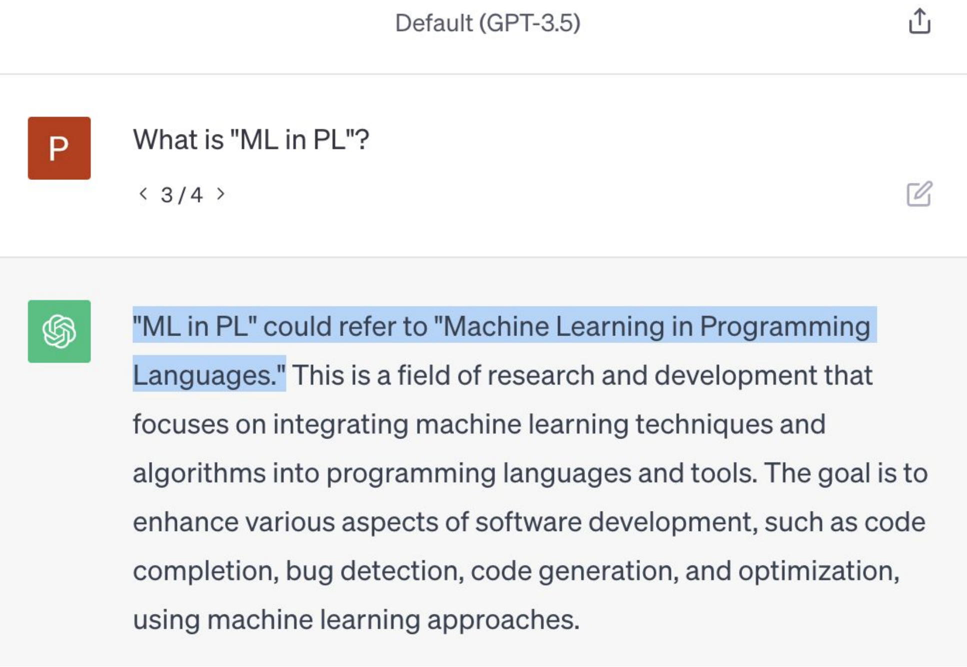
The State of Polish Semantic Search in the Era of LLMs

Piotr Rybak

Institute of Computer Science

Polish Academy of Sciences

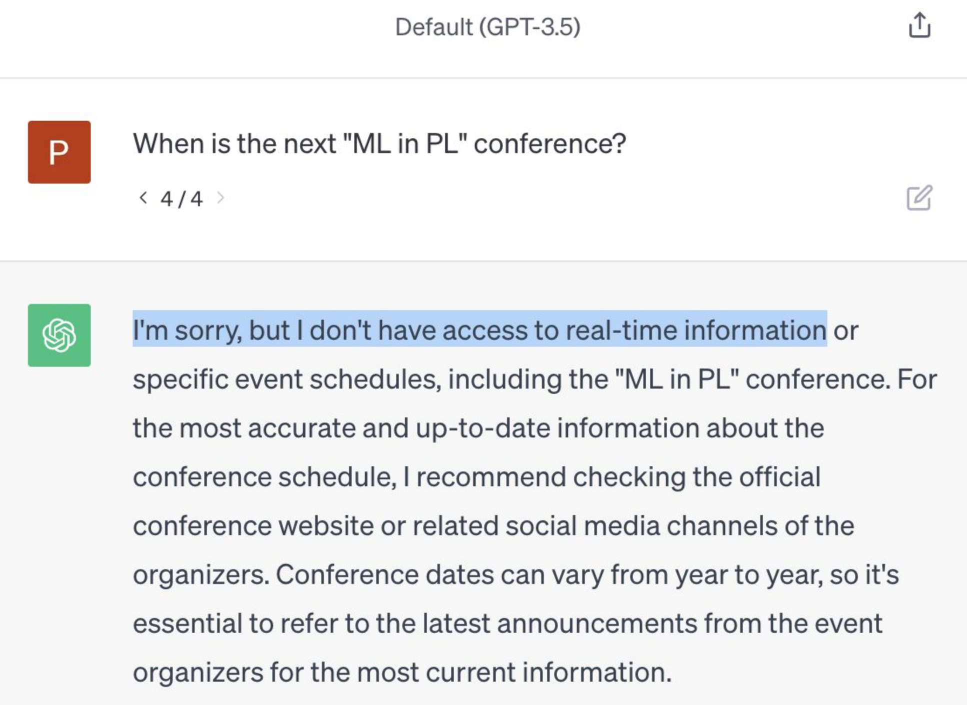
LLMs are great, right?



What is "ML in PL"?

"ML in PL" could refer to "Machine Learning in Programming Languages." This is a field of research and development that focuses on integrating machine learning techniques and algorithms into programming languages and tools. The goal is to enhance various aspects of software development, such as code completion, bug detection, code generation, and optimization, using machine learning approaches.

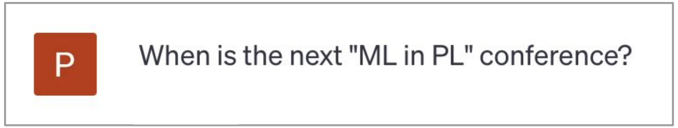
LLMs are great, right?



When is the next "ML in PL" conference?

I'm sorry, but I don't have access to real-time information or specific event schedules, including the "ML in PL" conference. For the most accurate and up-to-date information about the conference schedule, I recommend checking the official conference website or related social media channels of the organizers. Conference dates can vary from year to year, so it's essential to refer to the latest announcements from the event organizers for the most current information.

Retrieval-Augmented Generation

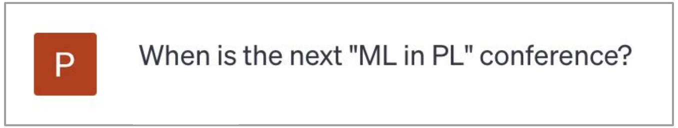


When is the next "ML in PL" conference?

↓

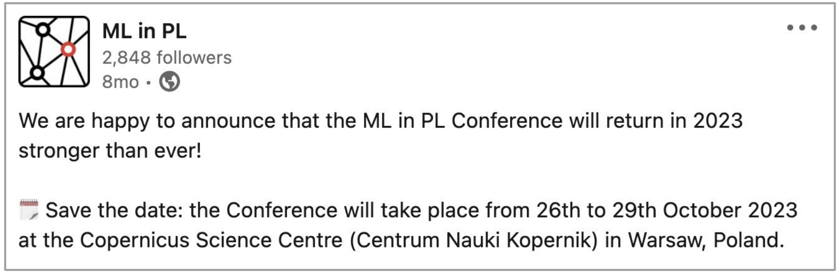
Google

Retrieval-Augmented Generation



When is the next "ML in PL" conference?

Google



ML in PL

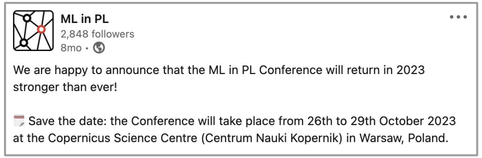
2,848 followers

8mo •

We are happy to announce that the ML in PL Conference will return in 2023 stronger than ever!

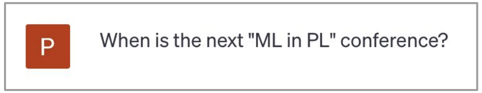
🗓️ Save the date: the Conference will take place from 26th to 29th October 2023 at the Copernicus Science Centre (Centrum Nauki Kopernik) in Warsaw, Poland.

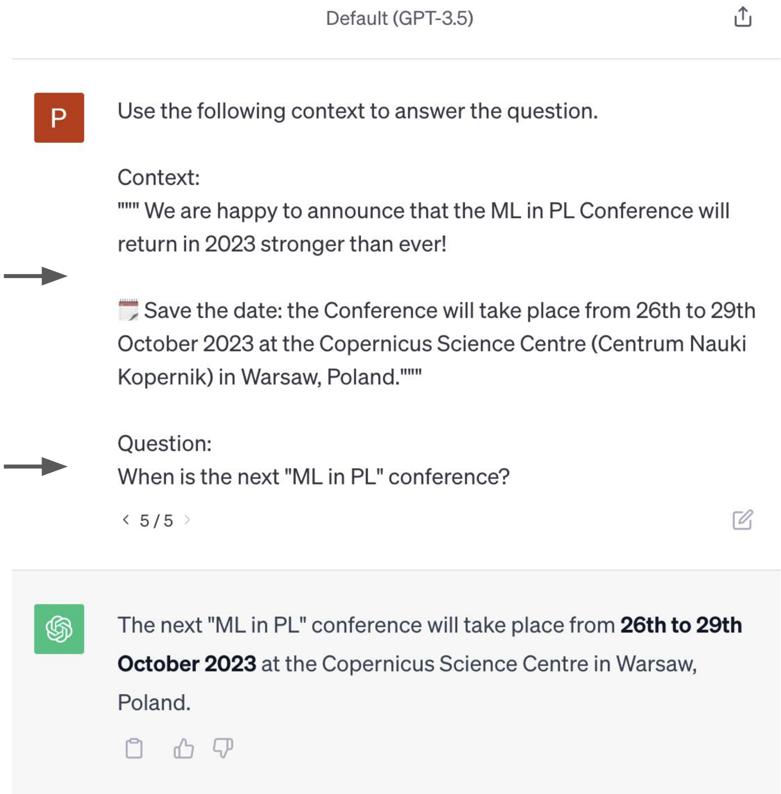
Retrieval-Augmented Generation


ML in PL
2,848 followers
8mo ·

We are happy to announce that the ML in PL Conference will return in 2023 stronger than ever!

