COS 484

Natural Language Processing

L19: Question Answering

Spring 2023

Announcements

- A4 deadline extended by 48 hours
- This lecture: question answering
- The next two lectures are guest lectures:

Prof. He He (NYU)

Natural language generation

Prof. Karthik Narasimhan

Language grounding

- No lecture on April 26th (next Wednesday) - we will schedule meetings for project discussion!

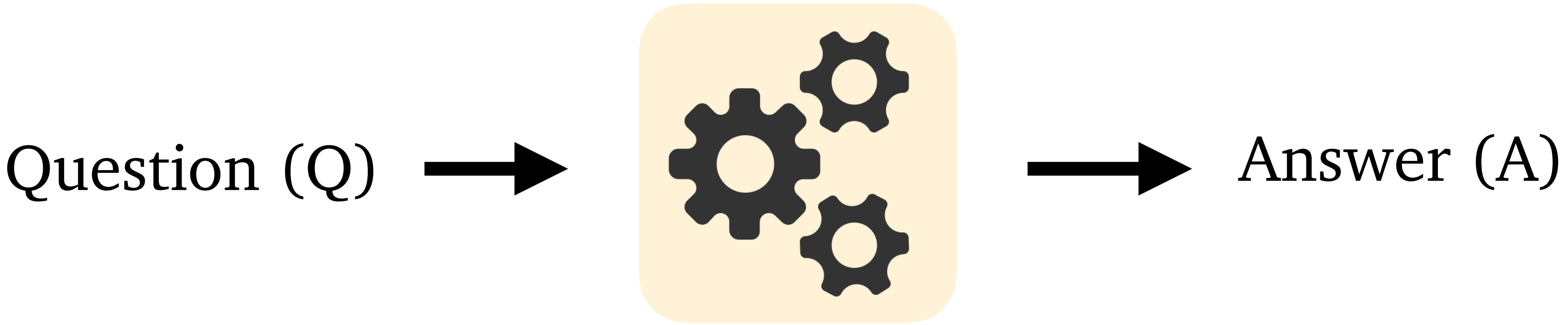
This lecture

1. What is question answering?

2. Reading comprehension
   ✓ How to answer questions over **a single passage of text**

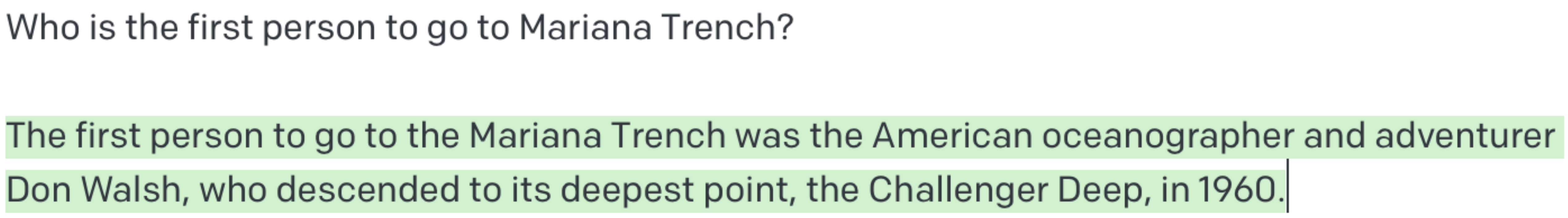
3. Open-domain question answering
   ✓ How to answer questions over **a large collection of documents**

1. What is question answering?



Question (Q) → Answer (A)

