Zeta Alpha

Dense Neural Retrieval for Scientific Documents at Zeta Alpha

ECIR 2022 Industry Day, Stavanger - April 14th 2022

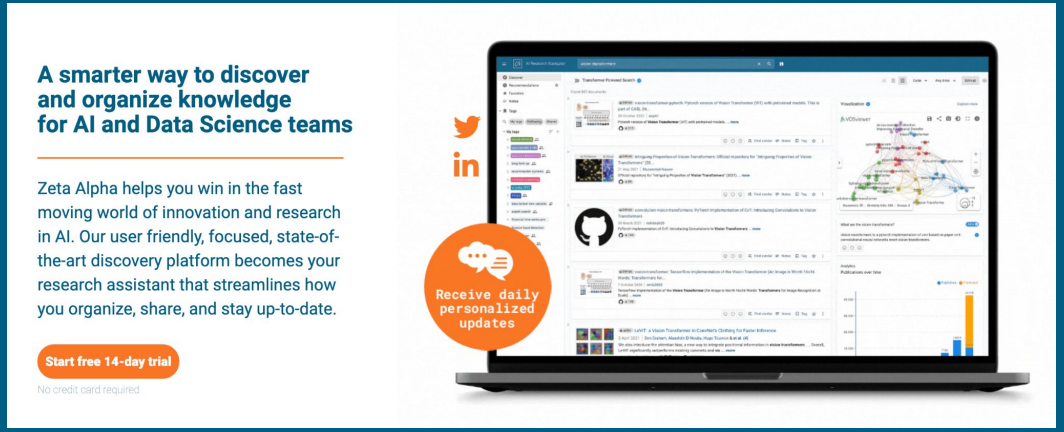
Jakub Zavrel, Marzieh Fadaee, Artem Grotov and Rodrigo Nogueira

ECIR 2022

Zeta Alpha

Made in Amsterdam, since 2019.

Leverage Neural Language Understanding to help people make better AI supported decisions.



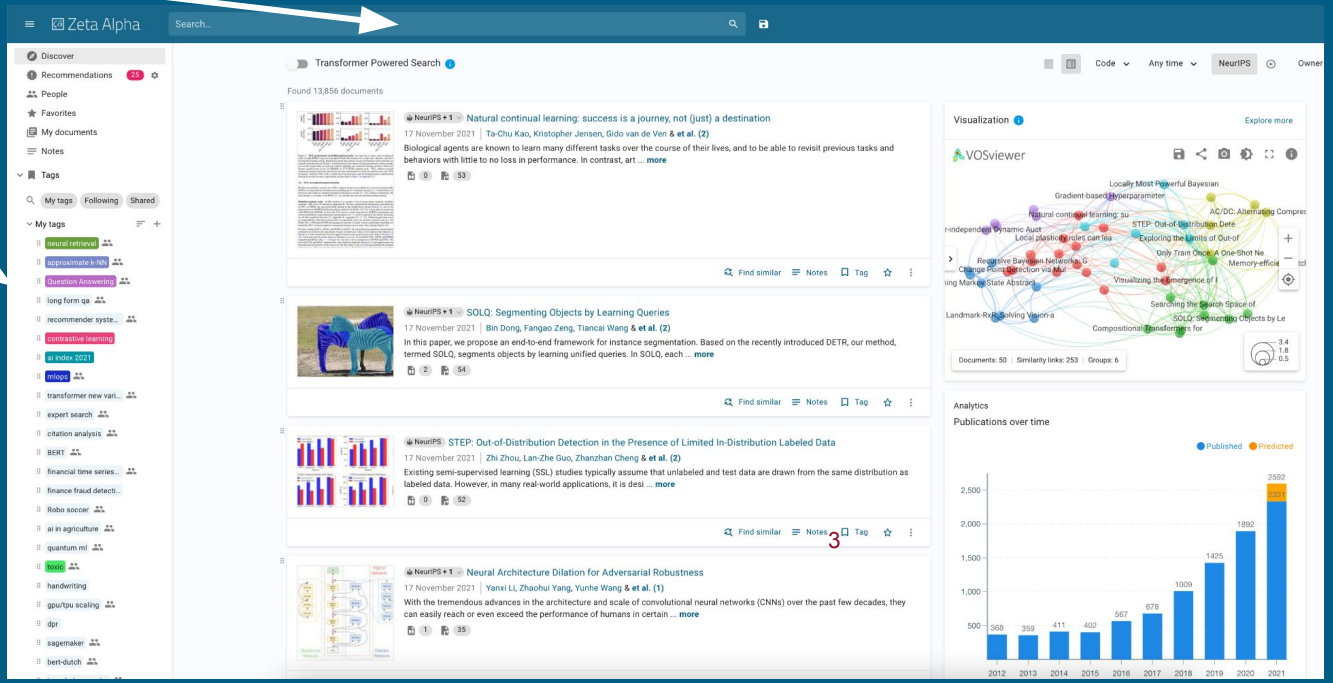
Zeta Alpha AI discovery platform

① Discover content using neural search, find similar, visualization, code popularity and social media.

② Organize your work in personalized topic tags

③ Receive timely need to know recommendations tailored to your interests

④ Share and re-use knowledge with your team



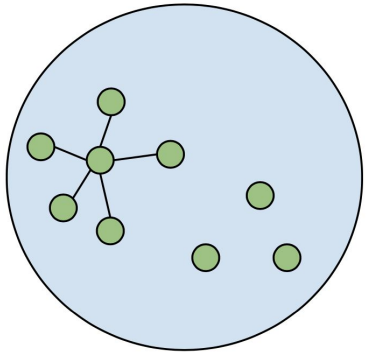
AI focused technical content from arXiv, conferences, companies, blogs, news, github code, twitter

+ import private data

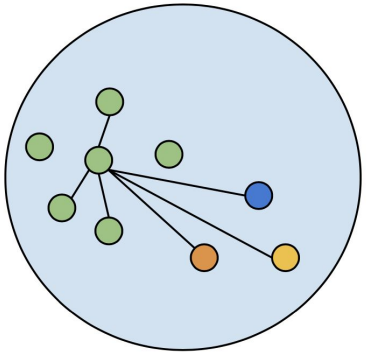
Zeta Alpha

Why Neural Search?

- Semantic understanding of data as opposed to surface keywords: bridge *the* ***lexical gap***
- Context and relationships crucial in interpreting meaning: ***handles complex and relational queries***
- Unstructured data accessible without classification and taxonomies, even ***multi-lingual*** and ***cross-lingual***
- ***Multi-modal*** capabilities: potential to combine text, audio, images and video