🗓️ Save the date: the Conference will take place from 26th to 29th October 2023 at the Copernicus Science Centre (Centrum Nauki Kopernik) in Warsaw, Poland.


When is the next "ML in PL" conference?



Use the following context to answer the question.

Context:
""" We are happy to announce that the ML in PL Conference will return in 2023 stronger than ever!

🗓️ Save the date: the Conference will take place from 26th to 29th October 2023 at the Copernicus Science Centre (Centrum Nauki Kopernik) in Warsaw, Poland."""

Question:
When is the next "ML in PL" conference?

The next "ML in PL" conference will take place from **26th to 29th October 2023** at the Copernicus Science Centre in Warsaw, Poland.

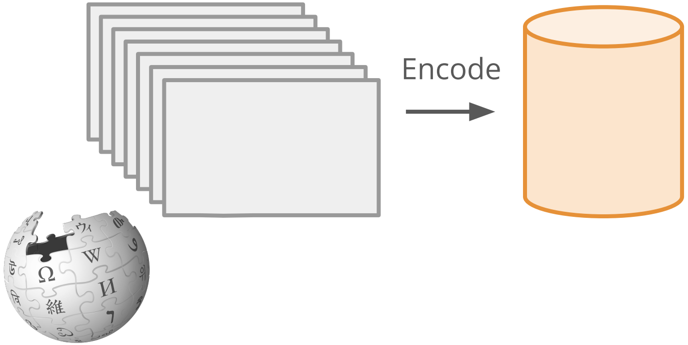
Internal documents

P When is the next medical appointment for Piotr Rybak?

P What is the revenue of the Warsaw office?

P How to deploy service ABC to the dev environment?

Search architecture

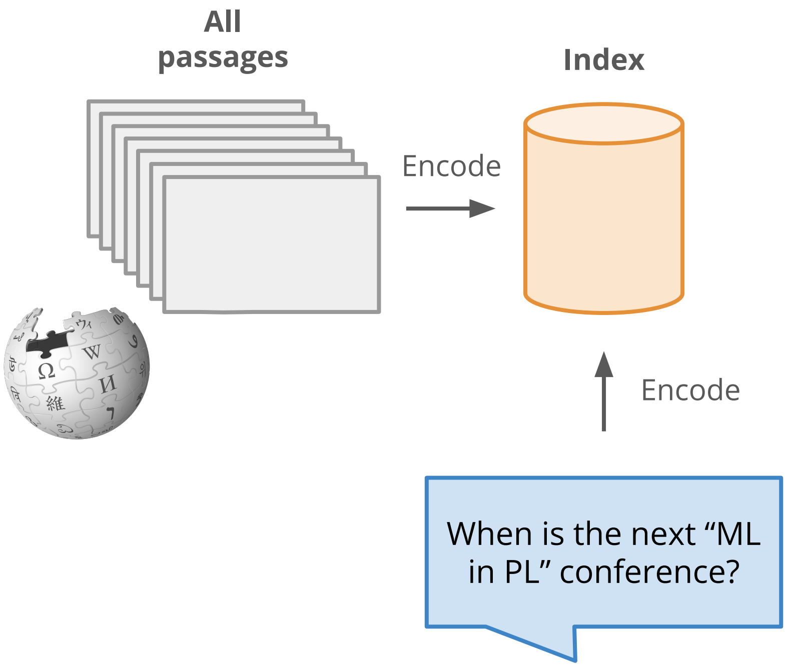


All passages

Encode

Index

Search architecture



All passages

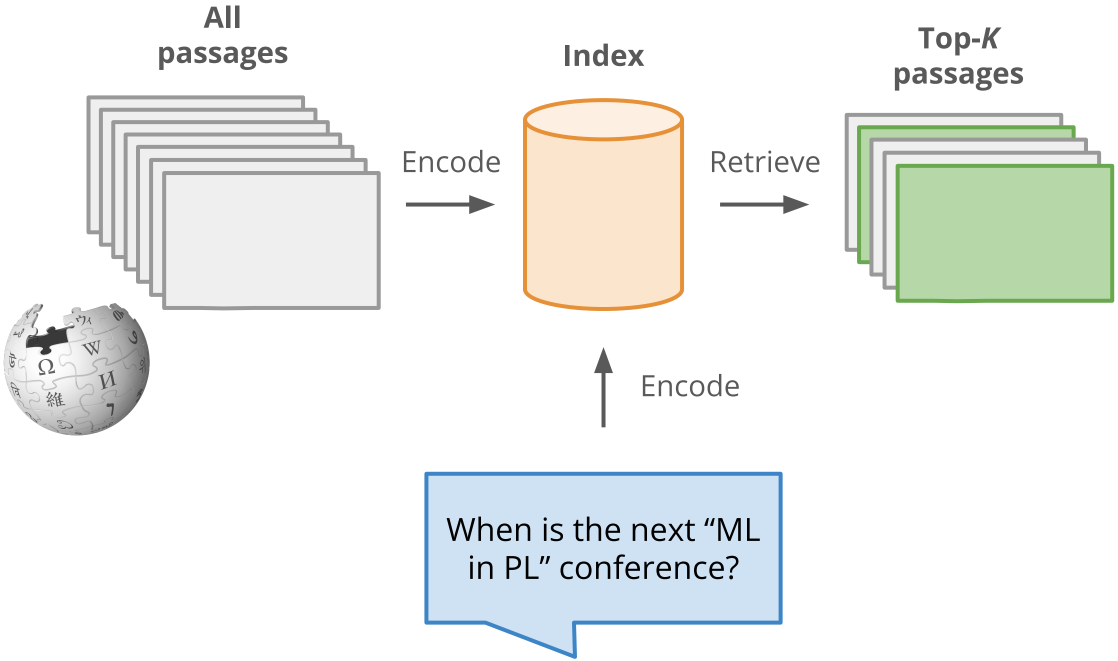
Encode

Index

Encode

When is the next “ML in PL” conference?

Search architecture



All passages

Encode

Index

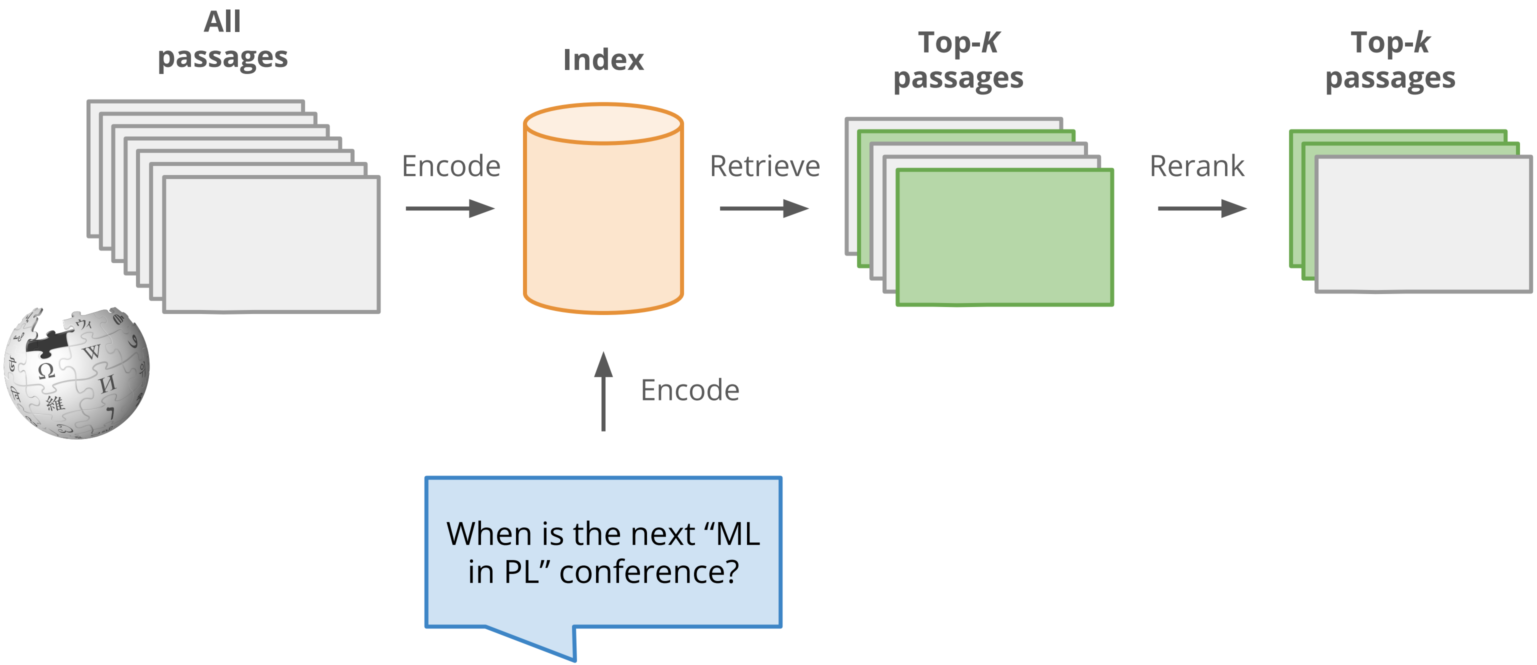
Retrieve

Top-*K* passages

Encode

When is the next “ML in PL” conference?

Search architecture



All passages

Encode

Index

Retrieve

Top-*K* passages

Rerank

Top-*k* passages

Encode

When is the next “ML in PL” conference?

Lexical vs Semantic

When is the **next** “**ML in PL**” **conference**?

