

[5] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, Arnold Overwijk. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. ICLR 2021.

[6] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, Allan Hanbury. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. SIGIR 2021: 113-122.

References

[7] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeurIPS Datasets and Benchmarks 2021.

[8] Kexin Wang, **Nandan Thakur**, Nils Reimers, Iryna Gurevych. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. NAACL-HLT 2022: 2345-2360.

[9] Xinyu Zhang*, **Nandan Thakur***, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, Jimmy Lin. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. TACL 2023.

[10] Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Wieting, Jimmy Lin, Daniel Cer. Leveraging LLMs for Synthesizing Training Data Across Many Languages in Multilingual Dense Retrieval. CoRR abs/2311.05800 (2023).

[11] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, Furu Wei. Text Embeddings by Weakly-Supervised Contrastive Pre-training. CoRR abs/2212.03533 (2022).

References

[12] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighof. C-Pack: Packaged Resources To Advance General Chinese Embedding. CoRR abs/2309.07597 (2023).

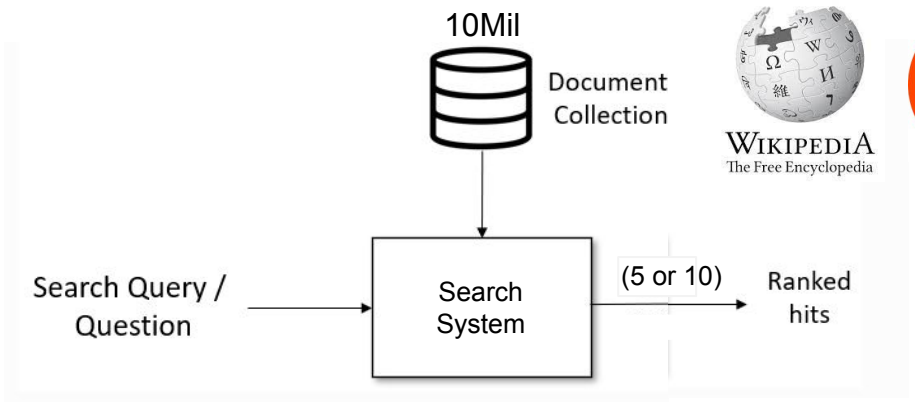
[13] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, Furu Wei. Improving Text Embeddings with Large Language Models. CoRR abs/2401.00368 (2024).

[14] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, Jimmy Lin. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. CoRR abs/2310.08319 (2023).

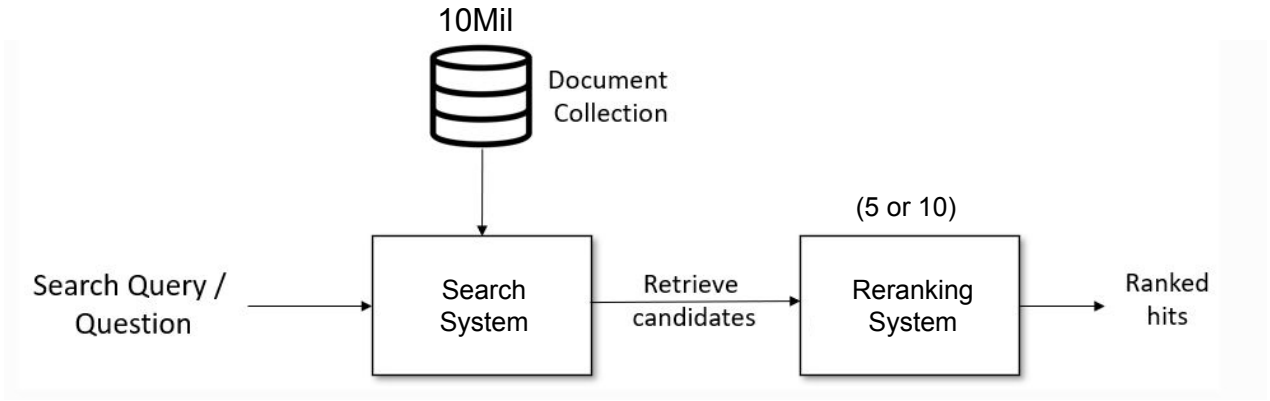
[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022.

[16] Nandan Thakur, Luiz Bonifacio, Xinyu Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, Jimmy Lin. NoMIRACL: Knowing When You Don't Know for Robust Multilingual Retrieval-Augmented Generation. CoRR abs/2312.11361 (2023).

Breaking down popular 🔍 IR Tasks



Retrieval



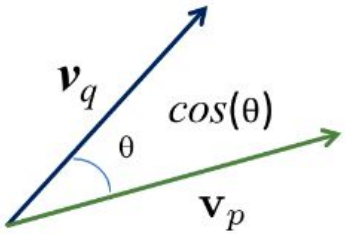
Retrieve and Rerank

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… 1 …. 1 …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.*

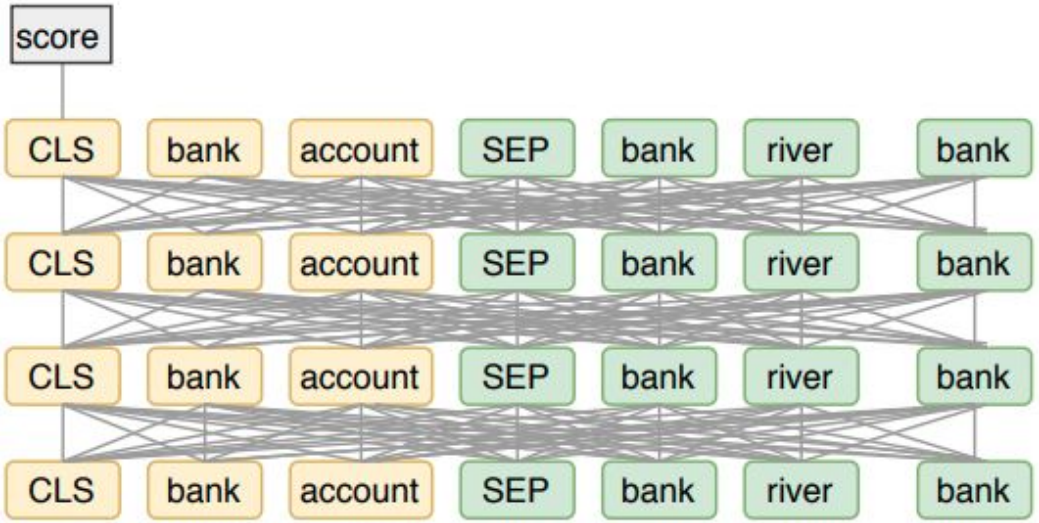
Ref: Danqi Chen, ACL 2020 OpenQA Tutorial
https://github.com/danqi/acl2020-openqa-tutorial/blob/master/slides/part5-dense-retriever-e2e-training.pdf