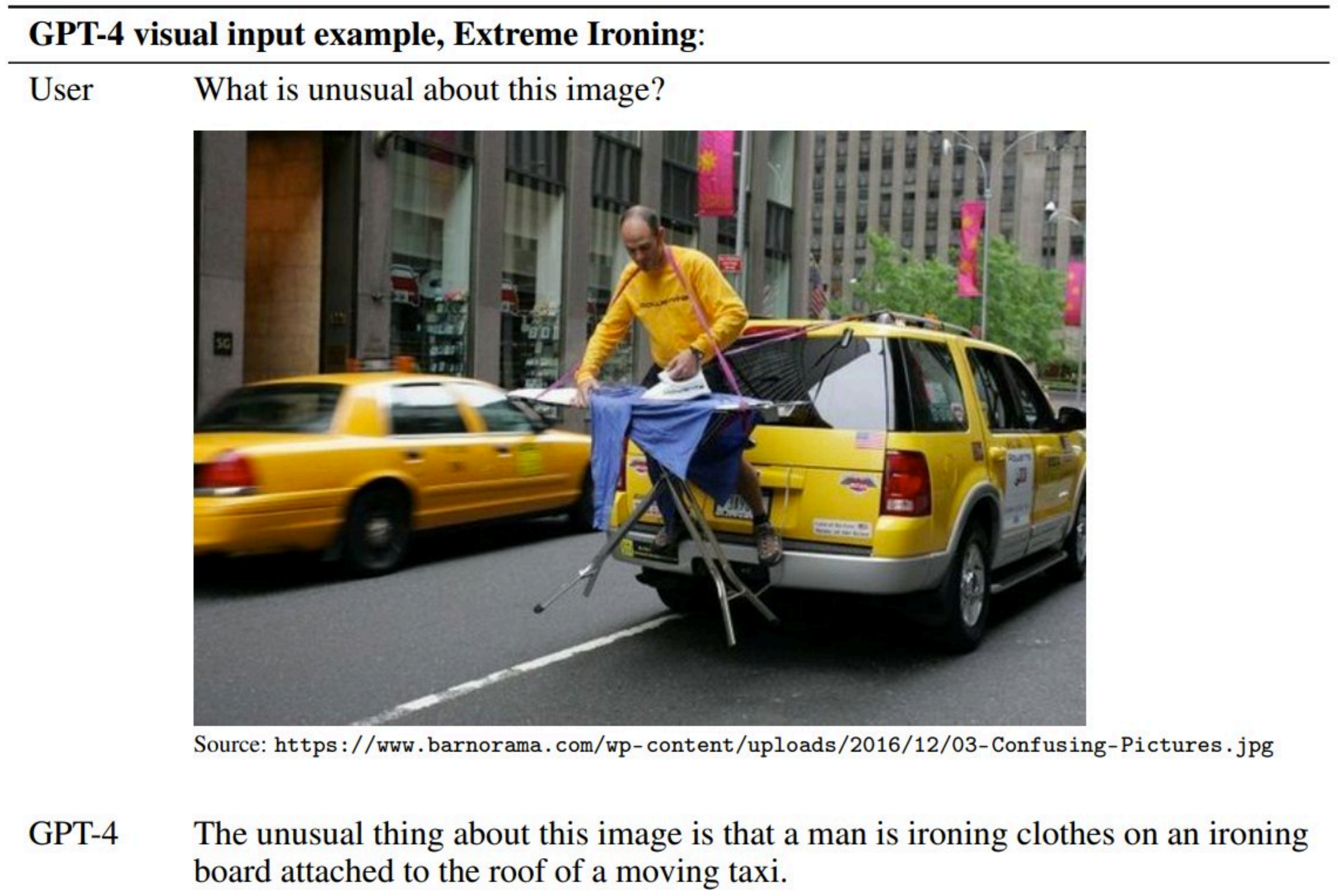
The goal of question answering is to build systems that **automatically** answer questions posed by humans in a **natural language**



Who is the first person to go to Mariana Trench?

The first person to go to the Mariana Trench was the American oceanographer and adventurer Don Walsh, who descended to its deepest point, the Challenger Deep, in 1960.

Q. Are you happy with the answer from users' perspective?



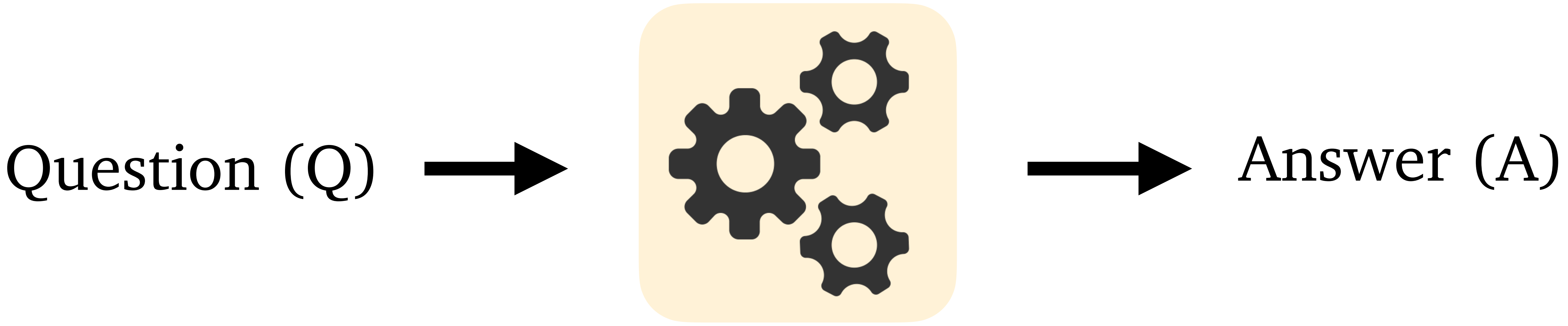
GPT-4 visual input example, Extreme Ironing:

User What is unusual about this image?