We are happy to announce that the **ML in PL Conference** will return in 2023 stronger than ever!

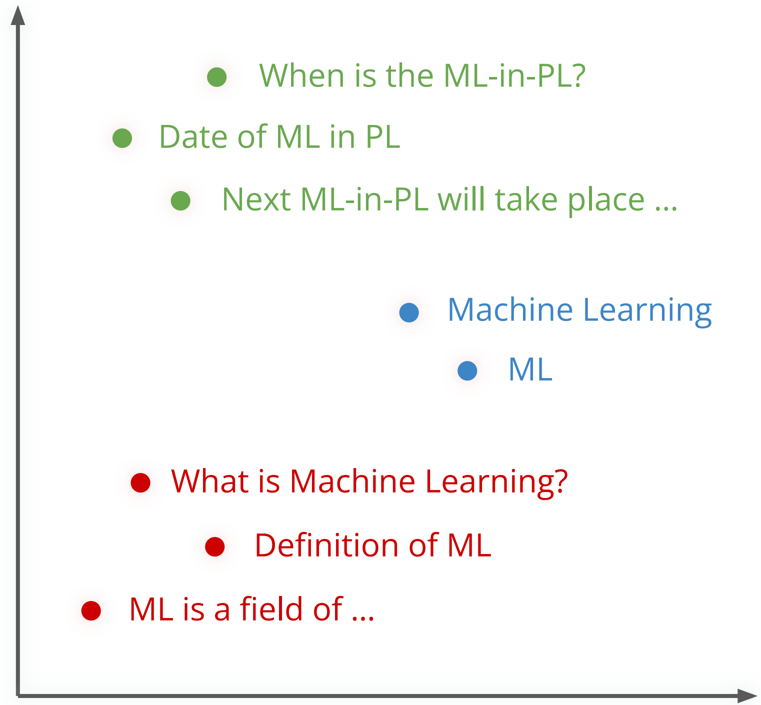
📅 Save the date: the **next** **Conference** will take place from 26th to 29th October 2023 at the Copernicus Science Centre (Centrum Nauki Kopernik) in Warsaw Poland.

Lexical vs Semantic

When is the **next** “**ML in PL**” **conference**?

We are happy to announce that the **ML in PL Conference** will return in 2023 stronger than ever!

📅 Save the date: the **next Conference** will take place from 26th to 29th October 2023 at the Copernicus Science Centre (Centrum Nauki Kopernik) in Warsaw Poland.



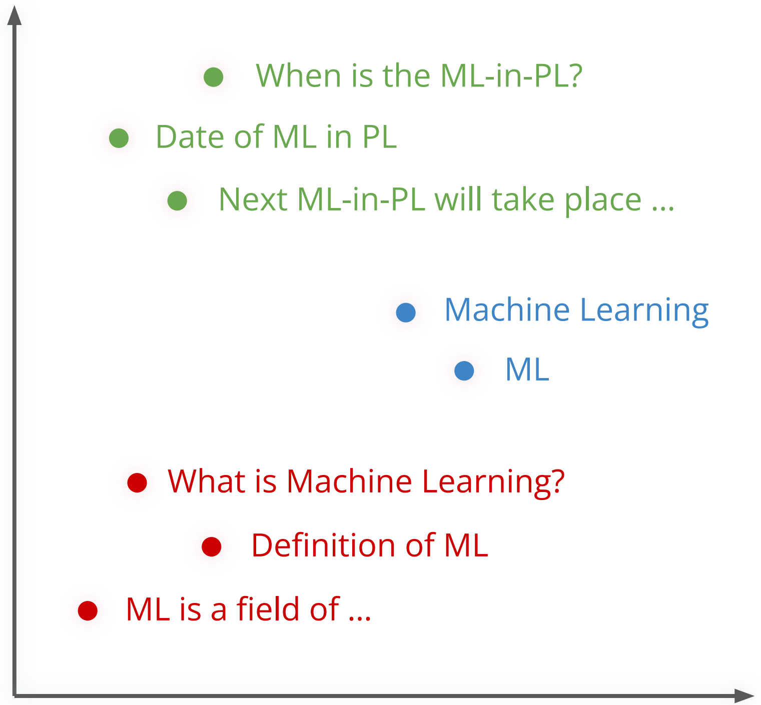
- When is the ML-in-PL?
- Date of ML in PL
- Next ML-in-PL will take place …
- Machine Learning
- ML
- What is Machine Learning?
- Definition of ML
- ML is a field of …

Lexical vs Semantic

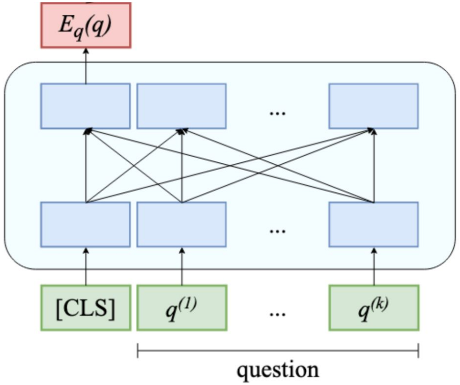
When is the **next** “**ML in PL**” **conference**?

We are happy to announce that the **ML in PL Conference** will return in 2023 stronger than ever!

📆 Save the date: the **next Conference** will take place from 26th to 29th October 2023 at the Copernicus Science Centre (Centrum Nauki Kopernik) in Warsaw Poland.



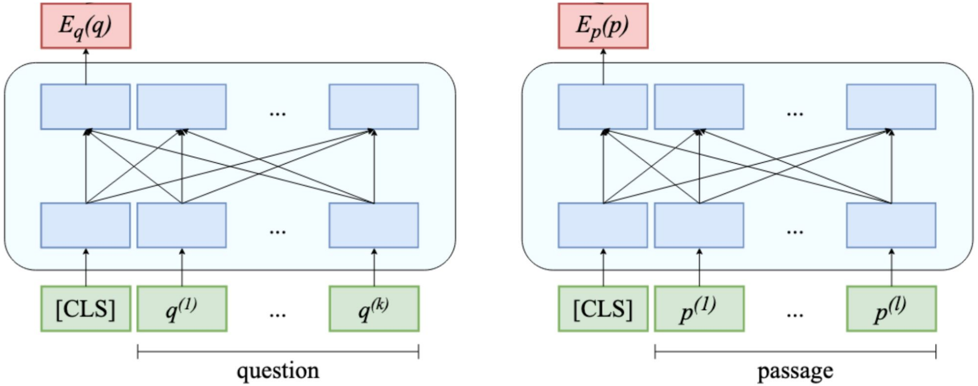
Dense Passage Retrieval



Who is the bad guy in
lord of the rings?

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering

Dense Passage Retrieval

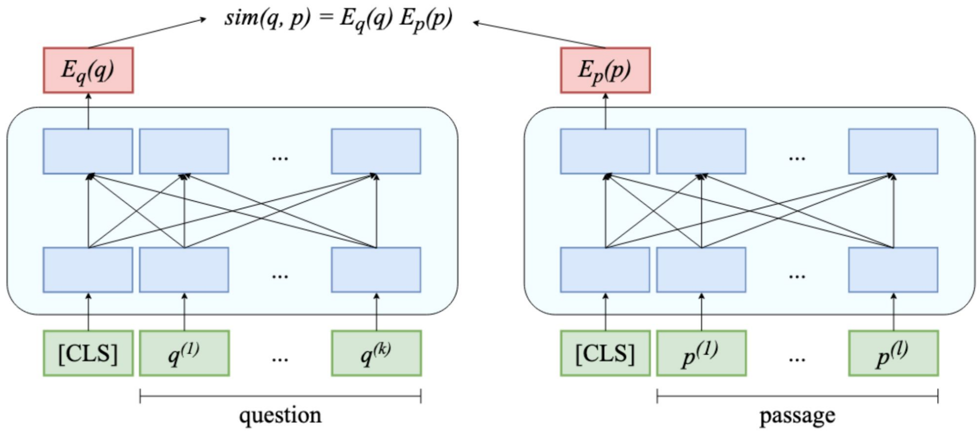


Who is the bad guy in lord of the rings?

Sala Baker is best known for portraying the villain Sauron in the Lord of the Rings trilogy

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering

Dense Passage Retrieval

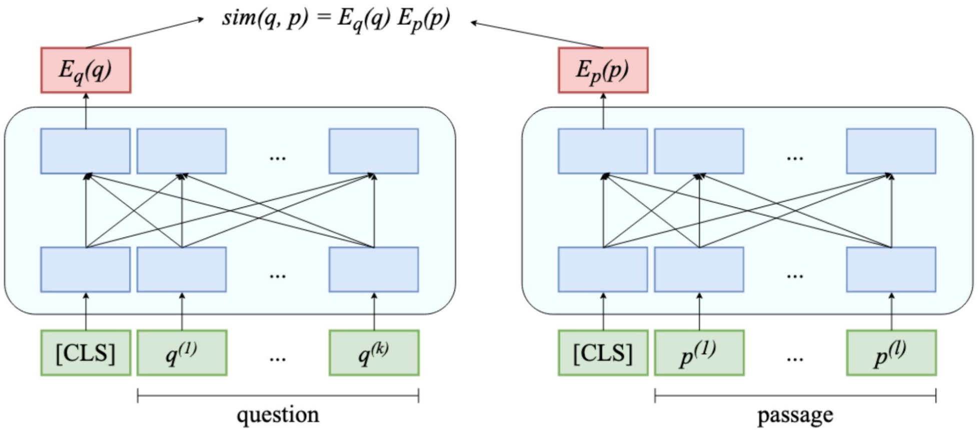


Who is the bad guy in lord of the rings?