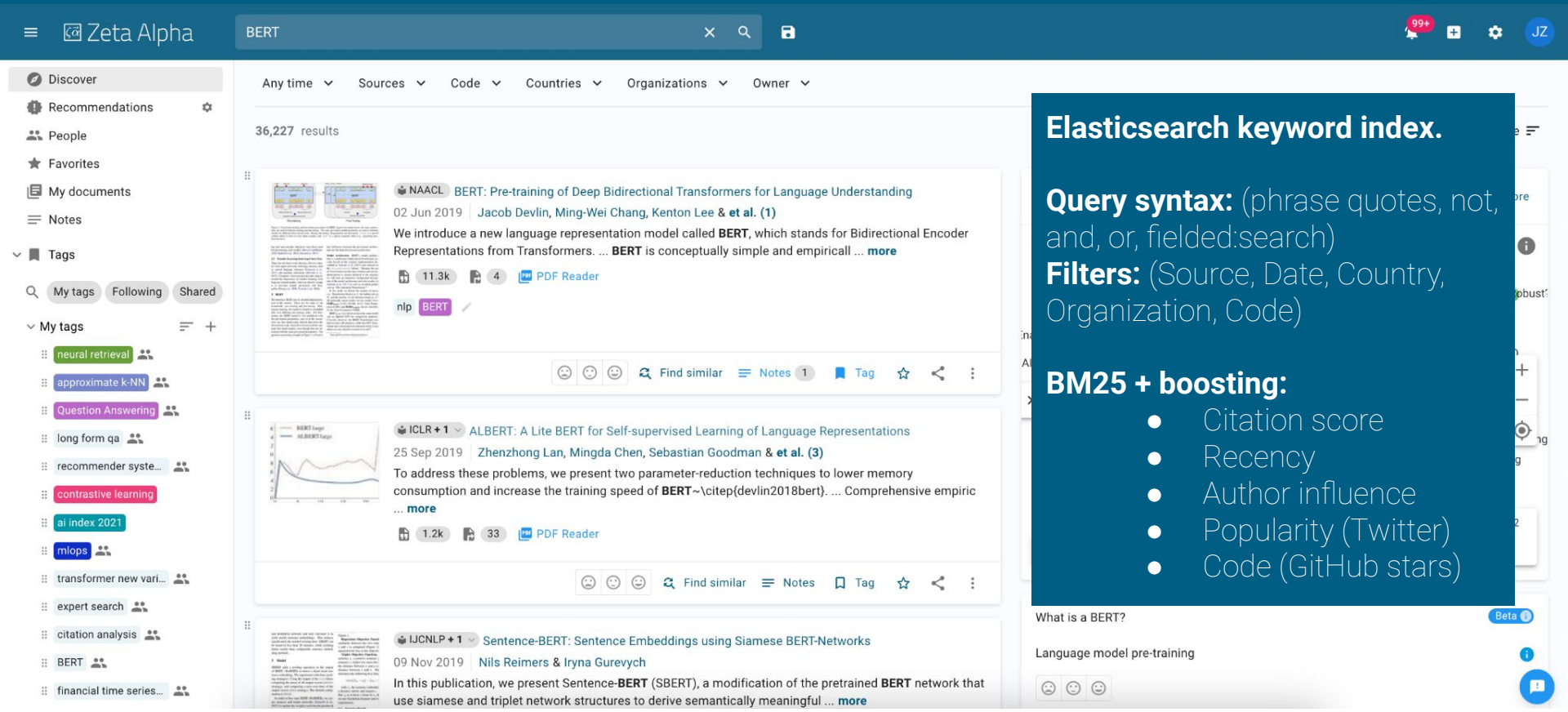


searching



expanding the horizon of exploration

A well tuned BM25 is hard to beat...



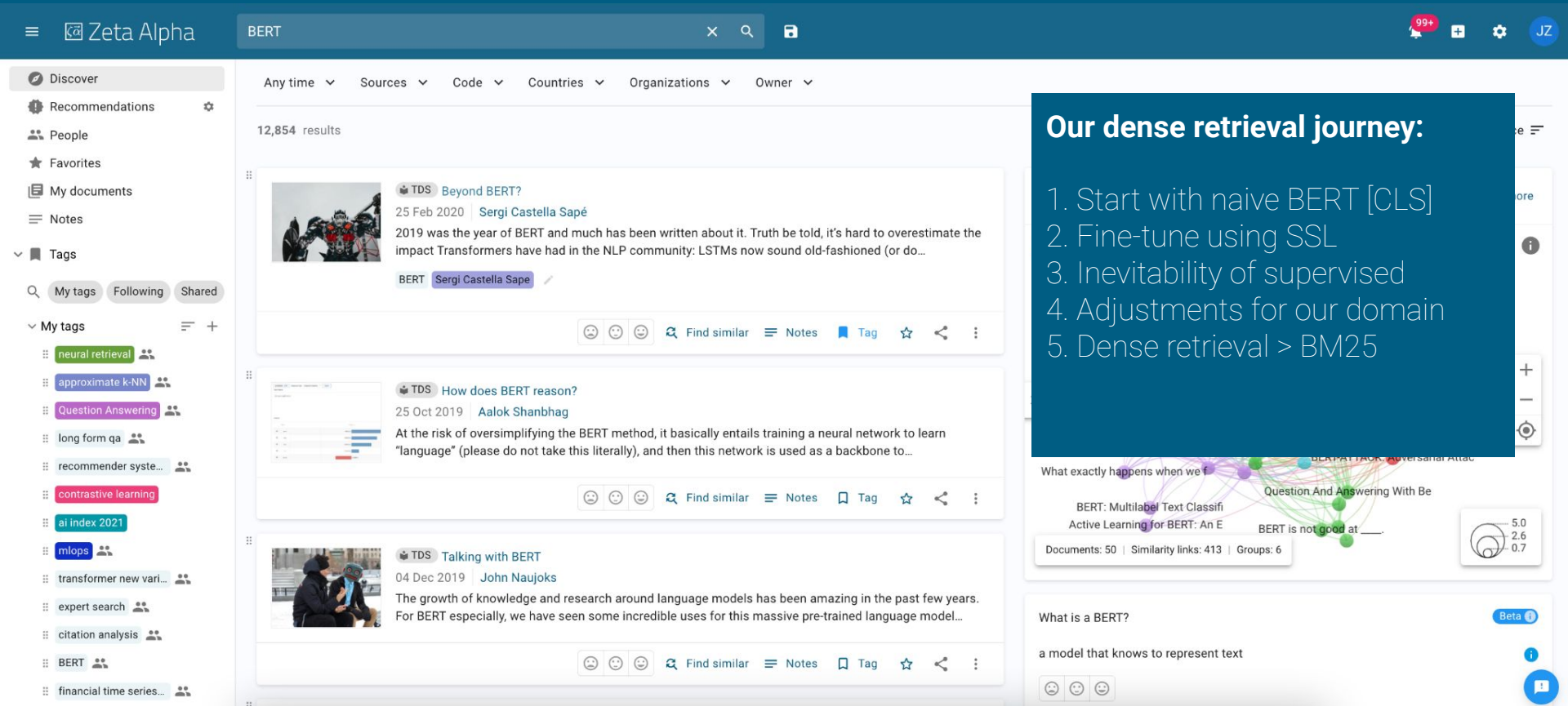
Elasticsearch keyword index.

Query syntax: (phrase quotes, not, and, or, fielded:search)
Filters: (Source, Date, Country, Organization, Code)

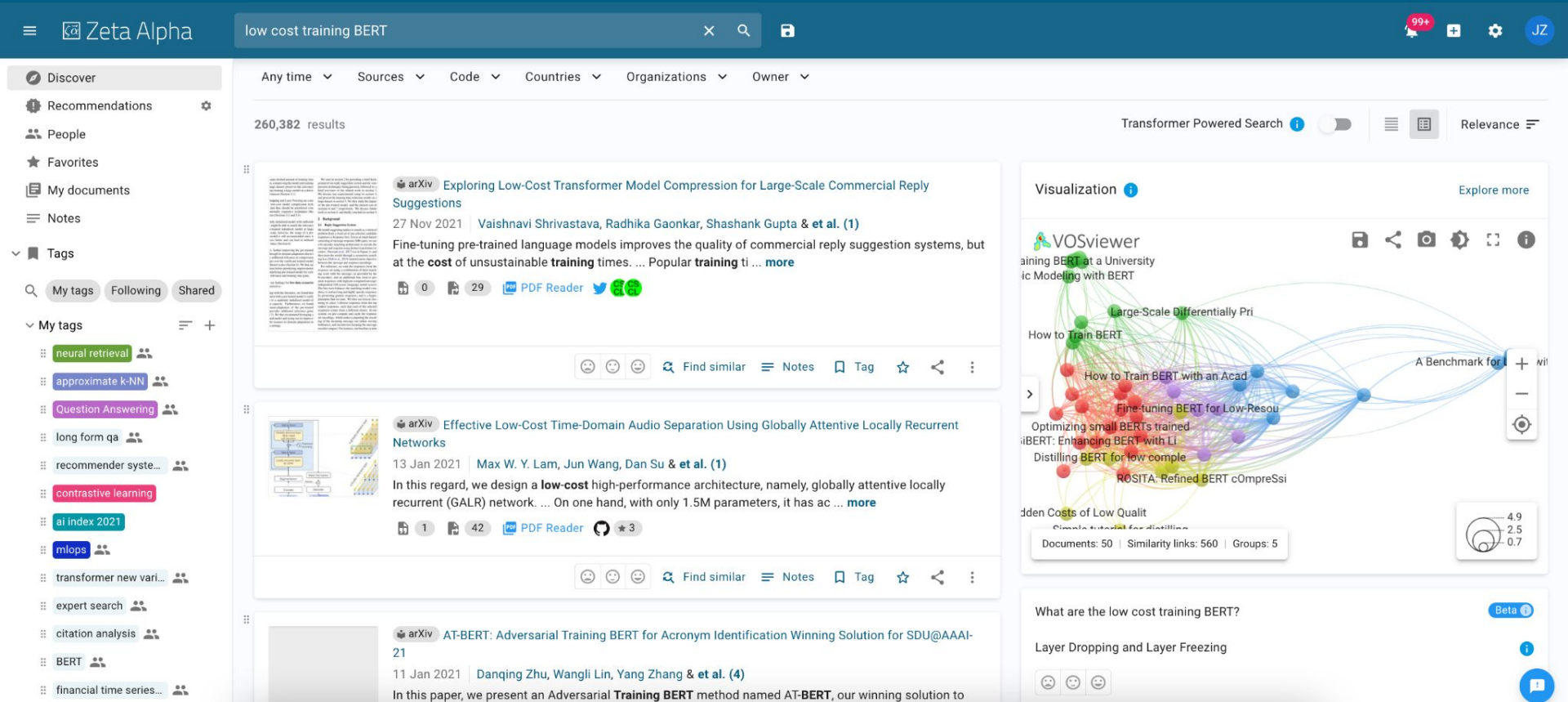
BM25 + boosting:

- Citation score
- Recency
- Author influence
- Popularity (Twitter)
- Code (GitHub stars)

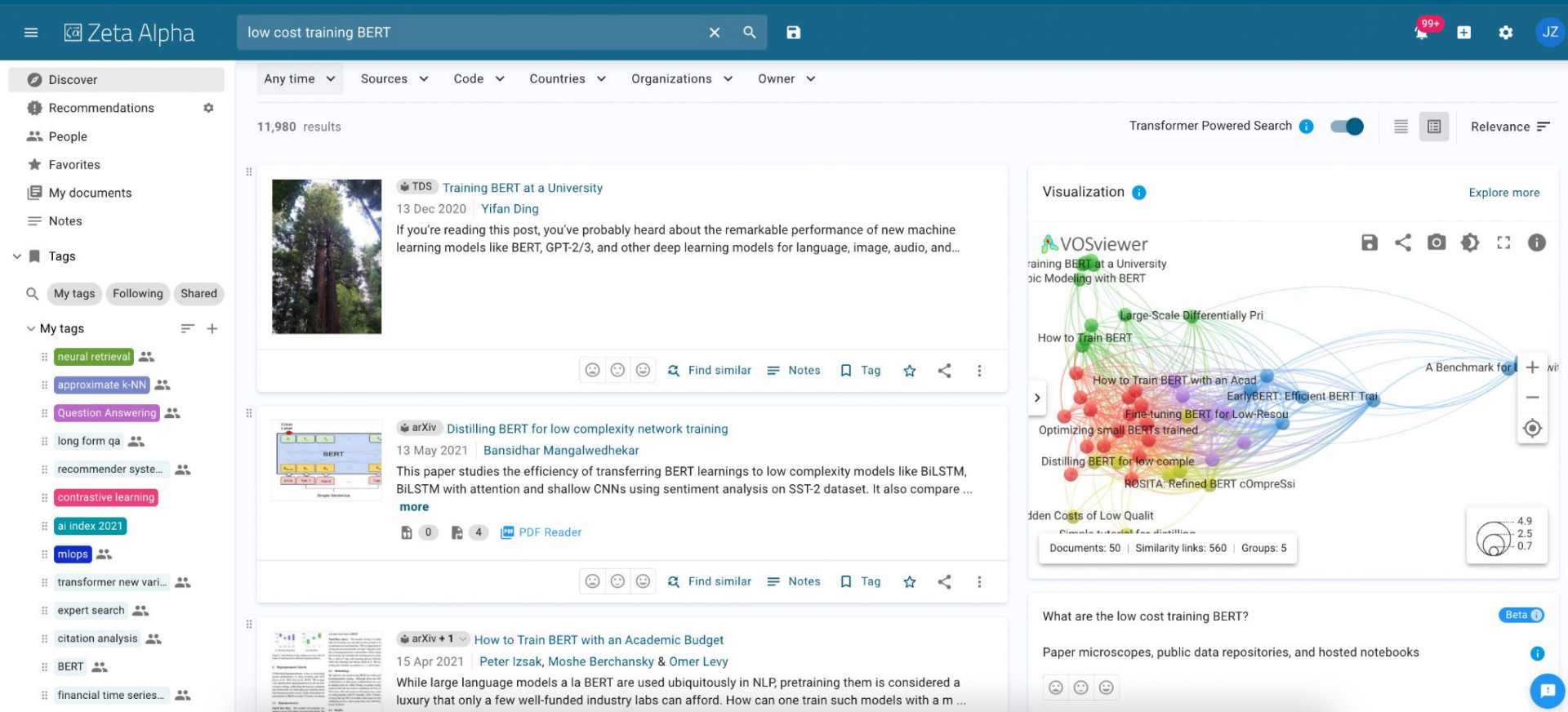
Dense Neural Retrieval, it's different...



BM25 struggles on more complex queries



Dense retrieval “gets it”



Zeta Alpha domain specific evaluation

We benchmark on ~100 queries in AI domain from 5 types: *short phrases, knowledge graph questions, quora questions, freq. user queries, and paper titles.*

Model	P@10	R@10	F1@10	MRR@10
ZA keyword retrieval	0.71	0.17	0.27	0.91
ZA dense retrieval	0.84	0.21	0.34	0.94

Challenges dense retrieval:

- **Calibration** (always gets an answer)
- **Explainability**
- **q2doc vs doc2doc** seem to require different embeddings.

Combining dense and keyword search

User does not ask for two separate retrieval mechanisms, but for relevant results.

Dense en keyword retrieval
top-10 overlap only ~ 15%.

How can we combine the rankings?

Sum:

$$score(q,d) = sim(q,d)/sim_{max}(q,d) + BM25(q,d)/BM25_{max}(q,d)$$

Mul:

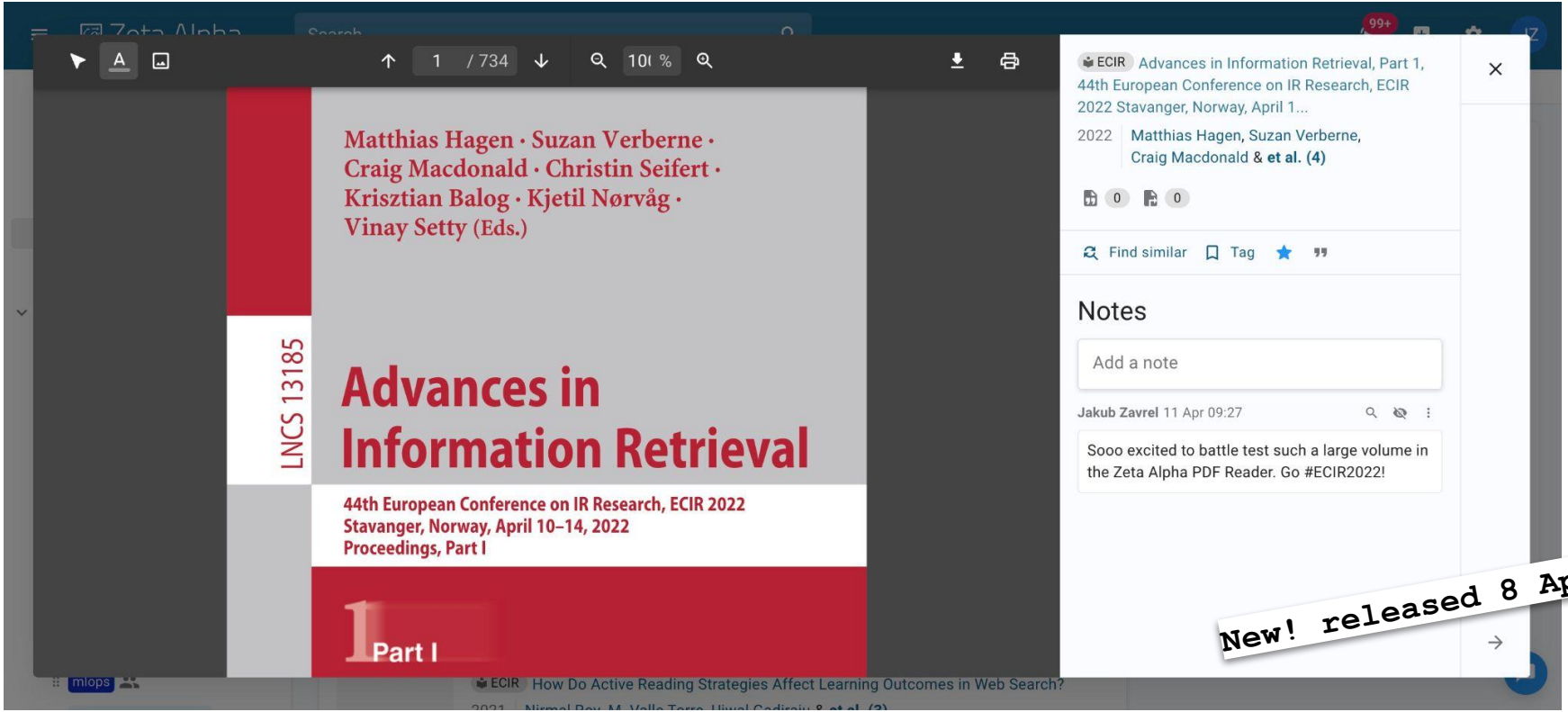
$$score(q,d) = \text{sim}(q,d) \times \text{BM25}(q,d)$$