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

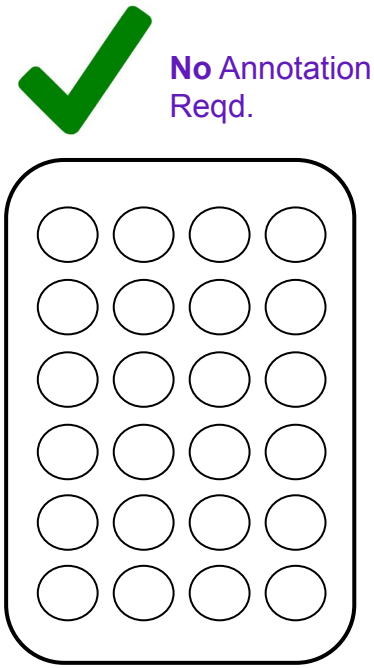
Performance (Cross-Encoder > Bi-Encoder > BM25)

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

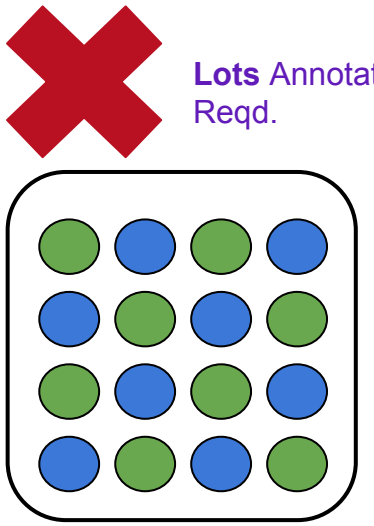
Why Zero-Shot Evaluation is Important?

Generating High-Quality Labeled Training Data is cumbersome!



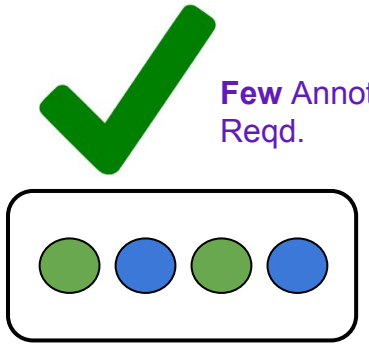
No Annotation Reqd.

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?

Bi-Encoders / Cross-Encoders

Trained on

Beir
Benchmarking IR

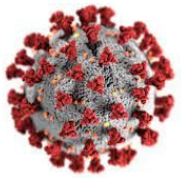
Domains with Training Data

Evaluated on

I can answer any question you have!

Domains without Training Data

Evaluated on



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

BEIR: Evaluation Benchmark for IR Systems

Diverse, Zero-shot retrieval benchmark with 18 datasets and task

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	Bio-Medical	TREC-COVID [63]	✓	3-level	—-	—-	50	171,332	493.5	10.60	160.77
Information	Bio-Medical	NFCorpus [7]	✓	3-level	110,575	324	323	3,633	38.2	3.30	232.26
Retrieval (IR)	Bio-Medical	BioASQ [59]	✓	Binary	32,916	—-	500	14,914,602	4.7	8.05	202.61
Question	Wikipedia	NQ [32]	✓	Binary	132,803	—-	3,452	2,681,468	1.2	9.16	78.88
Answering	Wikipedia	HotpotQA [74]	✓	Binary	170,000	5,447	7,405	5,233,329	2.0	17.61	46.30
(QA)	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	News	TREC-NEWS [56]	✓	5-level	—-	—-	57	594,977	19.6	11.14	634.79
Retrieval	News	Robust04 [62]	✗	3-level	—-	—-	249	528,155	69.9	15.27	466.40
Argument	Misc.	ArguAna [65]	✓	Binary	—-	—-	1,406	8,674	1.0	192.98	166.80
Retrieval	Misc.	Touché-2020 [6]	✓	3-level	—-	—-	49	382,545	19.0	6.55	292.37
Duplicate-Question	StackEx.	CQADupStack [23]	✓	Binary	—-	—-	13,145	457,199	1.4	8.59	129.09
Retrieval	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
	Wikipedia	FEVER [58]	✓	Binary	140,085	6,666	6,666	5,416,568	1.2	8.13	84.76
Fact Checking	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)		128	350ms	–	20GB
(3)	BM25	–	–	20ms	0.4GB
(4)	Bi-Encoders	768	14ms	125ms	3GB
(5)		768	20ms	275ms	3GB
(6)		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.

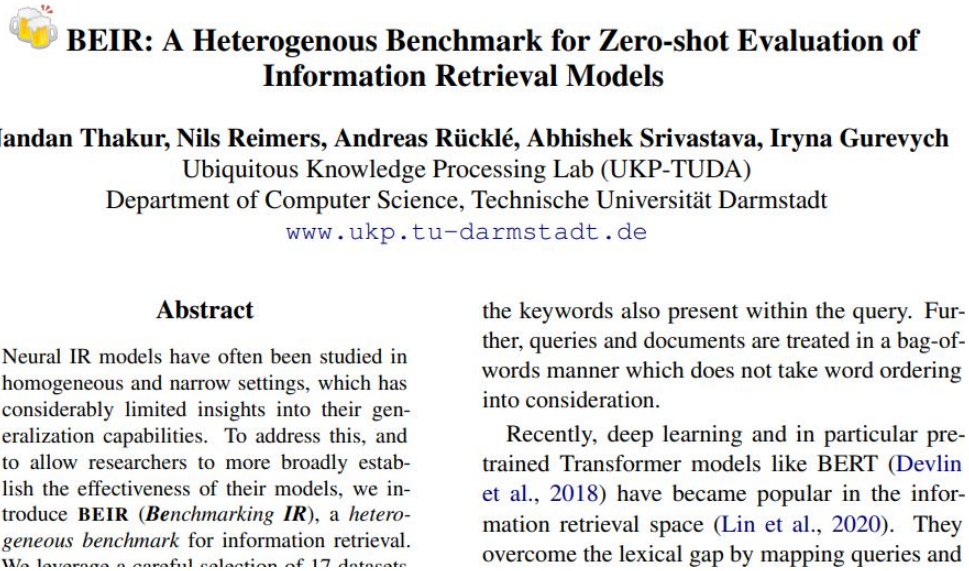
2. Bi-Encoders map query and document to a dense vector space, efficient and practical. However, they fail to perform well in zero-shot setting and are unable to generalize well!