Sala Baker is best known for portraying the villain Sauron in the Lord of the Rings trilogy

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering

Dense Passage Retrieval



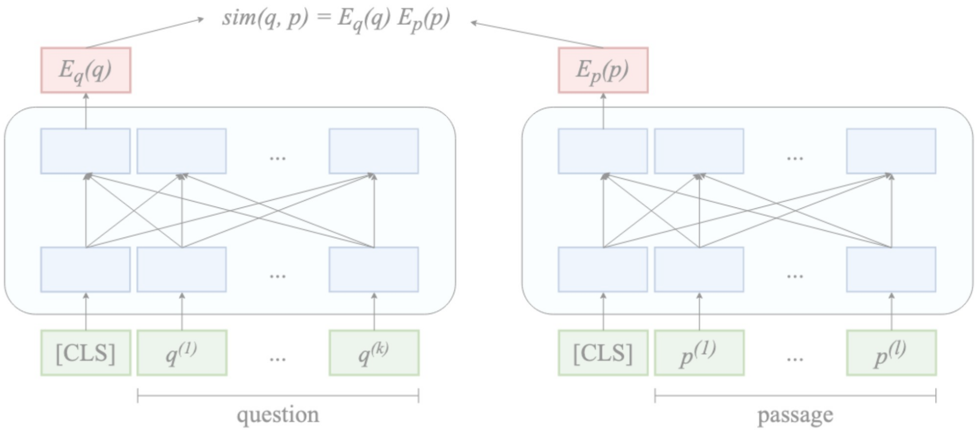
$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i,p_i^+)}}{e^{\mathrm{sim}(q_i,p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i,p_{i,j}^-)}}$$

Who is the bad guy in lord of the rings?

Sala Baker is best known for portraying the villain Sauron in the Lord of the Rings trilogy

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering

Dense Passage Retrieval



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

Who is the bad guy in lord of the rings?

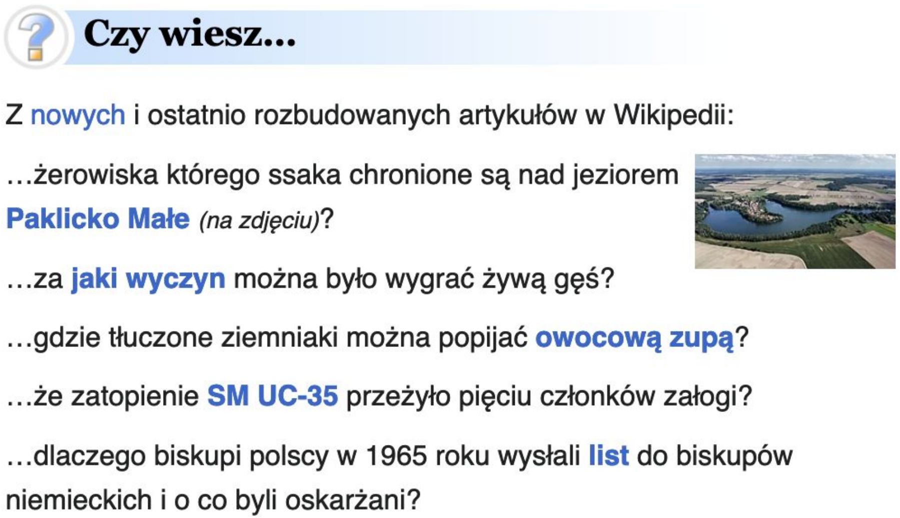
Sala Baker is best known for portraying the villain Sauron in the Lord of the Rings trilogy

Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering

Czy wiesz [2013]

> **4 721** question-*article* pairs

> **250** question-*passage* pairs



Czy wiesz...

Z nowych i ostatnio rozbudowanych artykułów w Wikipedii:

...żerowiska którego ssaka chronione są nad jeziorem **Paklicko Małe** *(na zdjęciu)*?

...za **jaki wyczyn** można było wygrać żywą gęś?

...gdzie tłuczone ziemniaki można popijać **owocową zupą**?

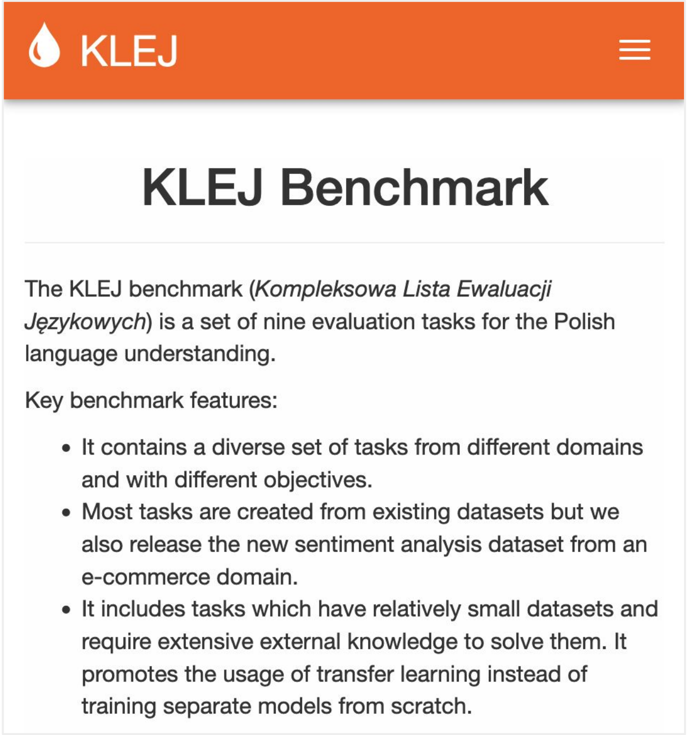
...że zatopienie **SM UC-35** przeżyło pięciu członków załogi?

...dlaczego biskupi polscy w 1965 roku wysłali **list** do biskupów niemieckich i o co byli oskarżani?

http://nlp.pwr.wroc.pl/en/tools-and-resources/resources/czy-wiesz-question-answering-dataset

Czy wiesz [2020]

- > **4 721** question-*article* pairs
- > **250** question-*passage* pairs
- > **1 070** additional question-*passage* pairs



KLEJ Benchmark

The KLEJ benchmark (*Kompleksowa Lista Ewaluacji Językowych*) is a set of nine evaluation tasks for the Polish language understanding.

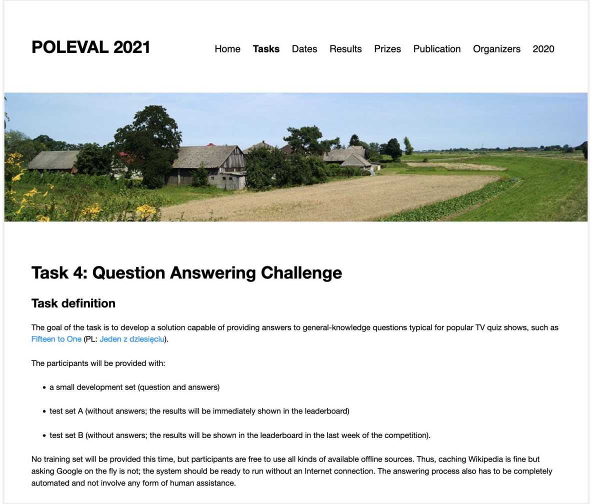
Key benchmark features:

- It contains a diverse set of tasks from different domains and with different objectives.
- Most tasks are created from existing datasets but we also release the new sentiment analysis dataset from an e-commerce domain.
- It includes tasks which have relatively small datasets and require extensive external knowledge to solve them. It promotes the usage of transfer learning instead of training separate models from scratch.

https://klejbenchmark.com

PolEval [2021]

> 6 000 question-*answer* pairs but **no passages** or articles



Task 4: Question Answering Challenge

Task definition

The goal of the task is to develop a solution capable of providing answers to general-knowledge questions typical for popular TV quiz shows, such as Fifteen to One (PL: Jeden z dziesięciu).

The participants will be provided with:

- a small development set (question and answers)
- test set A (without answers; the results will be immediately shown in the leaderboard)
- test set B (without answers; the results will be shown in the leaderboard in the last week of the competition).

No training set will be provided this time, but participants are free to use all kinds of available offline sources. Thus, caching Wikipedia is fine but asking Google on the fly is not; the system should be ready to run without an Internet connection. The answering process also has to be completely automated and not involve any form of human assistance.

Who is the hero in the Tomb Raider video game series?

Lara Croft

Is calcium a metal?

Yes

http://2021.poleval.pl/tasks/task4

PolQA [2022]

Who is the hero in the Tomb Raider video game series?

Lara Croft

Is calcium a metal?