RRF: (k=60)

$$RRF(q,d,M) = \sum_{m \in M} \frac{1}{k + \pi^m(q,d)}$$

	P@10	MRR@10	nDCG@10
Keyword	0.806	0.930	0.858
Dense	0.754	0.873	0.863
Sum	0.830	0.947	**0.883**
Mul	**0.846**	0.940	0.881
RRF	0.833	**0.957**	0.877

Just released: PDF Reader



Just released: PDF Reader + margin notes

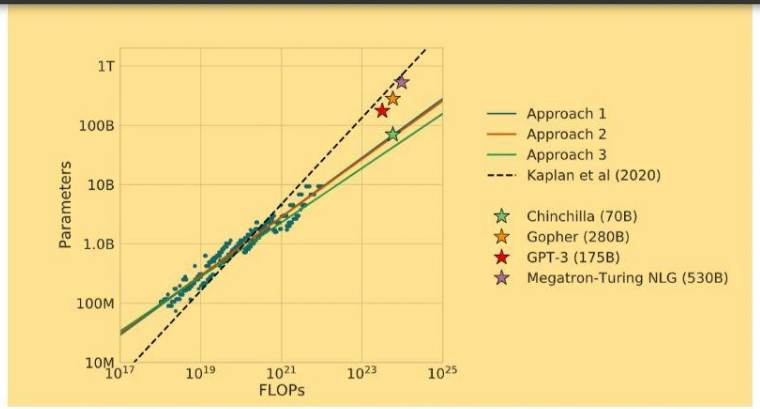
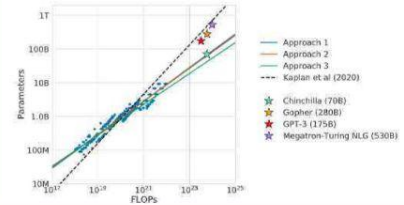


Figure 1 | **Overlaid predictions.** We overlay the predictions from our three different approaches, along with projections from Kaplan et al. (2020). We find that all three methods predict that current large models should be substantially smaller and therefore trained much longer than is currently done. In Figure A3, we show the results with the predicted optimal tokens plotted against the optimal number of parameters for fixed FLOP budgets. ***Chinchilla* outperforms *Gopher* and the other large models (see Section 4.2).**

In this work, we revisit the question: *Given a fixed FLOPs budget,*[1] *how should one trade-off model size and the number of training tokens?* To answer this question, we model the final pre-training loss[2] $L(N, D)$ as a function of the number of model parameters N, and the number of training tokens, D. Since the computational budget C is a deterministic function FLOPs(N, D) of the number of seen training tokens and model parameters, we are interested in minimizing L under the constraint

Jakub Zavrel 07 Apr 12:52



We have definitely entered the exaFLOPs era of Machine Learning.

Jakub Zavrel 07 Apr 12:49

Chinchilla outperforms Gopher and the other largemodels

In fact it also outperforms GPT-3

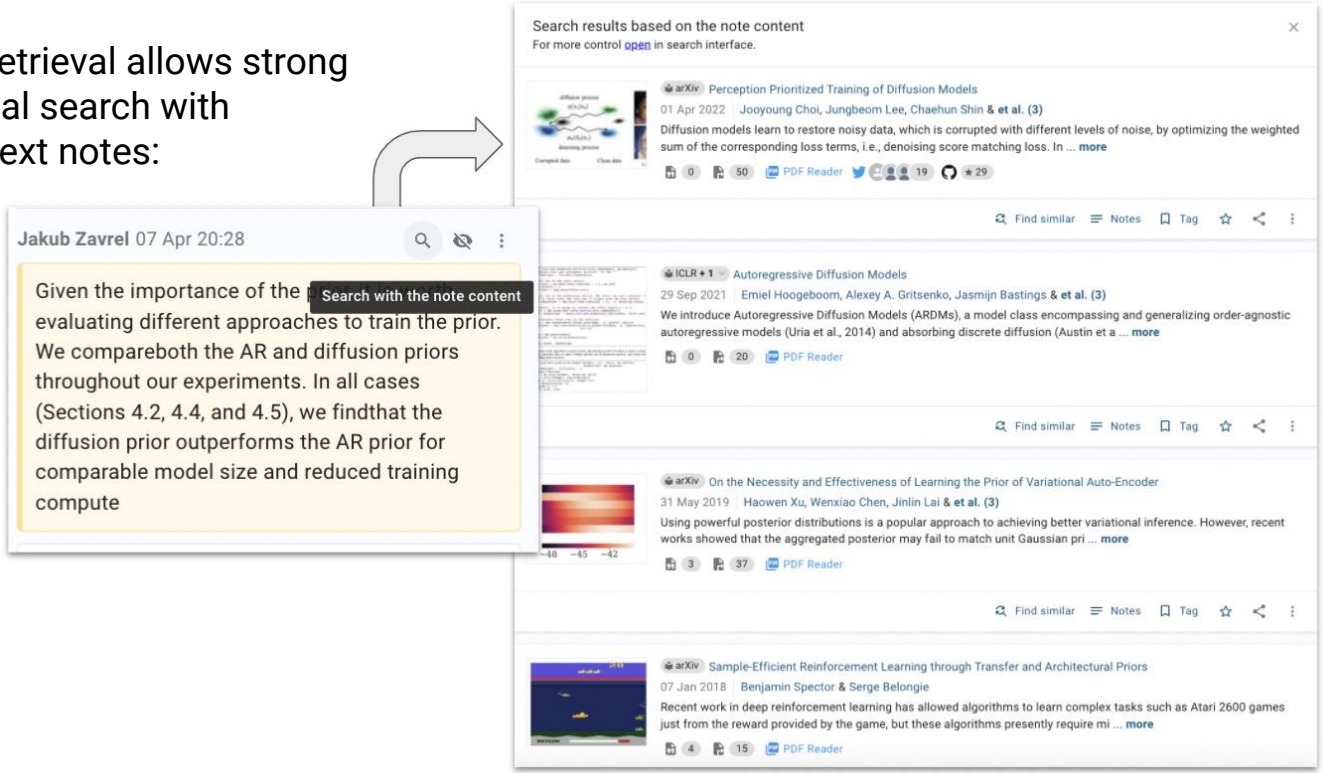
Jakub Zavrel 07 Apr 12:52

$$N_{opt}(C), D_{opt}(C) = \underset{N,D \text{ s.t. } \text{FLOPs}(N,D)=C}{\operatorname{argmin}} L(N, D).$$

?? really ??

Just released: PDF Reader + search in and with notes

Neural Retrieval allows strong contextual search with longish text notes:



How about using a neural reranker?