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

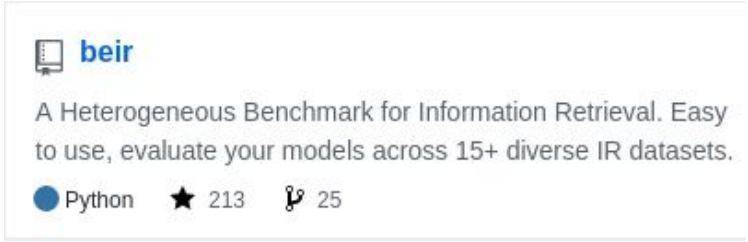
Abstract

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. We leverage a careful selection of 17 datasets

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



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

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

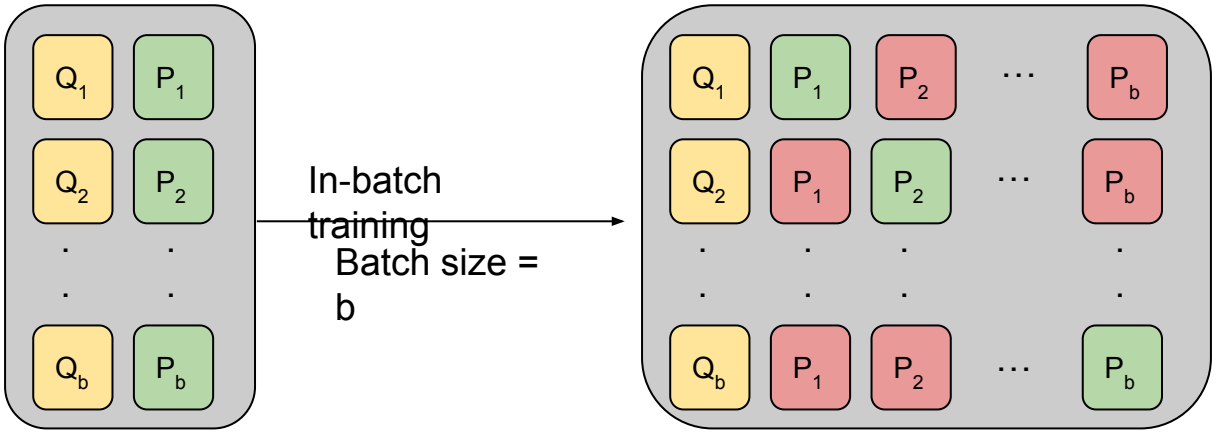
GLUE

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

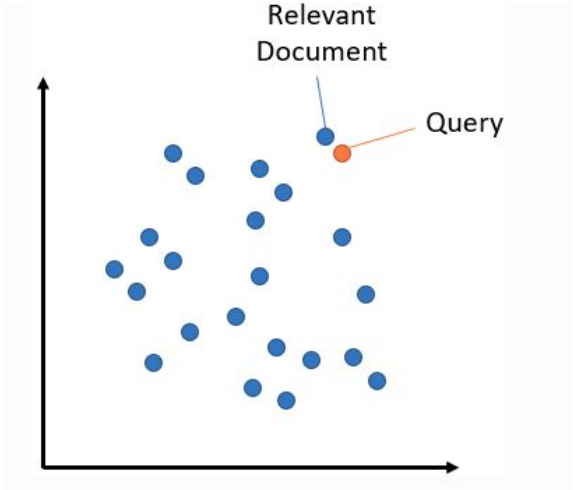
How to fine-tune Bi-Encoder model?

Method 1: Inbatch fine-tuning with random inbatch negatives

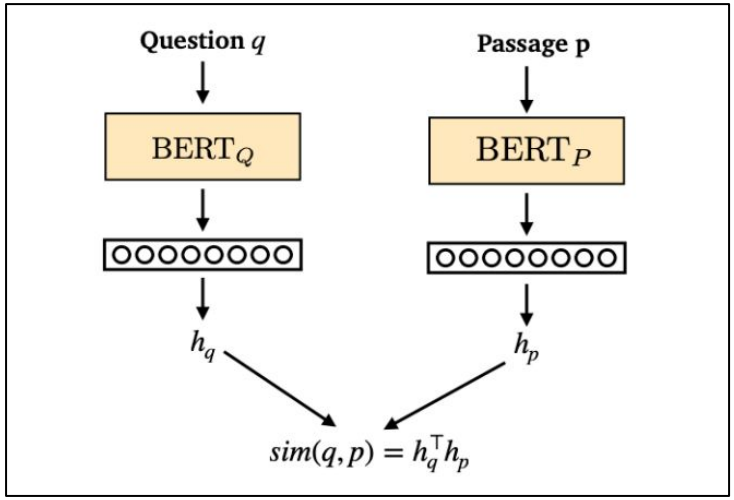


Cross-Entropy loss function

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



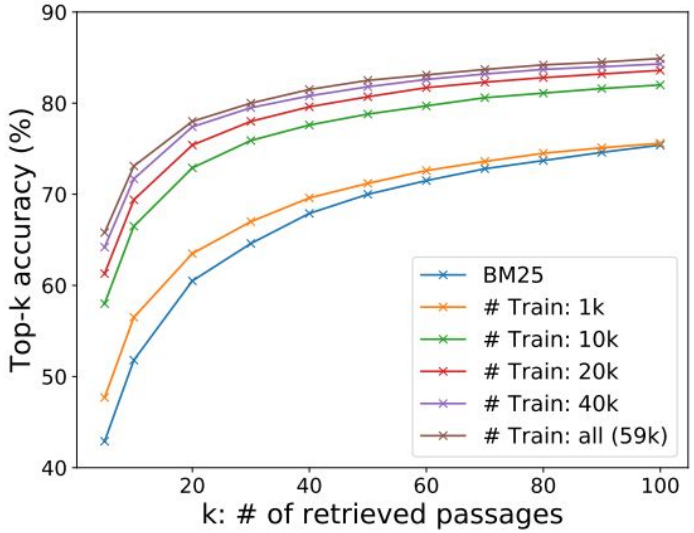
DPR: Dense Passage Retriever (kharpurkin et al. 2020)