Yes

Rybak et al. Improving Question Answering Performance through Manual Annotation: Costs Benefits and Strategies

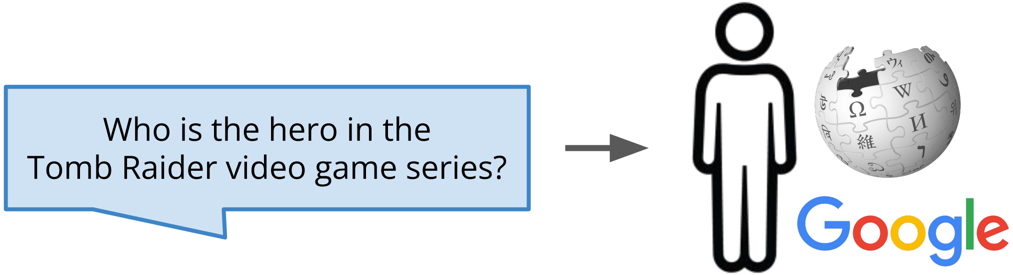
PolQA [2022]

Who is the hero in the Tomb Raider video game series?	**Lara Croft** Lara Croft is a character and the main protagonist of the video game franchise Tomb Raider.	Lara Croft
Is calcium a metal?	**Calcium** As an alkaline earth metal calcium is a reactive metal that forms a dark oxide-nitride layer when exposed to air.	Yes

Rybak et al. Improving Question Answering Performance through Manual Annotation: Costs Benefits and Strategies

PolQA [2022]

> **Standard** approach:

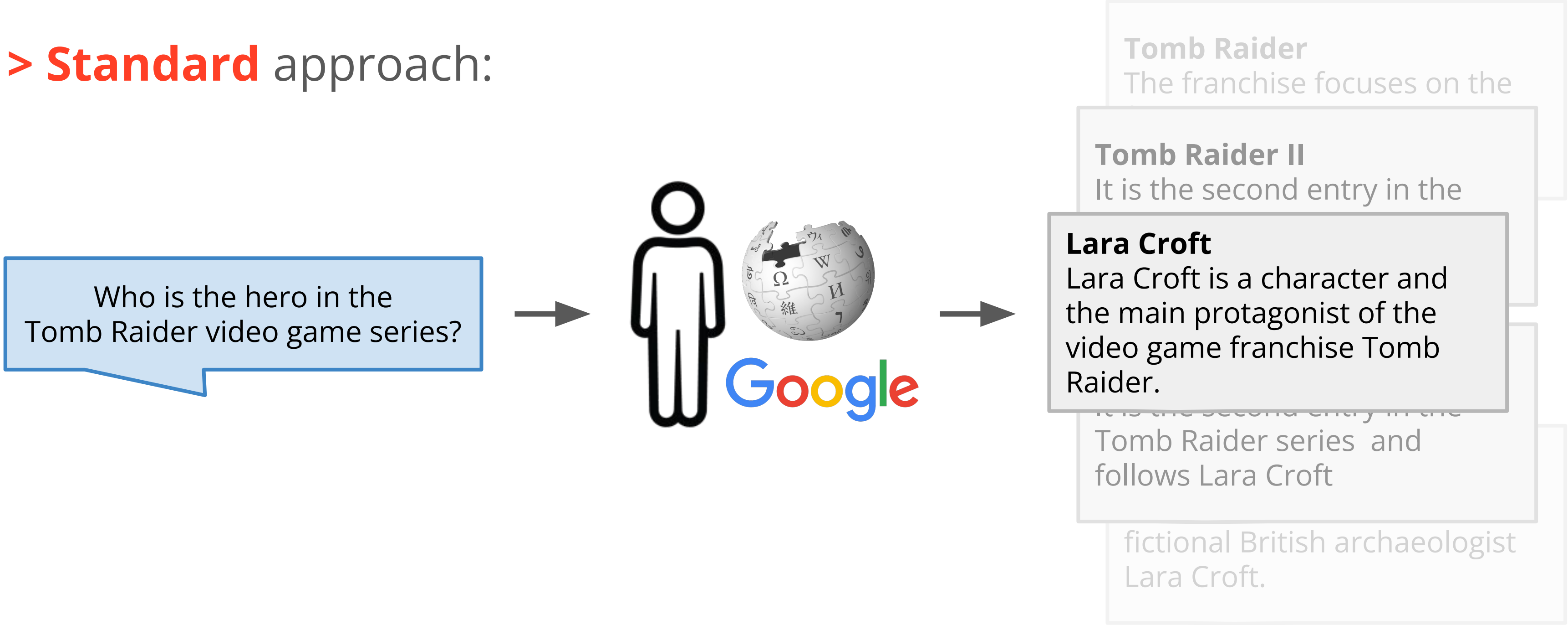


Who is the hero in the Tomb Raider video game series?

Rybak et al. Improving Question Answering Performance through Manual Annotation: Costs Benefits and Strategies

PolQA [2022]

> **Standard** approach:



Who is the hero in the Tomb Raider video game series?

Tomb Raider
The franchise focuses on the

Tomb Raider II
It is the second entry in the

Lara Croft
Lara Croft is a character and the main protagonist of the video game franchise Tomb Raider.

It is the second entry in the Tomb Raider series and follows Lara Croft

fictional British archaeologist Lara Croft.

Rybak et al. Improving Question Answering Performance through Manual Annotation: Costs Benefits and Strategies

PolQA [2022]

> **Efficient** approach:

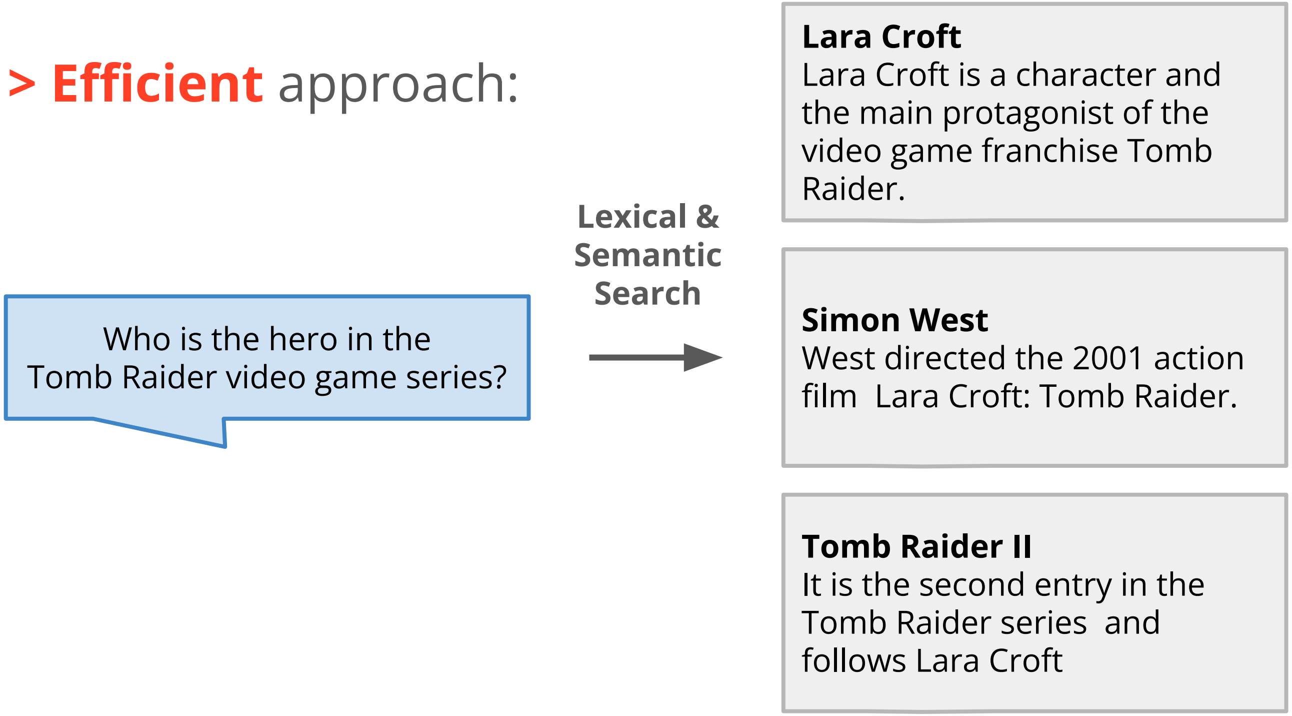
Who is the hero in the
Tomb Raider video game series?

Lexical & Semantic Search →

Rybak et al. Improving Question Answering Performance through Manual Annotation: Costs Benefits and Strategies

PolQA [2022]

> **Efficient** approach:



Who is the hero in the Tomb Raider video game series?

Lexical & Semantic Search

Lara Croft
Lara Croft is a character and the main protagonist of the video game franchise Tomb Raider.