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

1. What is question answering?

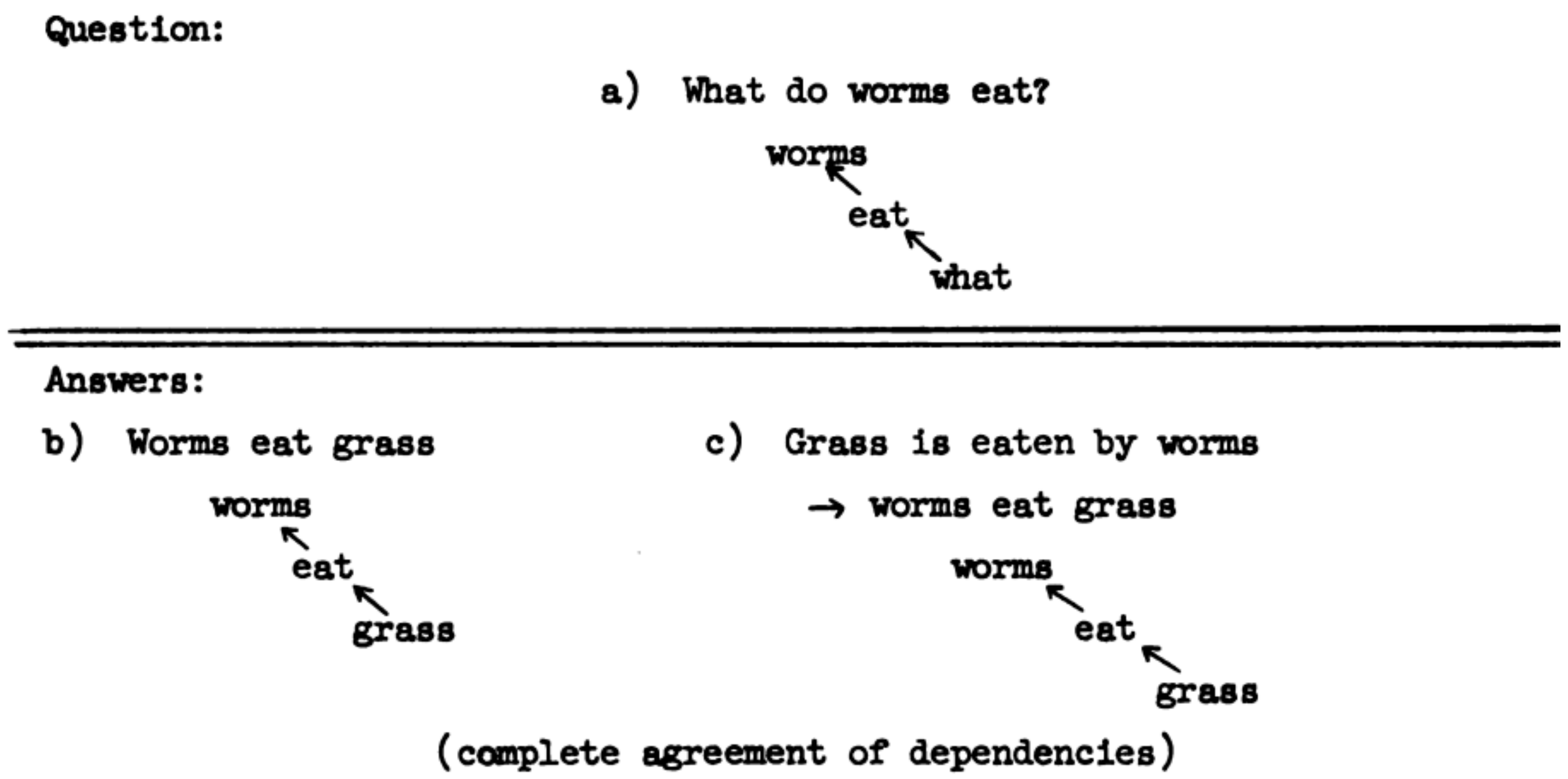


Question (Q) → Answer (A)

The goal of question answering is to build systems that **automatically** answer questions posed by humans in a **natural language**

The earliest QA systems dated back to 1960s!

(Simmons et al., 1964)



Question:

a) What do worms eat?