Add monoT5 based cross-encoder (Pradeep, Nogueira and Lin, 2021) on top of BM25

First stage	**Second stage**	*Prec@10*	*Recall@10*	*F1@10*	*MRR@10*
keyword	-	0.69	0.15	0.24	0.87
keyword	rerank 30	0.71	0.15	0.25	0.91
keyword	rerank 100	0.53	0.15	0.19	0.88

Marginal improvements or even degradation of results…

Outliers (i.e., noisy docs) can fool the reranker

+rerank 30

Alpha Go

Transformer Powered Search

Found 5,423 documents

arXiv From Gameplay to Symbolic Reasoning: Learning SAT Solver Heuristics in the Style of Alpha(Go) Zero

14 February 2018 | Fei Wang & Tiark Rompf

In this paper, we present our approach of casting symbolic reasoning as games, and directly harnessing the power of deep reinforcement learning in the style of **Alpha**(**Go**) Zero on symb ... **more**

2 11

Find similar Notes Tag

TDS Understanding AlphaGo: how AI thinks and learns (Advanced)

6 February 2020 | Shen Huang

This article will start with Deep Learning and **go** into the architecture of DeepMind AI, like **Alpha Go**. ... **more**

Towards Data Science

Find similar Notes Tag

GitHub CchessGo: Alpha Go for Chinese Chess

2 October 2018 | mengyangbai

Alpha Go for Chinese Chess ... **more**

Alpha Go

Transformer Powered Search

Found 5,423 documents

GitHub CchessGo: Alpha Go for Chinese Chess

2 October 2018 | mengyangbai

Alpha Go for Chinese Chess ... **more**

Find similar Notes Tag

GitHub sai

7 January 2019 | cjohnchen

Find similar Notes Tag

GitHub leela-zero

19 September 2018 | qinwang

Find similar Notes Tag

We lose our boosting effects

+rerank 30

bert

Transformer Powered Search

Code

Found 30,992 documents

NAACL BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

2 June 2019 | Jacob Devlin, Ming-Wei Chang, Kenton Lee & **et al. (1)**

We introduce a new language representation model called **BERT**, which stands for Bidirectional Encoder Representations from Transformers. ... **BERT** is conceptually simple and empiricall ... **more**

11.1k 4

Cited by 11068

Find similar Notes Tag

arXiv RoBERTa: A Robustly Optimized BERT Pretraining Approach

26 July 2019 | Yinhan Liu, Myle Ott, Naman Goyal & **et al. (7)**

We present a replication study of **BERT** pretraining (Devlin et al., 2019) that carefully measures the impact of many key hyperparameters and training data size. ... We find that **BERT** ... **more**

2.5k 43 6 1.9k

Find similar Notes Tag

bert

Transformer Powered Search

Code

Found 30,992 documents

AAAI + 1 K-BERT: Enabling Language Representation with Knowledge Graph.

5 May 2020 | Weijie Liu 0002, Peng Zhou, Zhe Zhao 0006 & **et al. (4)**

K-**BERT** can easily inject domain knowledge into the models by being equipped with a KG without pre-training by itself because it is capable of loading model parameters from the pre-trained **more**

100 0

Find similar Notes Tag

AAAI + 1 Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT.

5 May 2020 | Sheng Shen, Zhen Dong, Jiayu Ye & **et al. (5)**

However, **BERT** based models have a prohibitive memory footprint and latency. ... Among all tasks, we observed the highest performance loss for **BERT** fine-tuned on SQuAD. ... **more**

121 0

Find similar Notes Tag

Dense Retrieval + Reranker

Add monoT5 based cross-encoder (Pradeep, Nogueira and Lin, 2021) on top of ZA Dense (knn)

First stage	**Second stage**	*Prec@10*	*Recall@10*	*F1@10*	*MRR@10*
knn	-	0.83	0.19	0.31	0.93
knn	rerank 30	0.82	0.19	0.31	0.96
knn	rerank 50	0.82	0.19	0.31	0.97
knn	rerank 100	0.85	0.19	0.31	0.97

Marginal improvements, but at what cost?

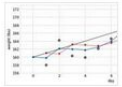
Better top results

+rerank 30

Alpha Go

Transformer Powered Search

Found 2,296 documents

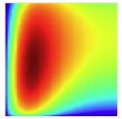


TDS Alpha beta filter

11 September 2020 | Nabil MADALI

The g and h coefficients refer to the two scaling factors, where g is the scaling we used for the measurement, and h is the scaling for the change in measurement over time . Measurement is typically...

Find similar Notes Tag



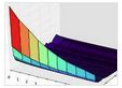
arXiv Bayesian Optimization in AlphaGo

17 December 2018 | Yutian Chen, Aja Huang, Ziyu Wang & **et al. (4)**

During the development of AlphaGo, its many hyper-parameters were tuned with Bayesian optimization multiple times. This automatic tuning process resulted in substantial improvements in playing stren ... **more**

17 9

Find similar Notes Tag



arXiv Hyper-Parameter Sweep on AlphaZero General

19 March 2019 | Hui Wang, Michael Emmerich, Mike Preuss & **et al. (1)**

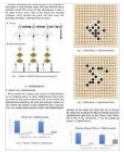
Since AlphaGo and AlphaGo Zero have achieved breakground successes in the game of Go, the programs have been generalized to solve other tasks. Subsequently, AlphaZero was developed to play Go, Chess ... **more**

4 27 ★154

Alpha Go

Transformer Powered Search

Found 2,073 documents



arXiv AlphaGomoku: An AlphaGo-based Gomoku Artificial Intelligence using Curriculum Learning

27 September 2018 | Zheng Xie, XingYu Fu & JinYuan Yu

In this project, we combine AlphaGo algorithm with Curriculum Learning to crack the game of Gomoku. Modifications like Double Networks Mechanism and Winning Value Decay are implemented to solve the ... **more**

1 13 ★48

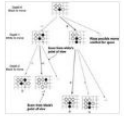
Find similar Notes Tag

TDS Understanding AlphaGo, how AI think and learn (Fundamentals)

3 February 2020 | Shen Huang

Join me in an adventure of learning how game AI works...

Find similar Notes Tag



TDS Game Theory Concepts Within AlphaGo

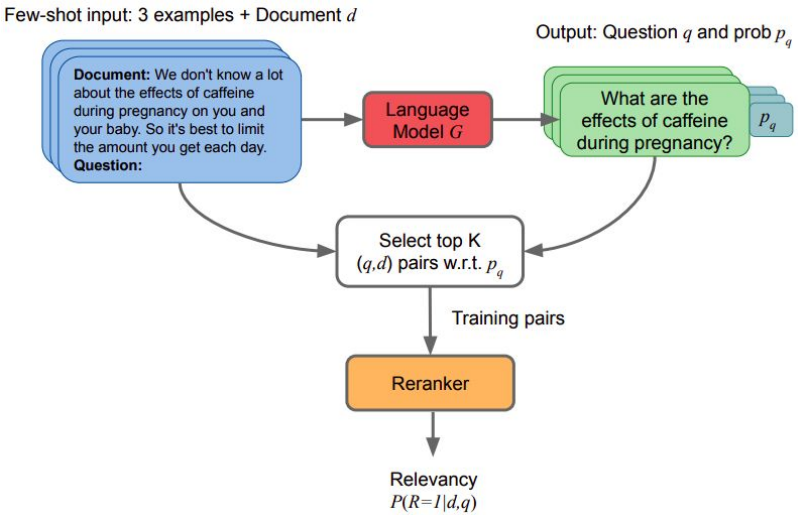
10 May 2020 | Gabe Stechschulte

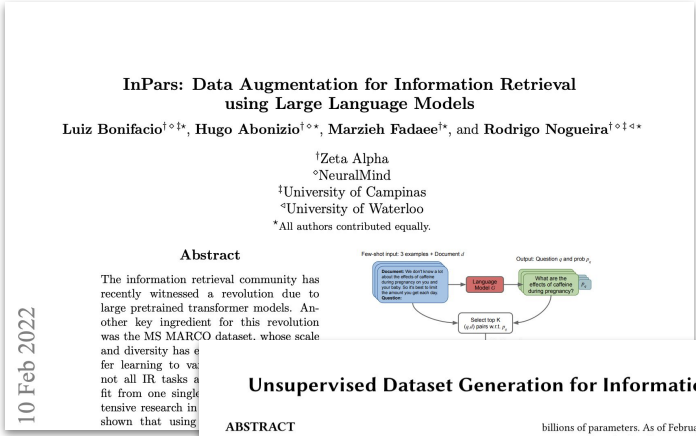
Go, arguably the world's oldest board game and originated in East Asia some 4,000 years ago, is a game played amongst two players. Go is often played with 181 black and 180 white stones on a square...