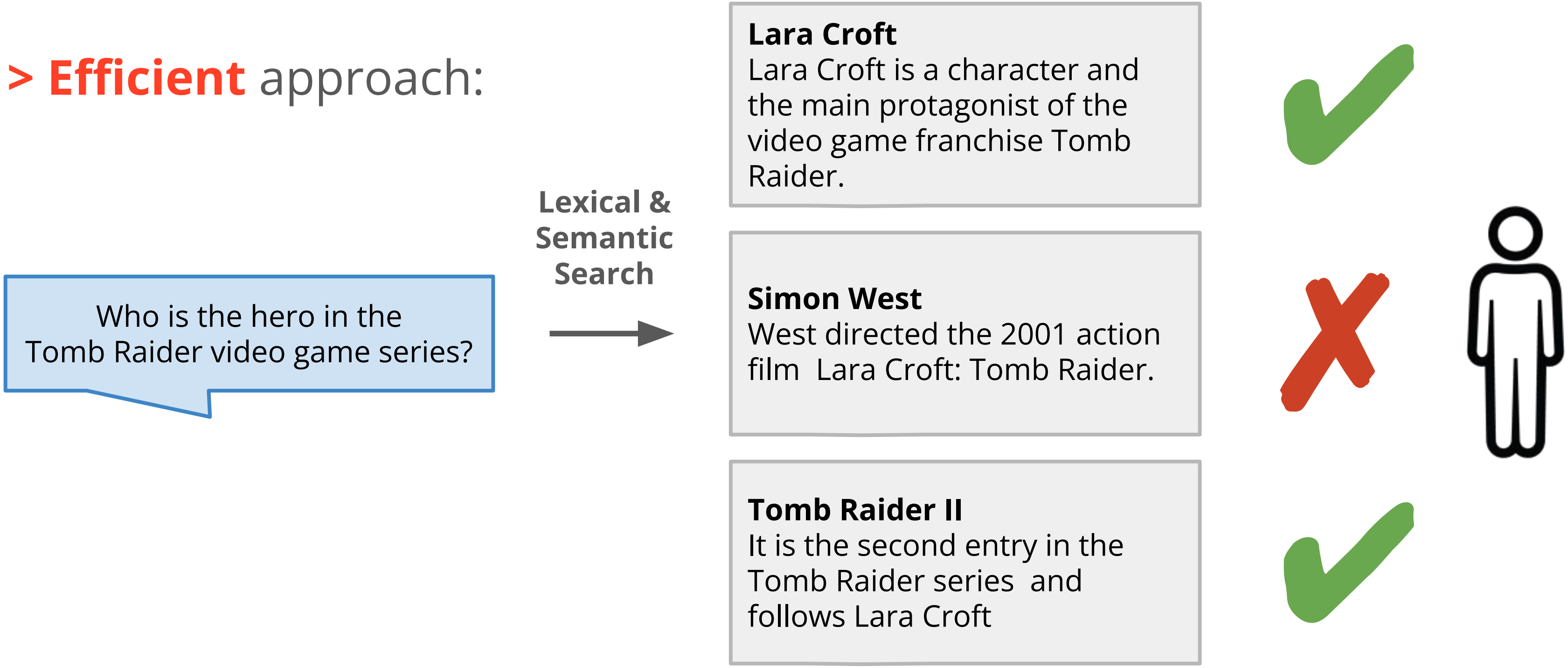
Simon West
West directed the 2001 action film Lara Croft: Tomb Raider.

Tomb Raider II
It is the second entry in the Tomb Raider series and follows Lara Croft

Rybak et al. Improving Question Answering Performance through Manual Annotation: Costs Benefits and Strategies

PolQA [2022]

> **Efficient** approach:



Who is the hero in the Tomb Raider video game series?

Lexical & Semantic Search

Lara Craft
Lara Croft is a character and the main protagonist of the video game franchise Tomb Raider. ✓

Simon West
West directed the 2001 action film Lara Croft: Tomb Raider. ✗

Tomb Raider II
It is the second entry in the Tomb Raider series and follows Lara Croft ✓

Rybak et al. Improving Question Answering Performance through Manual Annotation: Costs Benefits and Strategies

PolQA [2022]

- **> 7 000** questions
- **> 39k** positive pairs
- **> 48k** hard negative pairs

Model	**Annotation Time [s]**	**Accuracy@10**
Standard approach	376	51.5%
Efficient approach	68	62.0%

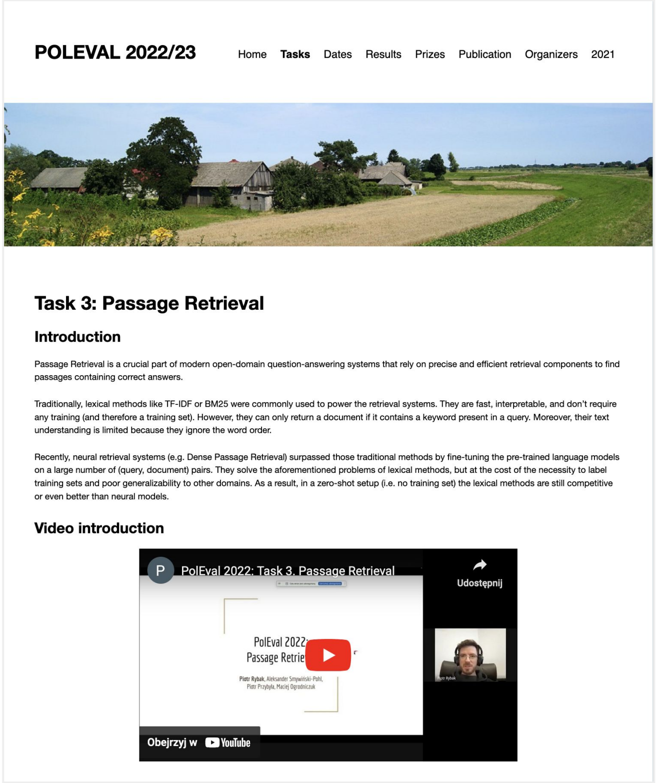
Evaluated on a test set gathered using Standard approach

Rybak et al. Improving Question Answering Performance through Manual Annotation: Costs Benefits and Strategies

PolEval [2022/2023]

> **Passage retrieval** task

> **PolQA** as a **train** set



POLEVAL 2022/23

Home **Tasks** Dates Results Prizes Publication Organizers 2021

Task 3: Passage Retrieval

Introduction

Passage Retrieval is a crucial part of modern open-domain question-answering systems that rely on precise and efficient retrieval components to find passages containing correct answers.

Traditionally, lexical methods like TF-IDF or BM25 were commonly used to power the retrieval systems. They are fast, interpretable, and don't require any training (and therefore a training set). However, they can only return a document if it contains a keyword present in a query. Moreover, their text understanding is limited because they ignore the word order.

Recently, neural retrieval systems (e.g. Dense Passage Retrieval) surpassed those traditional methods by fine-tuning the pre-trained language models on a large number of (query, document) pairs. They solve the aforementioned problems of lexical methods, but at the cost of the necessity to label training sets and poor generalizability to other domains. As a result, in a zero-shot setup (i.e. no training set) the lexical methods are still competitive or even better than neural models.

Video introduction

P
PolEval 2022: Task 3. Passage Retrieval
Udostępnij
PolEval 2022:
Passage Retrie
Piotr Rybak, Aleksander Smywiński-Pohl,
Piotr Przybyła, Maciej Ogrodniczuk
Piotr Rybak
Obejrzyj w YouTube

http://poleval.pl/tasks/task3

PolEval [2022/2023]

- **Passage retrieval** task
- **PolQA** as a **train** set
- Three **test** sets
  - PolQA
  - Customer Support FAQ
  - Legal Questions

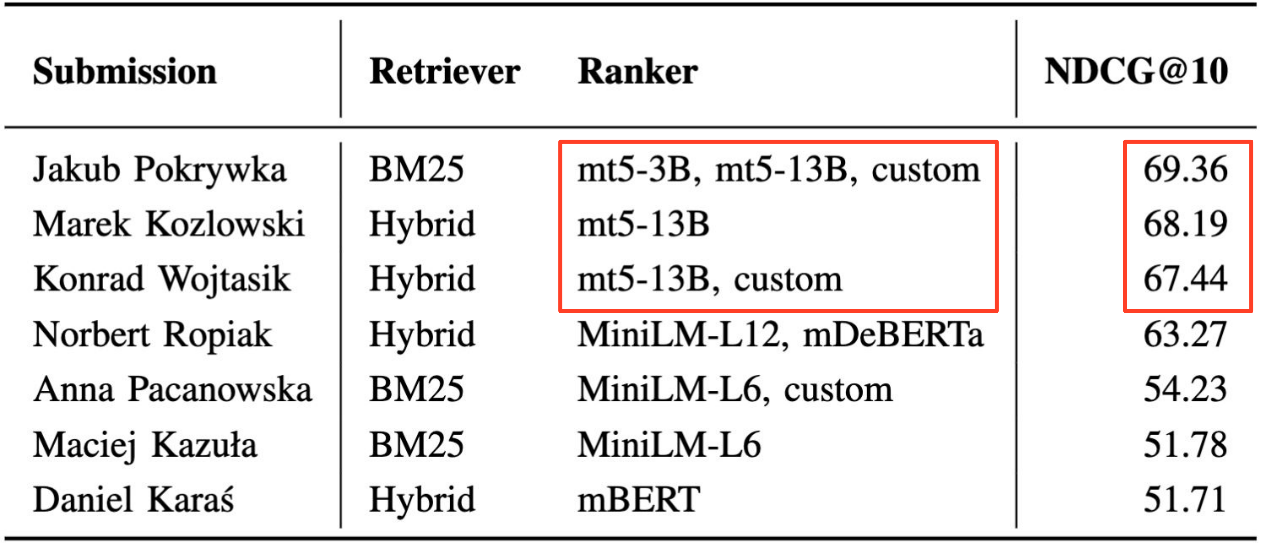
Who is the hero in the
Tomb Raider video game series?

Can I change my user name?