Answers:

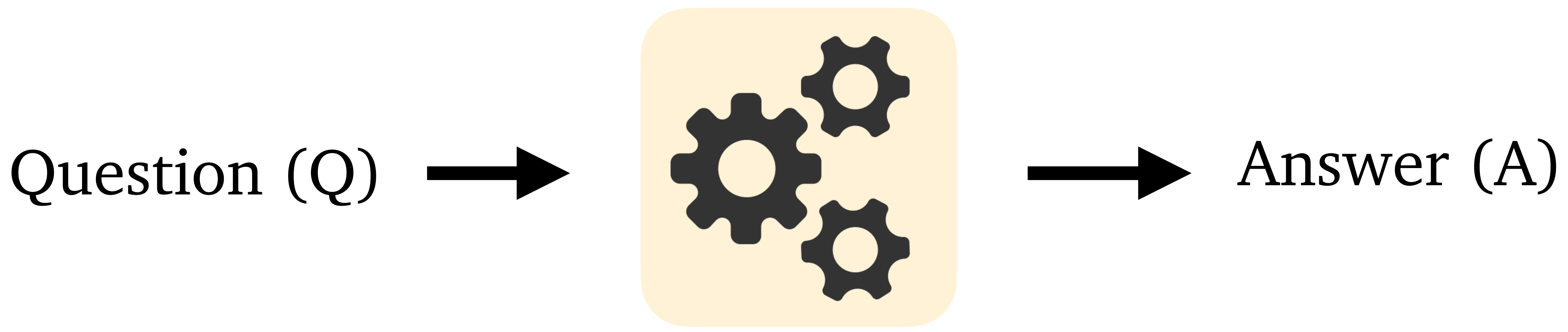
b) Worms eat grass

c) Grass is eaten by worms

→ worms eat grass

(complete agreement of dependencies)

Question answering: a taxonomy

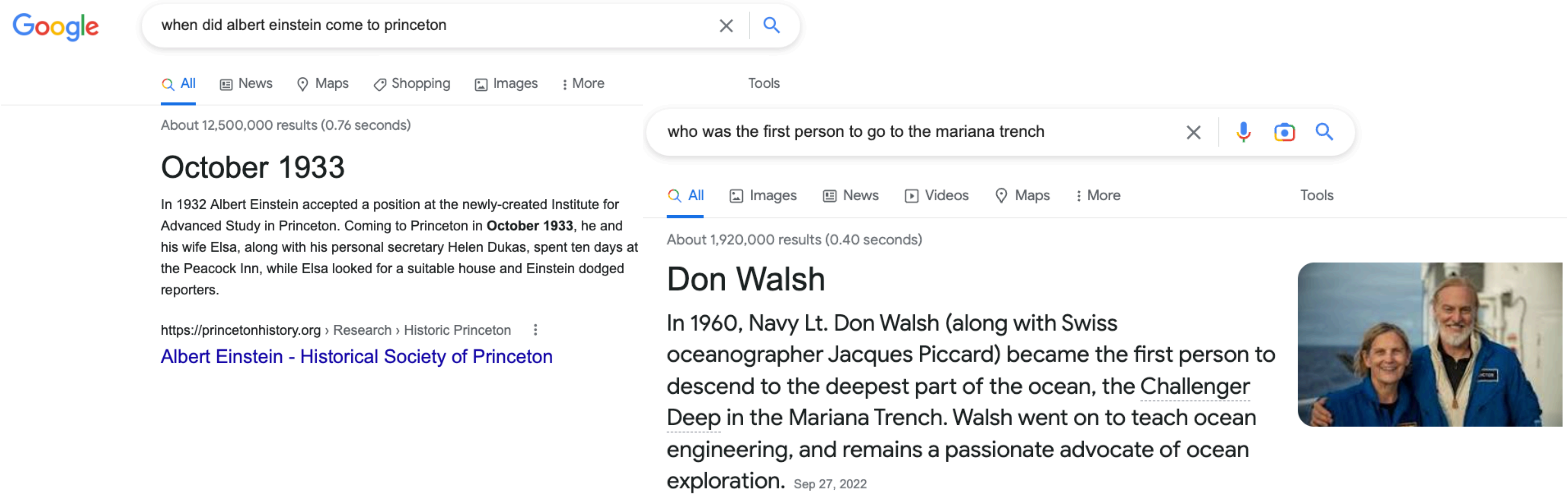


Question (Q)

Answer (A)

- What information source does a system build on?
    - A text passage, all Web documents, knowledge bases, tables, images..
- Question type
    - Factoid vs non-factoid, open-domain vs closed-domain, simple vs compositional, ..
- Answer type
    - A short segment of text, a paragraph, a list, yes/no, …