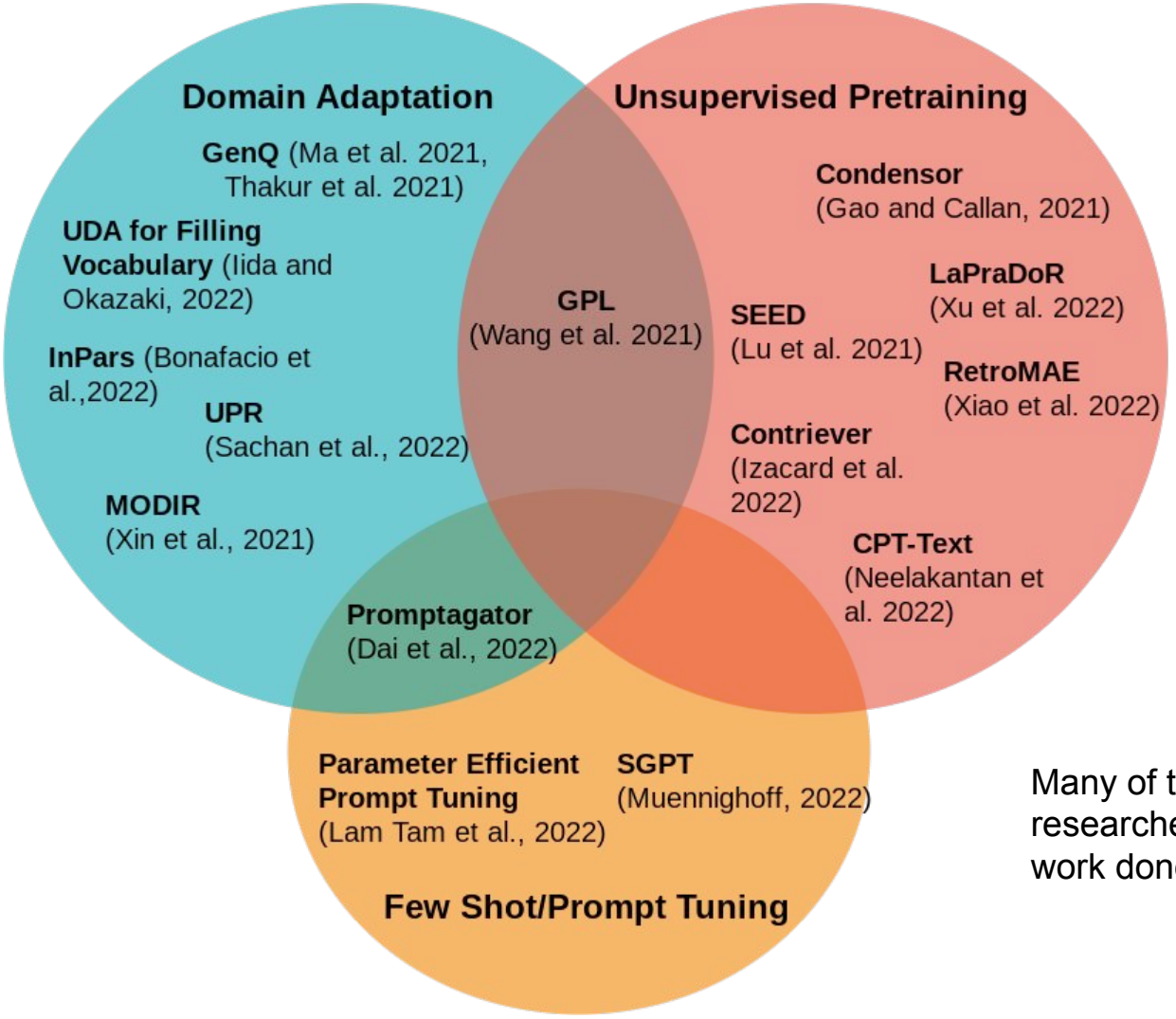
$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)$$
$$= -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}}$$

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

Natural Questions (Kwiatkowski et al., 2019)

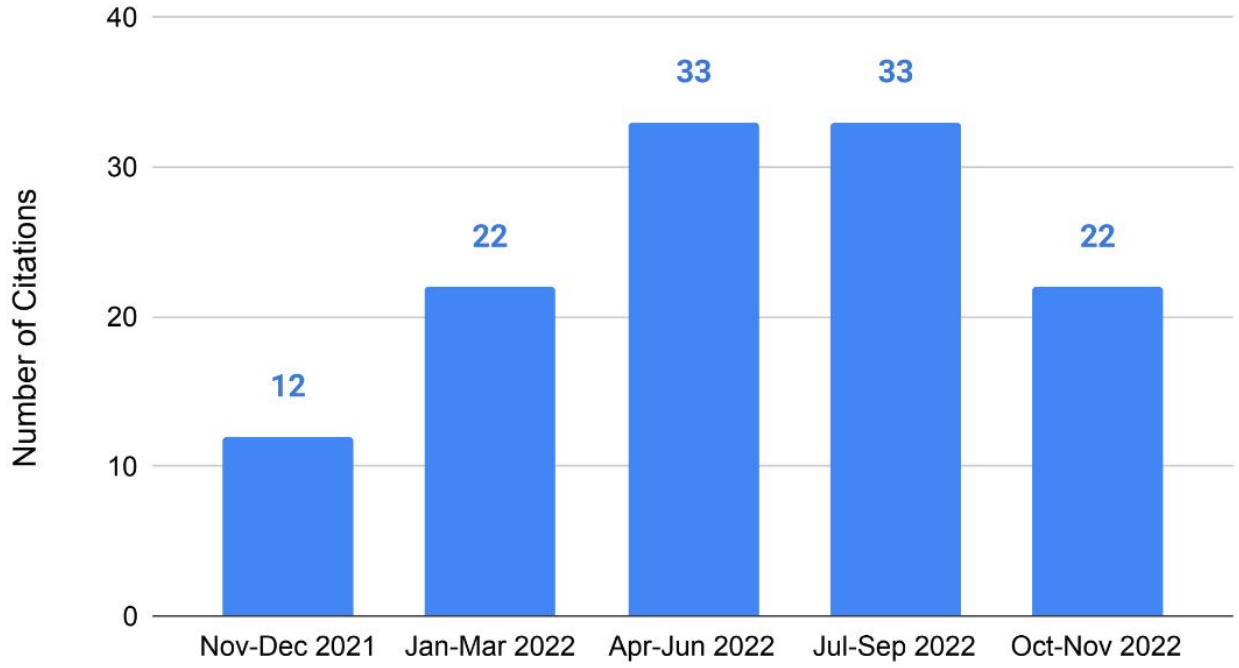


Summary of Recent Works to Improve Bi-Encoder Generalization



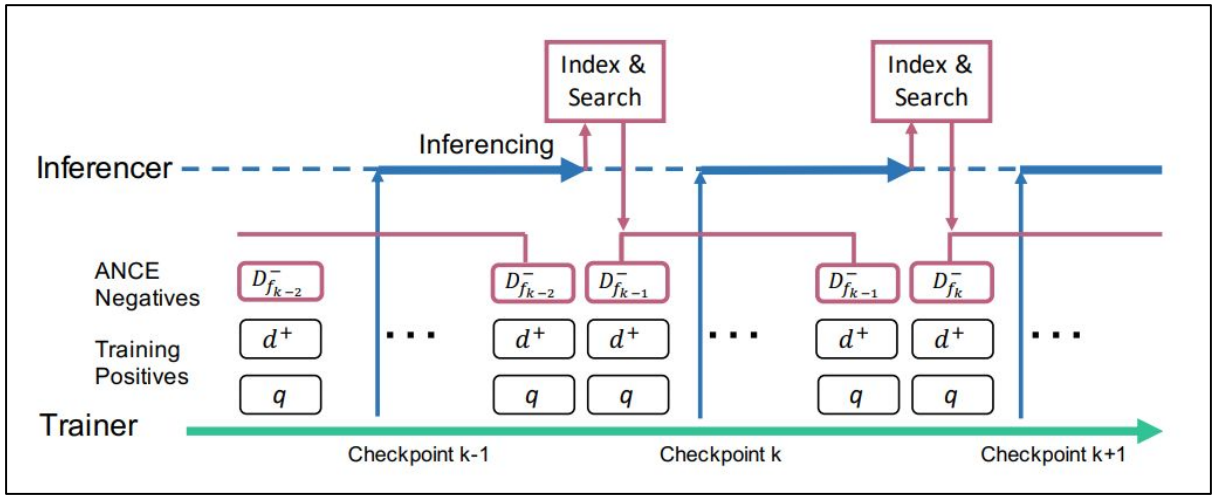
Many of these ideas (by other researchers) got inspired by work done in BEIR :)