How many budgets does a city with
county rights have?

http://poleval.pl/tasks/task3

PolEval [2022/2023]



Submission	Retriever	Ranker	NDCG@10
Jakub Pokrywka	BM25	mt5-3B, mt5-13B, custom	69.36
Marek Kozlowski	Hybrid	mt5-13B	68.19
Konrad Wojtasik	Hybrid	mt5-13B, custom	67.44
Norbert Ropiak	Hybrid	MiniLM-L12, mDeBERTa	63.27
Anna Pacanowska	BM25	MiniLM-L6, custom	54.23
Maciej Kazuła	BM25	MiniLM-L6	51.78
Daniel Karaś	Hybrid	mBERT	51.71

http://poleval.pl/tasks/task3

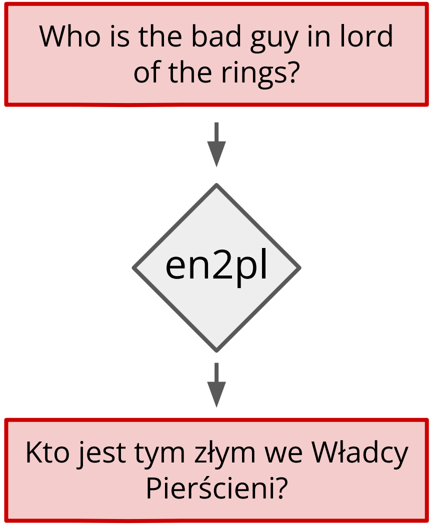
MAUPQA [2023.05]

- > **14** datasets
- > **1M** positive pairs
- > **7M** hard negative pairs

dataset	# questions	# positives	# negatives
1z10	22 835	22 014	139 471
czy-wiesz-v2	29 078	29 078	143 306
gpt3-cc	10 146	10 177	89 203
gpt3.5-cc	29 591	29 720	251 959
gpt3.5-wiki	29 674	29 748	115 564
mkqa	4 036	3 968	19 814
mqa	172 768	178 131	1 249 659
msmarco	389 987	416 763	3 006 996
multilingual-NLI	100 752	68 096	743 857
nq	135 781	139 976	797 436
poleval2021-pairs	1 977	2 088	17 608
poquad	56 588	46 187	299 865
templates	15 993	15 993	45 228
wiki-def	18 093	18 093	84 956
Total	1 017 299	1 010 032	7 004 922

https://huggingface.co/datasets/ipipan/maupqa

MAUPQA [2023.05]


Who is the bad guy in lord of the rings?

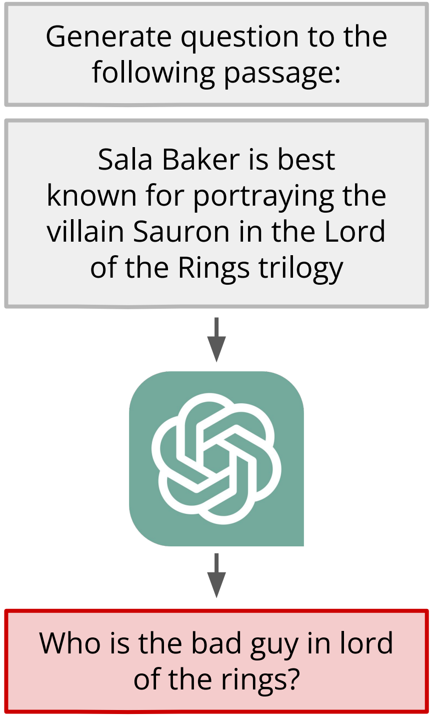
en2pl

Kto jest tym złym we Władcy Pierścieni?

dataset	# questions	# positives	# negatives
1z10	22 835	22 014	139 471
czy-wiesz-v2	29 078	29 078	143 306
gpt3-cc	10 146	10 177	89 203
gpt3.5-cc	29 591	29 720	251 959
gpt3.5-wiki	29 674	29 748	115 564
mkqa	4 036	3 968	19 814
mqa	172 768	178 131	1 249 659
msmarco	389 987	416 763	3 006 996
multilingual-NLI	100 752	68 096	743 857
nq	135 781	139 976	797 436
poleval2021-pairs	1 977	2 088	17 608
poquad	56 588	46 187	299 865
templates	15 993	15 993	45 228
wiki-def	18 093	18 093	84 956
Total	1 017 299	1 010 032	7 004 922

https://huggingface.co/datasets/ipipan/maupqa

MAUPQA [2023.05]



Generate question to the following passage:

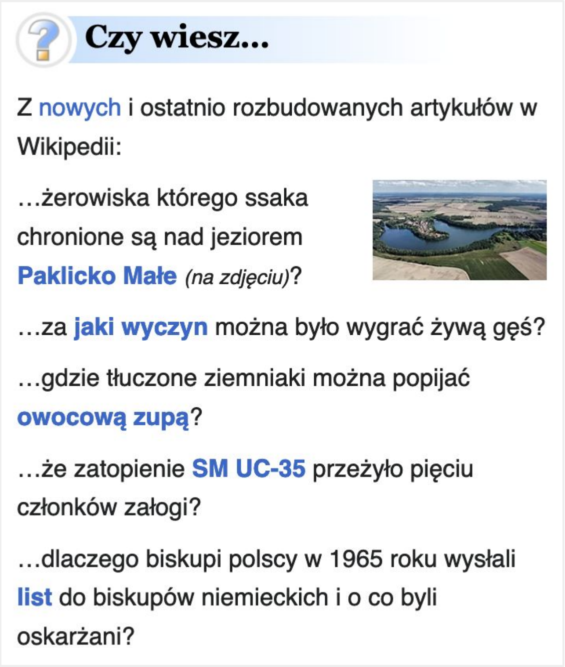
Sala Baker is best known for portraying the villain Sauron in the Lord of the Rings trilogy

Who is the bad guy in lord of the rings?

dataset	# questions	# positives	# negatives
1z10	22 835	22 014	139 471
czy-wiesz-v2	29 078	29 078	143 306
gpt3-cc	10 146	10 177	89 203
gpt3.5-cc	29 591	29 720	251 959
gpt3.5-wiki	29 674	29 748	115 564
mkqa	4 036	3 968	19 814
mqa	172 768	178 131	1 249 659
msmarco	389 987	416 763	3 006 996
multilingual-NLI	100 752	68 096	743 857
nq	135 781	139 976	797 436
poleval2021-pairs	1 977	2 088	17 608
poquad	56 588	46 187	299 865
templates	15 993	15 993	45 228
wiki-def	18 093	18 093	84 956
Total	1 017 299	1 010 032	7 004 922

https://huggingface.co/datasets/ipipan/maupqa

MAUPQA [2023.05]



Czy wiesz...

Z nowych i ostatnio rozbudowanych artykułów w Wikipedii:

…żerowiska którego ssaka chronione są nad jeziorem **Paklicko Małe** *(na zdjęciu)*?

…za **jaki wyczyn** można było wygrać żywą gęś?

…gdzie tłuczone ziemniaki można popijać **owocową zupą**?

…że zatopienie **SM UC-35** przeżyło pięciu członków załogi?

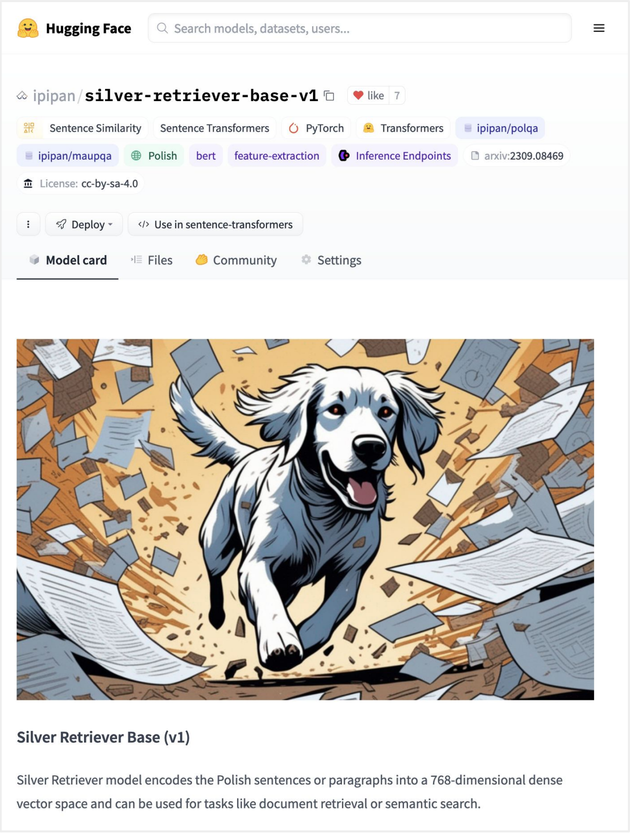
…dlaczego biskupi polscy w 1965 roku wysłali **list** do biskupów niemieckich i o co byli oskarżani?

dataset	# questions	# positives	# negatives
1z10	22 835	22 014	139 471
czy-wiesz-v2	29 078	29 078	143 306
gpt3-cc	10 146	10 177	89 203
gpt3.5-cc	29 591	29 720	251 959
gpt3.5-wiki	29 674	29 748	115 564
mkqa	4 036	3 968	19 814
mqa	172 768	178 131	1 249 659
msmarco	389 987	416 763	3 006 996
multilingual-NLI	100 752	68 096	743 857
nq	135 781	139 976	797 436
poleval2021-pairs	1 977	2 088	17 608
poquad	56 588	46 187	299 865
templates	15 993	15 993	45 228
wiki-def	18 093	18 093	84 956
Total	1 017 299	1 010 032	7 004 922

https://huggingface.co/datasets/ipipan/maupqa

Silver Retriever [2023.08]

- > Fine-tuned on **PolQA** & **MAUPQA**
- > Reranking models for **denoising**
- > PolEval22 test-B (NDCG@10):
  - > **64.43** for Silver Retriever
  - > **48.78** for lexical BM25



Silver Retriever Base (v1)

Silver Retriever model encodes the Polish sentences or paragraphs into a 768-dimensional dense vector space and can be used for tasks like document retrieval or semantic search.

https://huggingface.co/ipipan/silver-retriever-base-v1

BEIR-PL [2023.05]