Cross encoder reranking: Ouch, Performance!

top N	api call	Direct model call
10	0.9s	3.0s
30	1.9s	3.1s
50	4.2s	3.0s
100	4.7s	3.1s
1000	44.2s	3.4s

InPars: Data Augmentation for Unsupervised IR





10 Feb 2022

InPars: Data Augmentation for Information Retrieval using Large Language Models

Luiz Bonifacio[†◇‡*], Hugo Abonizio[‡◇*], Marzieh Fadaee[†*], and Rodrigo Nogueira[†◇‡◁*]

[†]Zeta Alpha
[◇]NeuralMind
[‡]University of Campinas
[◁]University of Waterloo
[*]All authors contributed equally.

Abstract

The information retrieval community has recently witnessed a revolution due to large pretrained transformer models. Another key ingredient for this revolution was the MS MARCO dataset, whose scale and diversity has e
fer learning to va
not all IR tasks a
fit from one singl
tensive research in
shown that using

Unsupervised Dataset Generation for Information Retrieval

ABSTRACT

The Information Retrieval (IR) community has recently witnessed a revolution due to large pretrained transformer models. Another key ingredient for this revolution was the MS MARCO dataset, whose scale and diversity has enabled zero-shot transfer learning to various tasks. However, not all IR tasks and domains can benefit from one single dataset equally. Extensive research in various NLP tasks has shown that using domain-specific training data, as opposed to a general-purpose one, improves the performance of neural models [43, 54]. In this work, we harness the few-shot capabilities of large pretrained language models as synthetic data generators for IR tasks. We show that models finetuned solely on our unsupervised dataset outperform strong baselines such as BM25 as well as recently proposed self-supervised dense retrieval methods. Code, models, and data are available at <hidden_url>.

CCS CONCEPTS

• **Information systems** → **Novelty in information retrieval**; • **Computing methodologies** → **Neural networks**.

KEYWORDS

Few-shot Models, Large Language Models, Generative Models, Question Generation, Synthetic Datasets, Multi-stage Ranking

ACM Reference Format:

. 2018. Unsupervised Dataset Generation for Information Retrieval. In *Wood-*

billions of parameters. As of February 2022, they charge 0.06 USD per 1000 tokens for their largest model. If each candidate document contains 250 tokens, naively using this API for a reranking task would cost approximately 15 USD per query.

Dense retrievers [13, 14] avoid this expensive reranking step by precomputing vector representations for each document in the collection prior to retrieval. When a query comes in, only its vector representations are computed, and a fast vector search framework can be used to retrieve the nearest document vectors to the vector representation of the query [12]. Despite being computationally cheaper at inference time, dense retrievers need one inference pass to compute the vector representation *of each* document in the collection, which also makes billion-parameter neural models impracticable to be used as dense retrievers.[1] Another challenge in developing neural models for IR is the lack of domain-specific training data. Manually constructing high-quality datasets is difficult as it requires queries from real users. While there are a few general-purpose labeled data available [17, 28], they are not always effective in generalizing to out-of-domain datasets [26, 44]. For this goal, zero-shot and few-shot learning models are in particular promising. However, a cost-effective manner of using large LMs in IR tasks is still an open question.

In this work, we propose a simple yet effective approach towards efficiently using large LMs in retrieval and obtain improvements across several IR datasets. Rather than using large LMs directly in the retrieval process, we harness them to generate labeled data

Ranking models are **finetuned on a synthetic dataset** built by augmenting **documents with queries** using **generative LLMs like GPT-3.**

This represents our recipe for **unsupervised domain adaptation.**
With **very good results on the BEIR benchmark…**

InPars: Data Augmentation for IR using LLM's

		MARCO	TREC-DL 2020		Robust04		NQ	TRECC
		MRR@10	MAP	nDCG@10	MAP	nDCG@20	nDCG@10	nDCG@10
	Unsupervised							
(1)	BM25	0.1874	0.2876	0.4876	0.2531	0.4240	0.3290	0.6880
(2)	Contriever (Izacard et al., 2021)	-	-	-	-	-	0.2580	0.2740
(3)	cpt-text (Neelakantan et al., 2022)	0.2270	-	-	-	-	-	0.4270
	OpenAI Search reranking 100 docs from BM25							
(4)	Ada (300M)	$	0.3141	0.5161	0.2691	0.4847	0.4092	0.6757
(5)	Curie (6B)	$	0.3296	0.5422	0.2785	0.5053	0.4171	0.7251
(6)	Davinci (175B)	$	0.3163	0.5366	0.2790	0.5103	$	0.6918
	InPars (ours)							
(7)	monoT5-220M	0.2585	0.3599	0.5764	0.2490	0.4268	0.3354	0.6666
(8)	monoT5-3B	**0.2967**	**0.4334**	**0.6612**	**0.3180**	**0.5181**	**0.5133**	**0.7835**
	Supervised [▷ **MARCO**]							
(9)	Contriever (Izacard et al., 2021)	-	-	-	-	-	0.4980	0.5960
(10)	cpt-text (Neelakantan et al., 2022)	-	-	-	-	-	-	0.6490
(11)	ColBERT-v2 (Santhanam et al., 2021)	0.3970	-	-	-	-	0.5620	0.7380
(12)	GPL (Wang et al., 2021)	-	-	-	-	-	-	0.7400
(13)	miniLM reranker	†0.3901	-	-	-	-	‡0.5330	‡0.7570
(14)	monoT5-220M (Nogueira et al., 2020)	0.3810	0.4909	0.7141	0.3279	0.5298	0.5674	0.7775
(15)	monoT5-3B (Nogueira et al., 2020)	0.3980	**0.5281**	**0.7508**	0.3876	0.6091	**0.6334**	0.7948
	InPars (ours) [▷ MARCO ▷ unsup in-domain]							
(16)	monoT5-3B	0.3894	0.5087	0.7439	**0.3967**	**0.6227**	0.6297	**0.8471**

InPars: out of domain data augmentation

Example 1:
Document: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day. If you are pregnant, limit caffeine to 200 milligrams each day. This is about the amount in 1½ 8-ounce cups of coffee or one 12-ounce cup of coffee.
Relevant Query: Is a little caffeine ok during pregnancy?

Example 2:
Document: Passiflora herbertiana. A rare passion fruit native to Australia. Fruits are green-skinned, white fleshed, with an unknown edible rating. Some sources list the fruit as edible, sweet and tasty, while others list the fruits as being bitter and inedible.
Relevant Query: What fruit is native to Australia?

Example 3:
Document: The Canadian Armed Forces. 1 The first large-scale Canadian peacekeeping mission started in Egypt on November 24, 1956. 2 There are approximately 65,000 Regular Force and 25,000 reservist members in the Canadian military. 3 In Canada, August 9 is designated as National Peacekeepers' Day.
Relevant Query: How large is the Canadian military?

Example 4:
Document: {document_text}
Relevant Query:

Example 1:
Document: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day. If you are pregnant, limit caffeine to 200 milligrams each day. This is about the amount in 1½ 8-ounce cups of coffee or one 12-ounce cup of coffee.
Good Question: How much caffeine is ok for a pregnant woman to have?
Bad Question: Is a little caffeine ok during pregnancy?

Example 2:
Document: Passiflora herbertiana. A rare passion fruit native to Australia. Fruits are green-skinned, white fleshed, with an unknown edible rating. Some sources list the fruit as edible, sweet and tasty, while others list the fruits as being bitter and inedible.
Good Question: What is Passiflora herbertiana (a rare passion fruit) and how does it taste like?
Bad Question: What fruit is native to Australia?

Example 3:
Document: The Canadian Armed Forces. 1 The first large-scale Canadian peacekeeping mission started in Egypt on November 24, 1956. 2 There are approximately 65,000 Regular Force and 25,000 reservist members in the Canadian military. 3 In Canada, August 9 is designated as National Peacekeepers' Day.
Good Question: Information on the Canadian Armed Forces size and history.
Bad Question: How large is the Canadian military?