BEIR Benchmark Outreach on Zero-shot English IR

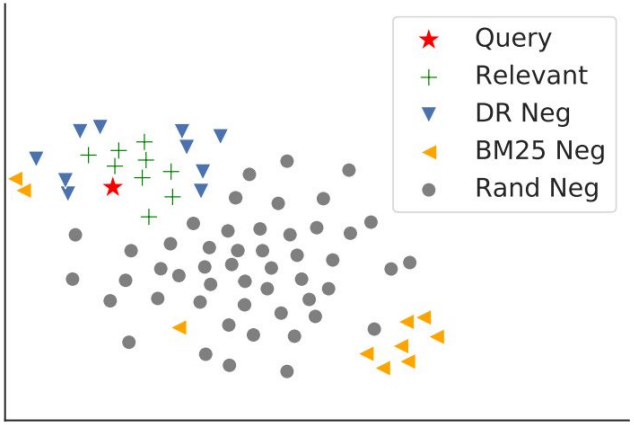


Arxiv	**60**	ECIR	**4**
SIGIR	**10**	ACL	**3**
CIKM	**7**	FINDINGS	**3**
NAACL	**6**	NAACL-HLT	**2**

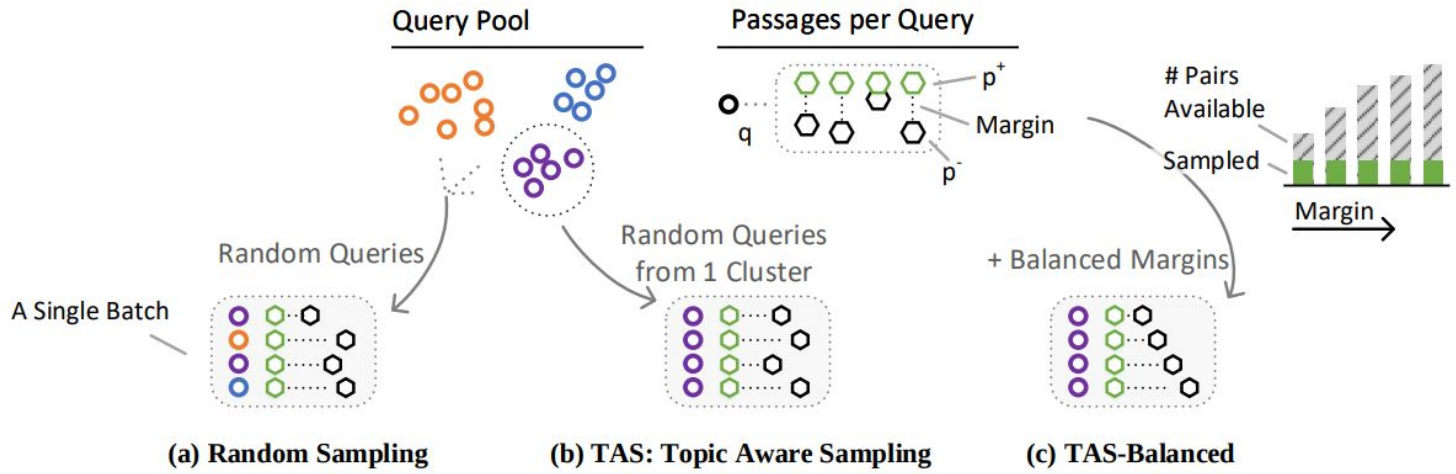
ANCE: Approximate Nearest Neighbor Negative Contrastive Learning (Xiong et al. 2021)

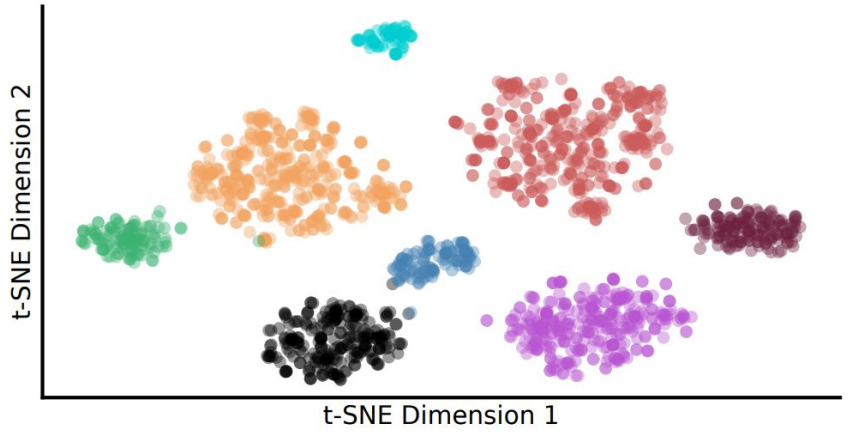


$$\theta^* = \text{argmin}_\theta \sum_q \sum_{d^+ \in D^+} \sum_{d^- \in D^-_{\text{ANCE}}} l(f(q, d^+), f(q, d^-)),$$