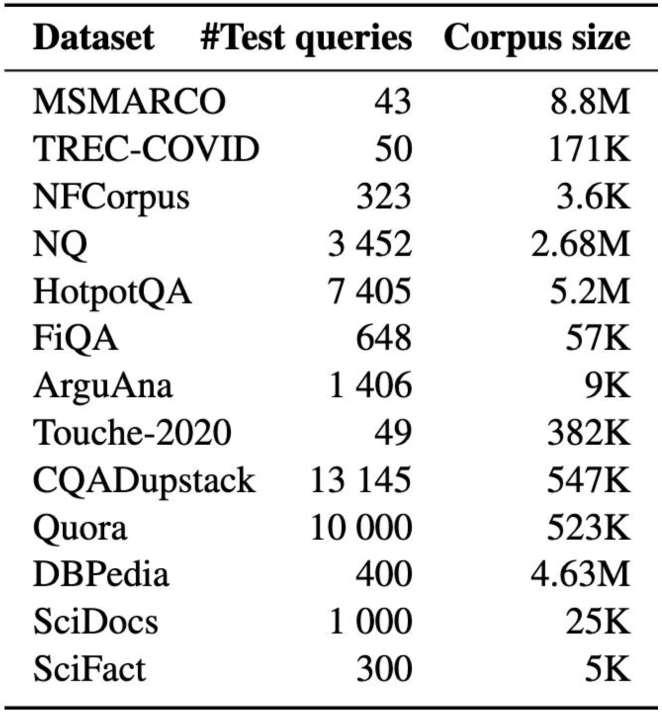
- > **Machine translated**

  English benchmark

- > **13** datasets

- > **9** evaluated models

- > Polish **reranking** models

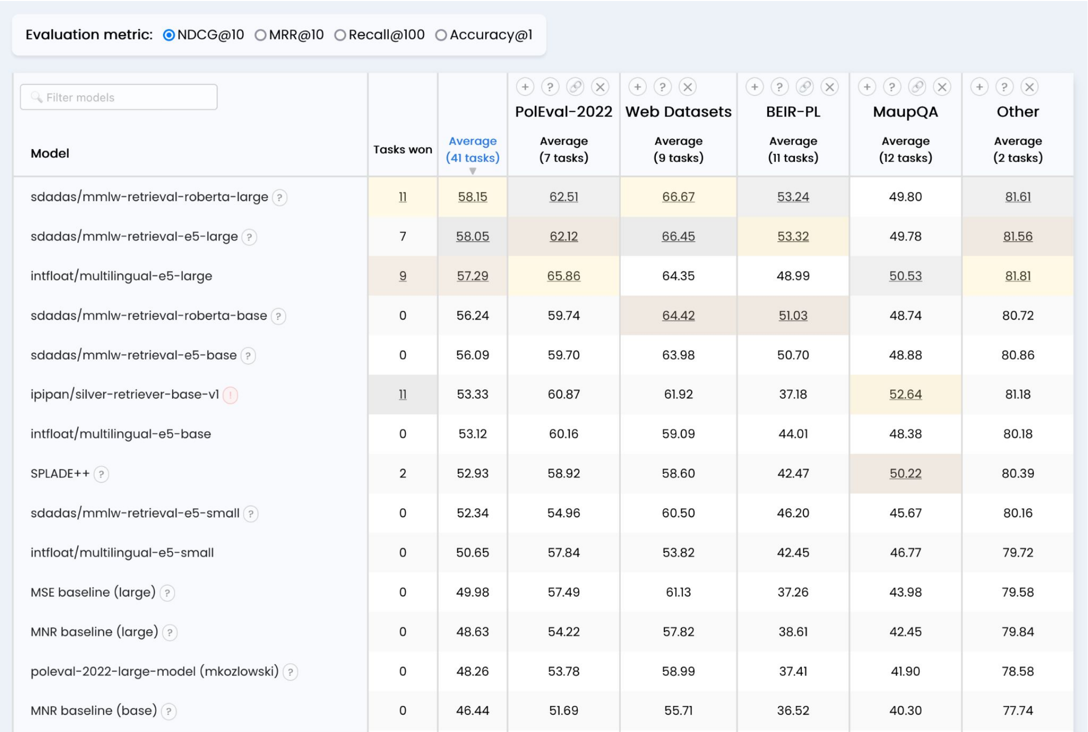


Dataset	#Test queries	Corpus size
MSMARCO	43	8.8M
TREC-COVID	50	171K
NFCorpus	323	3.6K
NQ	3 452	2.68M
HotpotQA	7 405	5.2M
FiQA	648	57K
ArguAna	1 406	9K
Touche-2020	49	382K
CQADupstack	13 145	547K
Quora	10 000	523K
DBPedia	400	4.63M
SciDocs	1 000	25K
SciFact	300	5K

Wojtasik et al., BEIR-PL: Zero Shot Information Retrieval Benchmark for the Polish Language

Polish Information Retrieval Benchmark [2023.09]

- > **41** datasets
  - > PolEval 2022
  - > BEIR-PL
  - > MAUPQA
  - > Crawled QA sites
- > **29** evaluated models



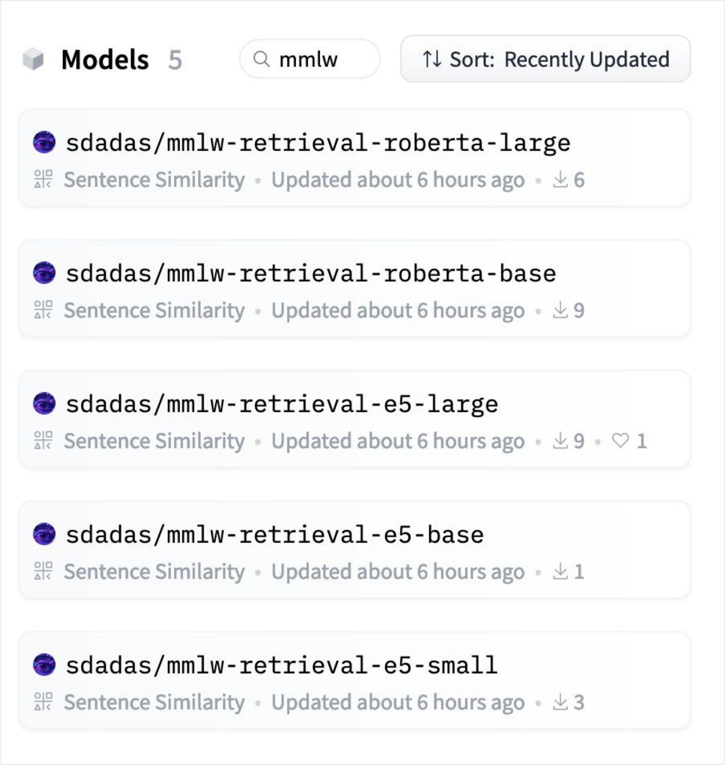
Evaluation metric: ● NDCG@10 ○ MRR@10 ○ Recall@100 ○ Accuracy@1

			PolEval-2022	Web Datasets	BEIR-PL	MaupQA	Other
Model	Tasks won	Average (41 tasks)	Average (7 tasks)	Average (9 tasks)	Average (11 tasks)	Average (12 tasks)	Average (2 tasks)
sdadas/mmlw-retrieval-roberta-large	11	58.15	62.51	66.67	53.24	49.80	81.61
sdadas/mmlw-retrieval-e5-large	7	58.05	62.12	66.45	53.32	49.78	81.56
intfloat/multilingual-e5-large	9	57.29	65.86	64.35	48.99	50.53	81.81
sdadas/mmlw-retrieval-roberta-base	0	56.24	59.74	64.42	51.03	48.74	80.72
sdadas/mmlw-retrieval-e5-base	0	56.09	59.70	63.98	50.70	48.88	80.86
ipipan/silver-retriever-base-v1	11	53.33	60.87	61.92	37.18	52.64	81.18
intfloat/multilingual-e5-base	0	53.12	60.16	59.09	44.01	48.38	80.18
SPLADE++	2	52.93	58.92	58.60	42.47	50.22	80.39
sdadas/mmlw-retrieval-e5-small	0	52.34	54.96	60.50	46.20	45.67	80.16
intfloat/multilingual-e5-small	0	50.65	57.84	53.82	42.45	46.77	79.72
MSE baseline (large)	0	49.98	57.49	61.13	37.26	43.98	79.58
MNR baseline (large)	0	48.63	54.22	57.82	38.61	42.45	79.84
poleval-2022-large-model (mkozlowski)	0	48.26	53.78	58.99	37.41	41.90	78.58
MNR baseline (base)	0	46.44	51.69	55.71	36.52	40.30	77.74

https://huggingface.co/spaces/sdadas/pirb

MMLW Retrieval [2023.10]

- **>** Small, Base, Large variants
- **>** Knowledge distillation from **English BGE** model
- **>** Fine-tuned on **translated MS MARCO**



sdadas/mmlw-retrieval-roberta-large
Sentence Similarity • Updated about 6 hours ago • 6

sdadas/mmlw-retrieval-roberta-base
Sentence Similarity • Updated about 6 hours ago • 9

sdadas/mmlw-retrieval-e5-large
Sentence Similarity • Updated about 6 hours ago • 9 • 1

sdadas/mmlw-retrieval-e5-base
Sentence Similarity • Updated about 6 hours ago • 1

sdadas/mmlw-retrieval-e5-small
Sentence Similarity • Updated about 6 hours ago • 3

https://huggingface.co/sdadas?search_models=mmlw

Summary

> In **2019**, we had **250** training examples and **0** models

Summary

- In **2019**, we had **250** training examples and **0** models
- In **2023**, we have:
  - over **1M** training examples
  - at least **8 retrieval** and **4 reranking** open-source models
  - **41** benchmark datasets

THANK YOU

piotr.cezary.rybak@gmail.com