Example 4:
Document: {document_text}
Good Question:

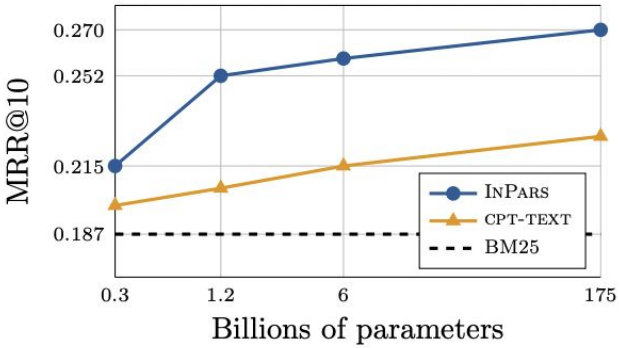
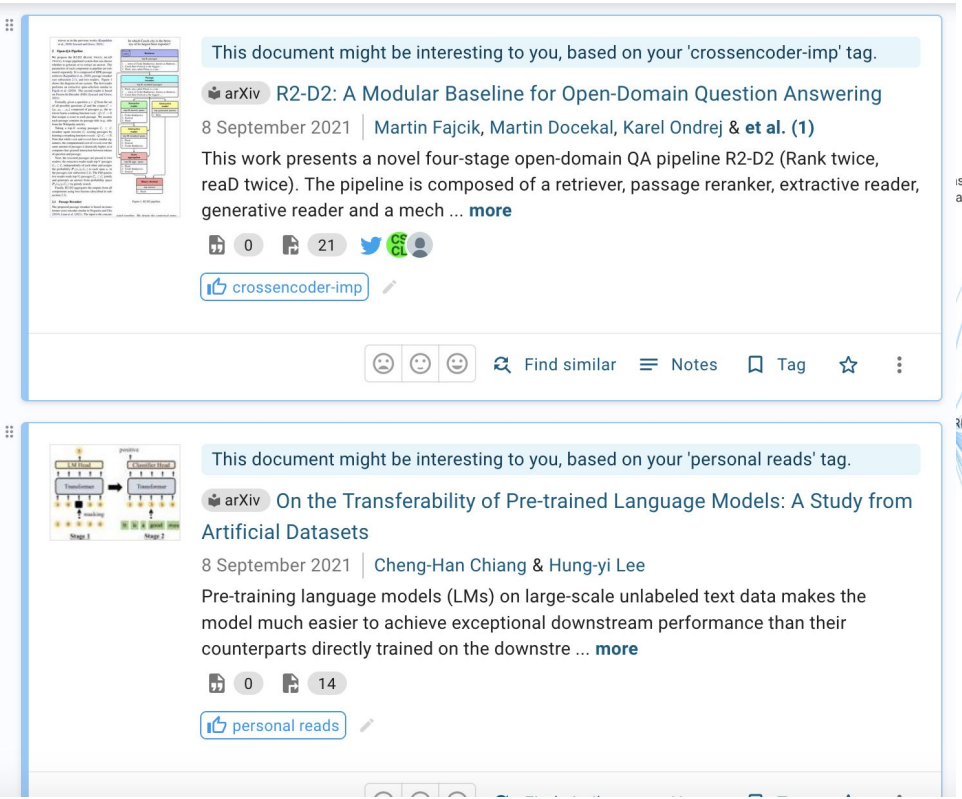
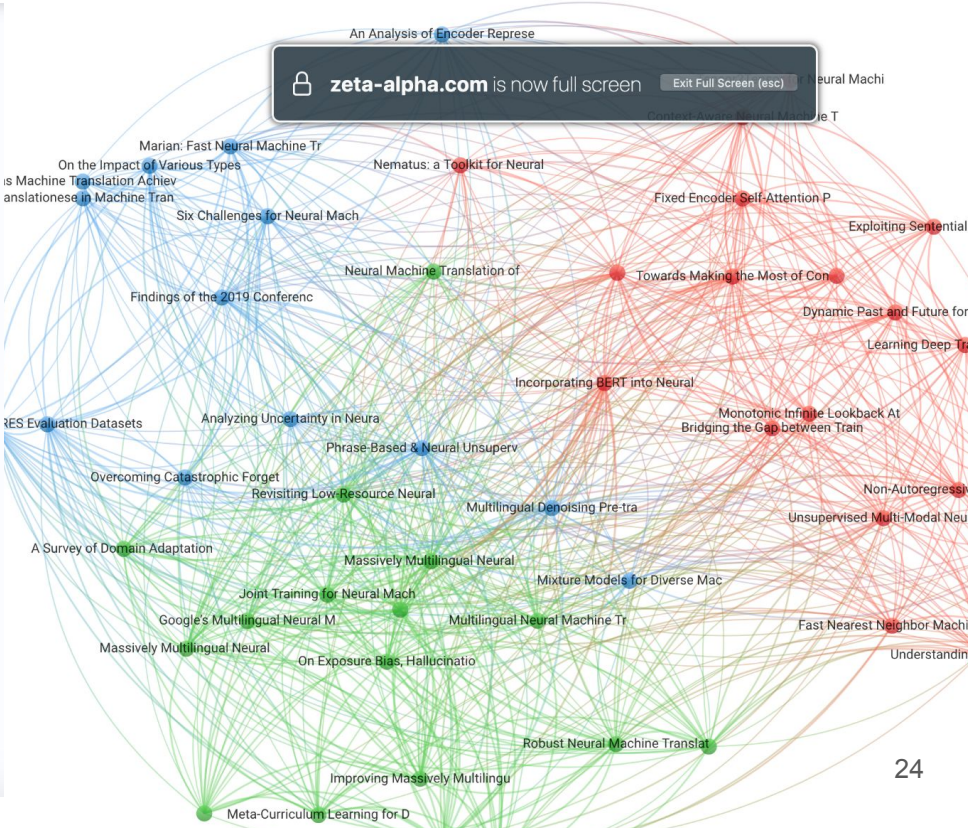


Figure 3: MRR@10 on the MS MARCO development set achieved by InPars using monoT5-220M reranker trained on synthetic questions generated by GPT-3 models of different sizes. Figures for cpt-text are from (Neelakantan et al., 2022). Note the log scale for the x-axis.

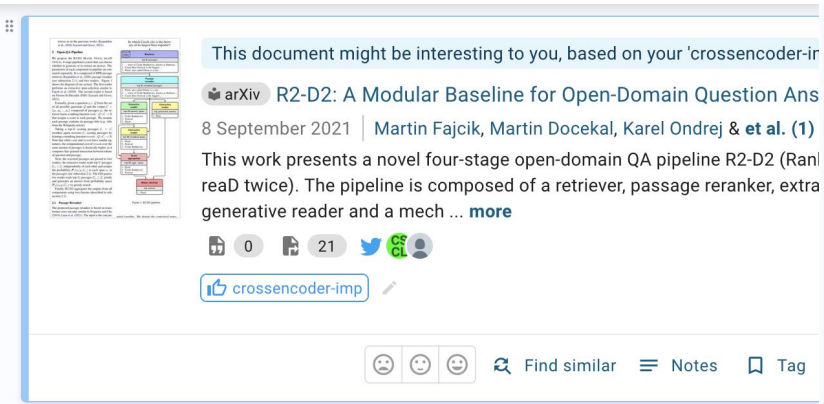
Many other uses of dense semantic embeddings

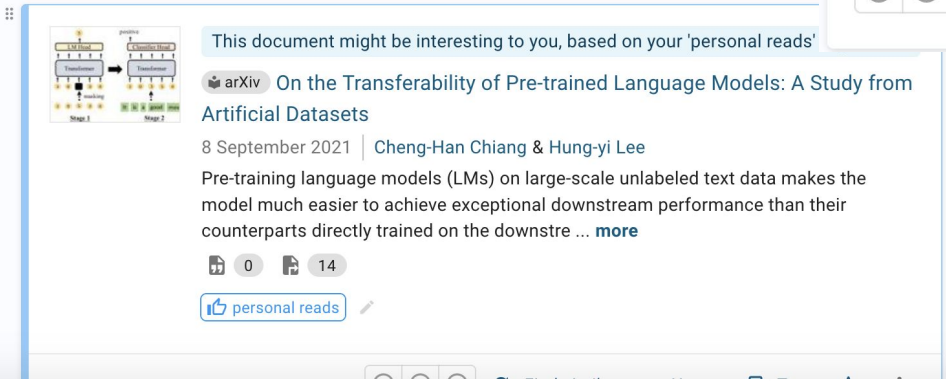
Many other uses of dense semantic embeddings