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





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

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

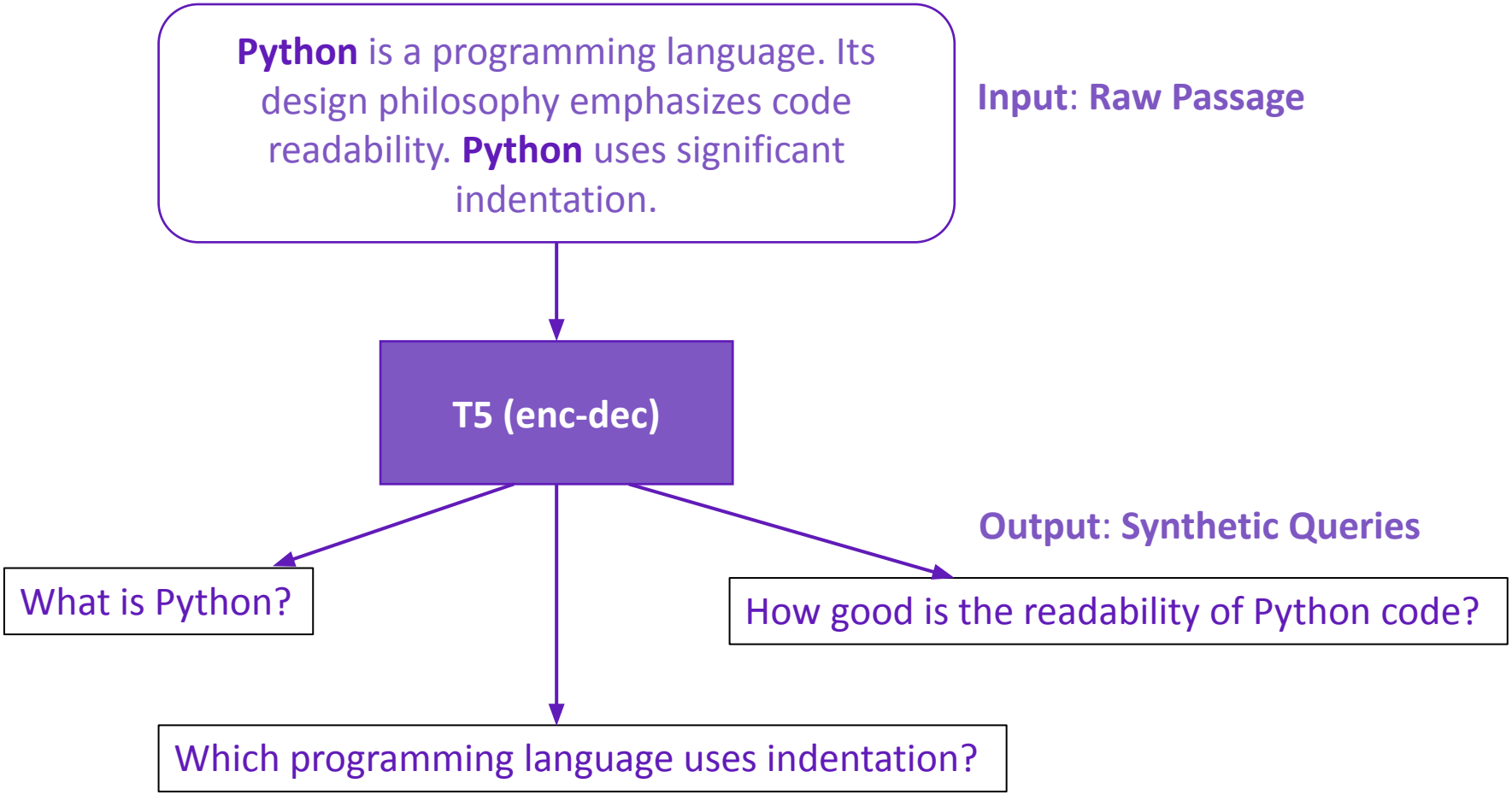
Performance of Bi-Encoders >> BM25

DPR (kharpurkin et al. 2020)	**BM25**	NQ Retrieval	⬆ **20**
ANCE (Xiong et al. 2021)	**BM25**	MSM	
TAS-B (Hofstät et al. 20			

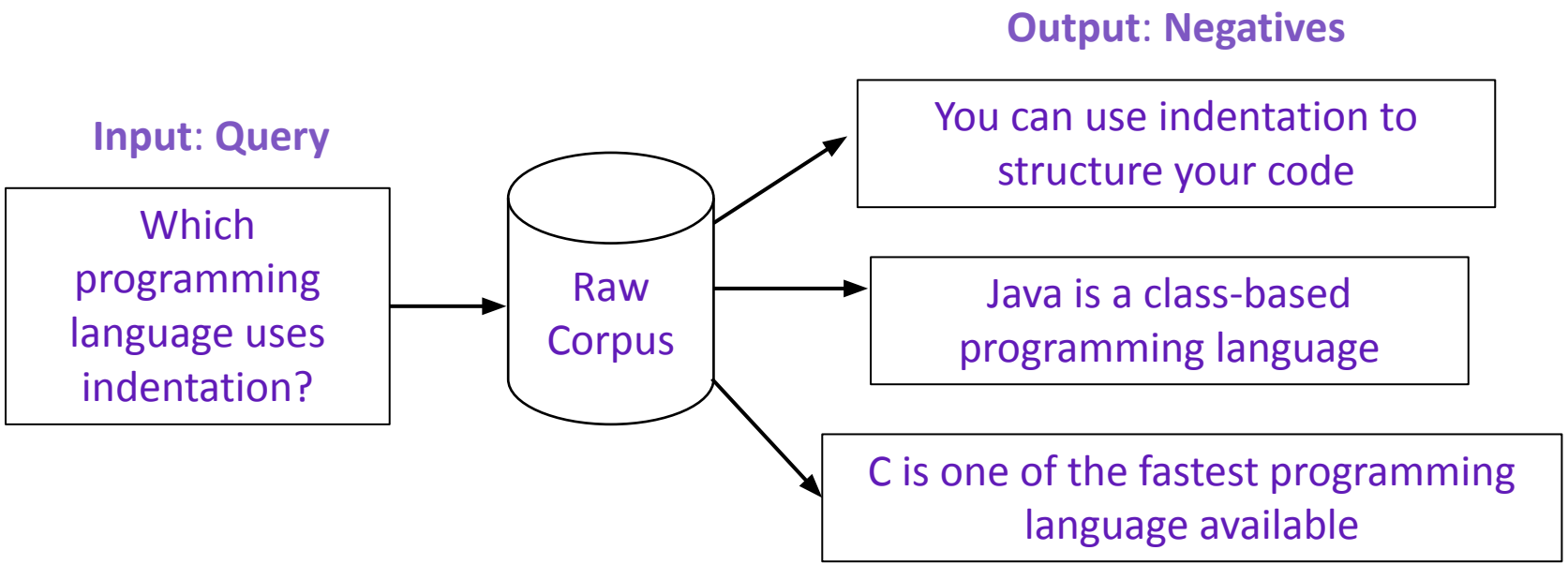
NO Broken Evaluation!

Training	Re-Ranking		Latency	TREC-DL'19			TREC-DL'20			MSMARCO DEV		
	Model	#	(ms)	nDCG@10	MRR@10	R@1K	nDCG@10	MRR@10	R@1K	nDCG@10	MRR@10	R@1K
Lo… 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}

GPL Step 1: Generate Queries



GPL Step 2: Mine Negatives

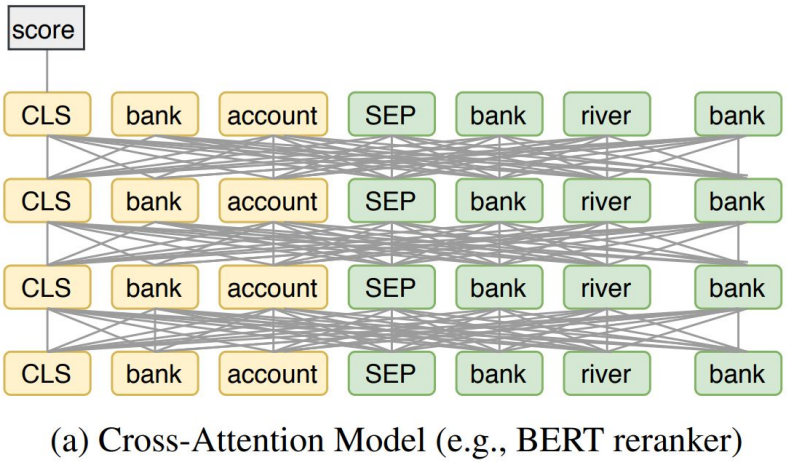


GPL Step 3: Label using Cross-Encoder

(Query, Doc1)

(Query, Doc2)

(Query, Doc3)



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

5.2

0.1

-3.8

Cross-Encoder

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.

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?

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)

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.

Issues with large-scale supervised training data

-
- 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!

Research Problems being faced recently

- 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!

Roadmap of my research journey

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

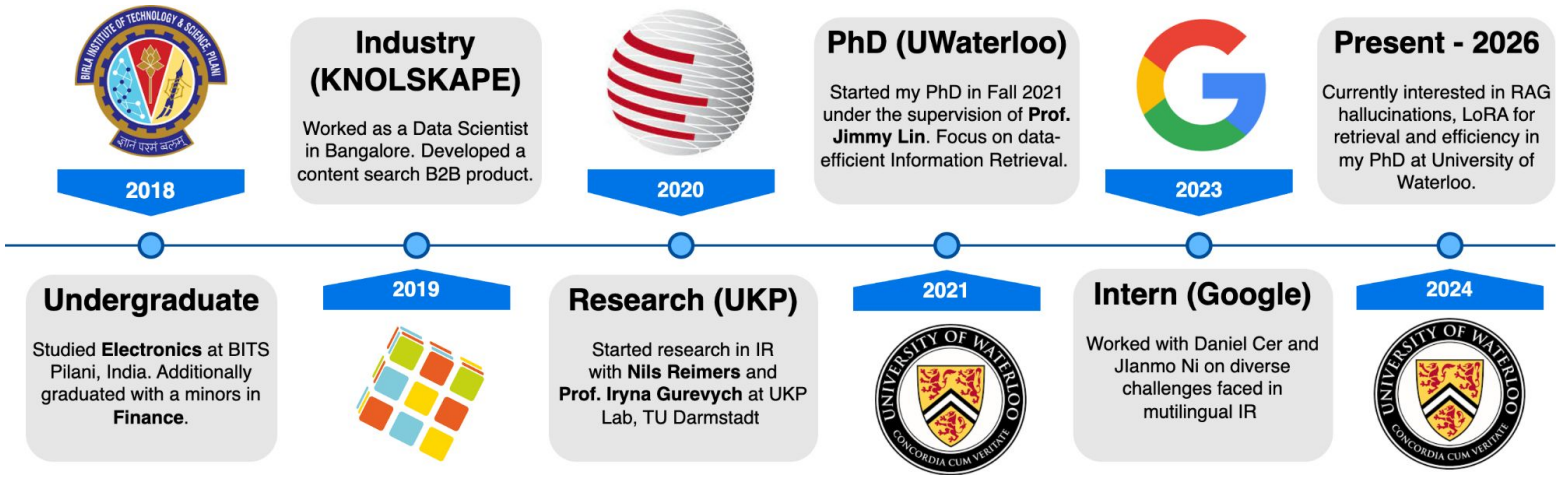


Figure: My research journey roadmap starting from 2018 until present.

MIRACL dataset statistics

- MIRACL contains over 100+ million passages, 18 languages (both low and high resource) and 343K training pairs overall.

Lang	ISO	Train		Dev		Test-A		Test-B		# Passages	# Articles	Avg. Q Len	Avg. P Len
		# Q	# J	# Q	# J	# Q	# J	# Q	# J				
Arabic	ar	3,495	25,382	2,896	29,197	936	9,325	1,405	14,036	2,061,414	656,982	6	54
Bengali	bn	1,631	16,754	411	4,206	102	1,037	1,130	11,286	297,265	63,762	7	56
English	en	2,863	29,416	799	8,350	734	5,617	1,790	18,241	32,893,221	5,758,285	7	65
Finnish	fi	2,897	20,350	1,271	12,008	1,060	10,586	711	7,100	1,883,509	447,815	5	41
Indonesian	id	4,071	41,358	960	9,668	731	7,430	611	6,098	1,446,315	446,330	5	49
Japanese	ja	3,477	34,387	860	8,354	650	6,922	1,141	11,410	6,953,614	1,133,444	17	147
Korean	ko	868	12,767	213	3,057	263	3,855	1,417	14,161	1,486,752	437,373	4	38
Russian	ru	4,683	33,921	1,252	13,100	911	8,777	718	7,174	9,543,918	1,476,045	6	46
Swahili	sw	1,901	9,359	482	5,092	638	6,615	465	4,620	131,924	47,793	7	36
Telugu	te	3,452	18,608	828	1,606	594	5,948	793	7,920	518,079	66,353	5	51
Thai	th	2,972	21,293	733	7,573	992	10,432	650	6,493	542,166	128,179	42	358
Spanish	es	2,162	21,531	648	6,443	–	–	1,515	15,074	10,373,953	1,669,181	8	66
Persian	fa	2,107	21,844	632	6,571	–	–	1,476	15,313	2,207,172	857,827	8	49
French	fr	1,143	11,426	343	3,429	–	–	801	8,008	14,636,953	2,325,608	7	55
Hindi	hi	1,169	11,668	350	3,494	–	–	819	8,169	506,264	148,107	10	69
Chinese	zh	1,312	13,113	393	3,928	–	–	920	9,196	4,934,368	1,246,389	11	121
German	de	–	–	305	3,144	–	–	712	7,317	15,866,222	2,651,352	7	58
Yoruba	yo	–	–	119	1,188	–	–	288	2,880	49,043	33,094	8	28
Total		40,203	343,177	13,495	130,408	7,611	76,544	17,362	174,496	106,332,152	19,593,919	10	77

Table taken from MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. (https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00595/117438/MIRACL-A-Multilingual-Retrieval-Dataset-Covering)

MIRACL cross-lingual transfer

Cross-lingual transfer capabilities are shown in language-specific retrieval models.

- Within language fine-tuning leads to best retrieval performance with mDPR model.
- Languages within same family or same sub-family overall transfer well among each other.

RQ: How much bilingual (mixed with English) data is present within Finnish Wikipedia? Is it because of **english tokens** for e.g. Finnish mDPR performs well across all evaluated languages?

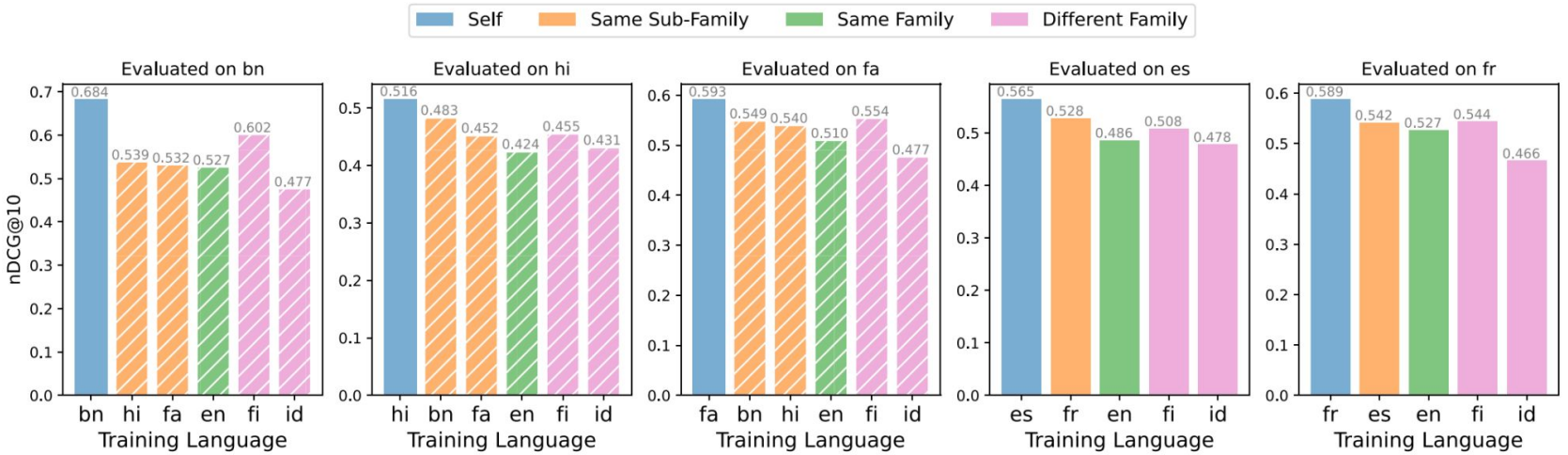


Table taken from MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. (https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00595/117438/MIRACL-A-Multilingual-Retrieval-Dataset-Covering)