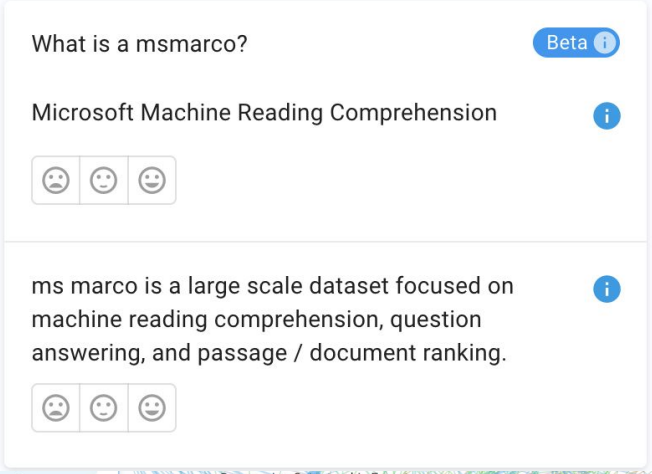




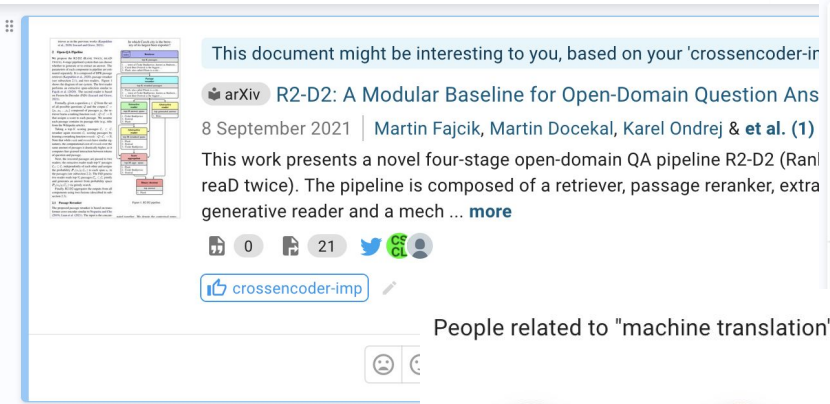
Many other uses of dense semantic embeddings

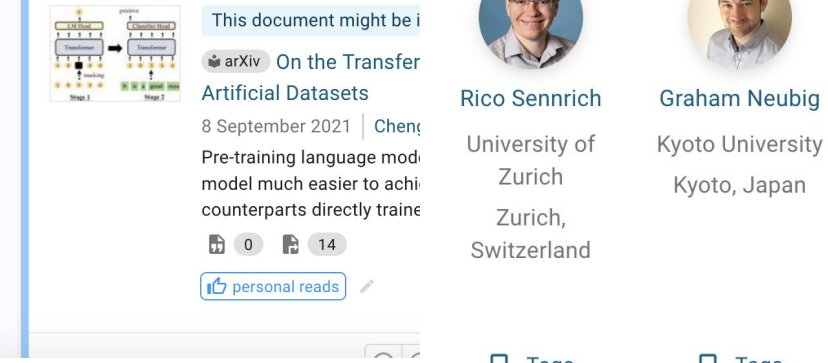


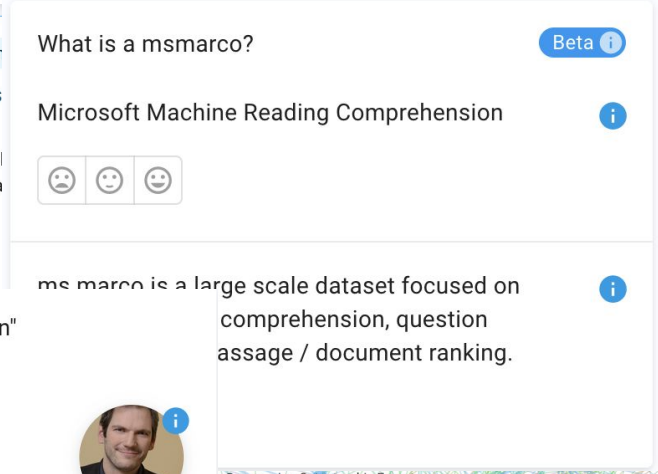


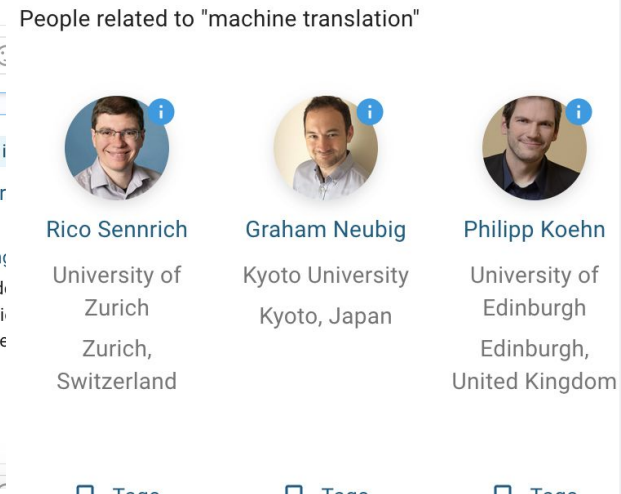


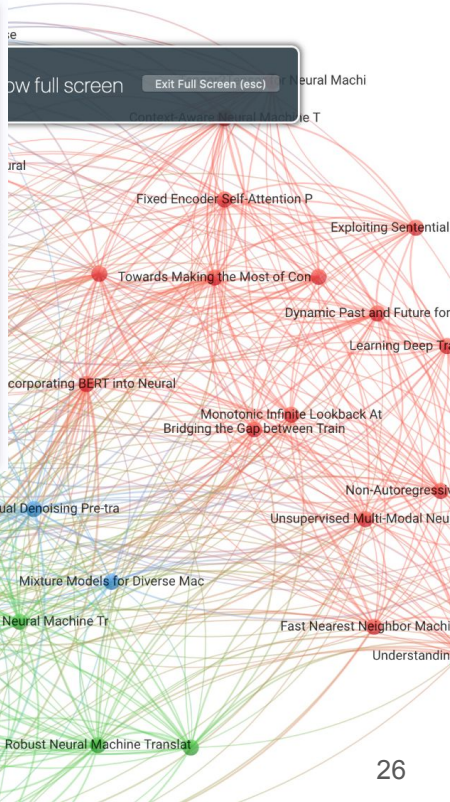
Many other uses of dense semantic embeddings

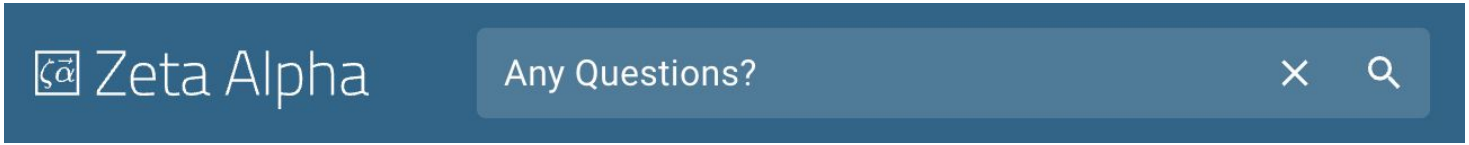






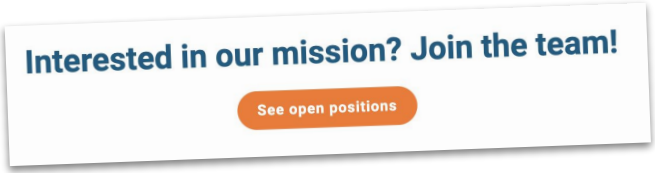






Publications

- **InPars: Data Augmentation for Information Retrieval using Large Language Models**
  2022 | Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee & Rodrigo Nogueira
  See paper
- **Building a Platform for Ensemble-based Personalized Research Literature Recommendations for AI and Data Science at Zeta Alpha**
  2021 | Jakub Zavrel, Artem Grotov, & Jonathan Mitnik
  See paper
- **mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset**
  2021 | Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo & Rodrigo Nogueira
  See paper
- **Pretrained Transformers for Text Ranking: BERT and Beyond**
  2021 | Jimmy Lin, Rodrigo Nogueira, & Andrew Yates
  Access book
- **A New Neural Search and Insights Platform for Navigating and Organizing AI Research**
  2020 | Marzieh Fadaee, Olga Gureenkova, Fernando Rejon Barrera, Carsten Schnober, Wouter Weerkamp, Jakub Zavrel
  See paper
- **Effective Distributed Representations for Academic Expert Search**
  2020 | Mark Berger, Jakub Zavrel, & Paul Groth
  See paper



Get more information, sign up to use the platform:

www.zeta-alpha.com