How to improve Bi-encoder generalization?

As training data is scarce, focus on unsupervised techniques!

(1) Unsupervised Domain Adaptation

- Generate synthetic queries and use query-passage pairs across each domain.
- Fine-tune a separate model to adapt across each domain/dataset.

(2) Unsupervised Pre-training

- Pretraining in a self-supervised fashion across (a lot) of raw data.
- Techniques also involve a light decoder setup, training in an autoencoder setup.

(3) 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.

Challenges 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)

Credits goes to Akari Asai. Taken from “Towards Better Multilingual Information Access” (https://akariasai.github.io/files/amazon_talk_akari_06312022.pdf)

Open research problems

RQ: How to improve generalizability of retrieval models?

- **Scaling up** models to include more parameters such as 7B perform better on BEIR.
- Uses a decoder-only LLM as a dense retrieval: **E5-Mistral-7B** [13], **RepLLAMA-7B** [14]
- **[EOS]** token *</s>* is used for representation. Efficient fine-tuning is done using **LoRA** [15].

$$V_T = \textbf{Decoder}(t_1, t_2, \ldots, t_k, \text{</s>})[-1]$$

	BM25	GTR-XXL	cpt-text-XL	Ada2	SGPT	RepLLaMA
model size	-	4.8B	175B	?	5.8B	7B
add. pretrain	-	Y	Y	?	Y	N
Arguana	39.7	54.0	43.5	56.7	51.4	48.6
Climate-FEVER	16.5	26.7	22.3	23.7	30.5	31.0
DBPedia	31.8	40.8	43.2	40.2	39.9	43.7
FEVER	65.1	74.0	77.5	77.3	78.3	83.4
FiQA	23.6	46.7	51.2	41.1	37.2	45.8
HotpotQA	63.3	59.9	68.8	65.4	59.3	68.5
NFCorpus	32.2	34.2	40.7	35.8	36.2	37.8
NQ	30.6	56.8	-	48.2	52.4	62.4
Quora	78.9	89.2	63.8	87.6	84.6	86.8
SCIDOCS	14.9	16.1	-	18.6	19.7	18.1
SciFact	67.9	66.2	75.4	73.6	74.7	75.6
TREC-COVID	59.5	50.1	64.9	81.3	87.3	84.7
Touche-2020	44.2	25.6	29.1	28.0	25.4	30.5
Average	43.7	49.3	-	52.1	52.1	**55.1**

Table taken from Fine-Tuning LLaMA for Multi-Stage Text Retrieval. (https://arxiv.org/abs/2310.08319)

# of datasets →	Class. 12	Clust. 11	PairClass. 3	Rerank 4	Retr. 15
Unsupervised Models					
Glove [35]	57.3	27.7	70.9	43.3	21.6
$\text{SimCSE}_{\text{bert-unsup}}$ [13]	62.5	29.0	70.3	46.5	20.3
Supervised Models					
$\text{SimCSE}_{\text{bert-sup}}$ [13]	67.3	33.4	73.7	47.5	21.8
Contriever [18]	66.7	41.1	82.5	53.1	41.9
GTR_{xxl} [32]	67.4	42.4	86.1	56.7	48.5
$\text{Sentence-T5}_{\text{xxl}}$ [31]	73.4	43.7	85.1	56.4	42.2
$\text{E5}_{\text{large-v2}}$ [46]	75.2	44.5	86.0	56.6	50.6
$\text{GTE}_{\text{large}}$ [23]	73.3	46.8	85.0	59.1	52.2
$\text{BGE}_{\text{large-en-v1.5}}$ [48]	76.0	46.1	87.1	60.0	54.3
Ours					
$\text{E5}_{\text{mistral-7b}}$ + full data	**78.5**	50.3	**88.3**	**60.2**	**56.9**
w/ synthetic data only	78.2	**50.5**	86.0	59.0	46.9
w/ synthetic + msmarco	78.3	49.9	87.1	59.5	52.2

Table taken from Improving Text Embeddings with Large Language Models. (https://arxiv.org/abs/2401.00368)

Open research problems

Retrieval-Augmented Generation (RAG) is popular in NLP research.

- **Retrieval Stage:** retrieval model provides top-k passages for a given query.
- **Generation Stage:** LLM model which provides a summarized answer from retrieved passages.

Issue: LLM hallucinates (makes up) an answer, when actually there is none present.

- **NoMIRACL** [16] is a multilingual dataset to evaluate **LLM hallucinations** in retrieval settings.
- To better evaluate RAG systems we are conducting a **TREC RAG challenge** in 2024! Stay tuned!

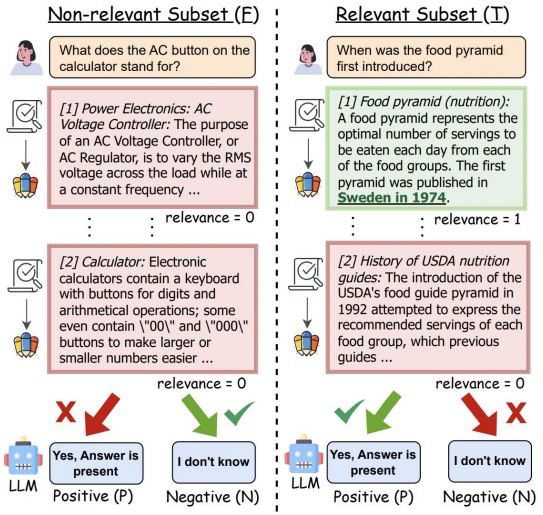


Figure taken from NoMIRACL: Knowing When You Don't Know for Robust Multilingual Retrieval-Augmented Generation. (https://arxiv.org/abs/2312.11361)

Generate synthetic data using LLM [10]

Human-labeled training data

- Expensive, cumbersome and labor-intensive and requires native expert annotators.

Alternative: Generative synthetic training data

- Cheaper and quicker to generate and easily extends across a wide variety of languages.
- Fine-tuning **multilingual T5 not feasible** due to scarce/uneven training data.