Lots of practical applications



when did albert einstein come to princeton

About 12,500,000 results (0.76 seconds)

October 1933

In 1932 Albert Einstein accepted a position at the newly-created Institute for Advanced Study in Princeton. Coming to Princeton in **October 1933**, he and his wife Elsa, along with his personal secretary Helen Dukas, spent ten days at the Peacock Inn, while Elsa looked for a suitable house and Einstein dodged reporters.

https://princetonhistory.org › Research › Historic Princeton

Albert Einstein - Historical Society of Princeton

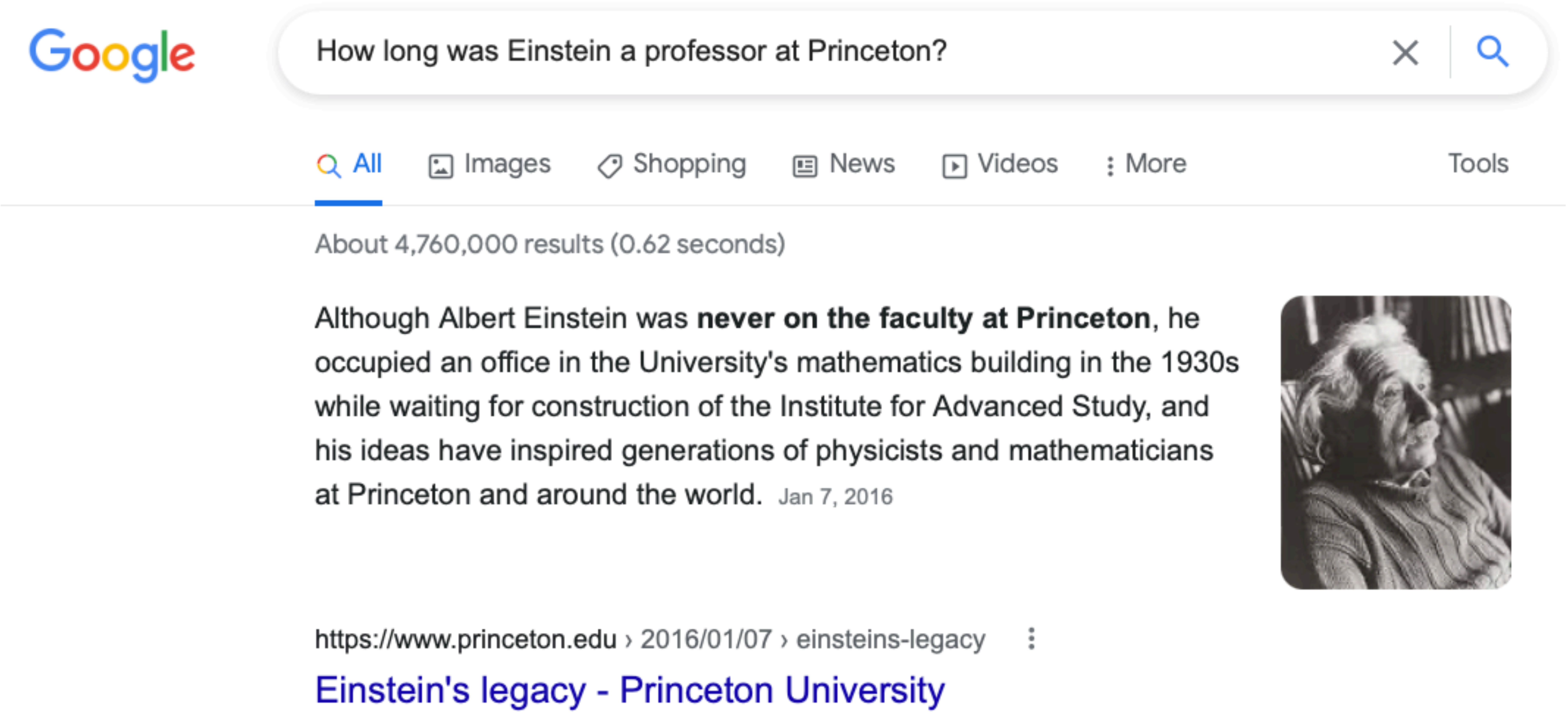
who was the first person to go to the mariana trench

About 1,920,000 results (0.40 seconds)

Don Walsh

In 1960, Navy Lt. Don Walsh (along with Swiss oceanographer Jacques Piccard) became the first person to descend to the deepest part of the ocean, the Challenger Deep in the Mariana Trench. Walsh went on to teach ocean engineering, and remains a passionate advocate of ocean exploration. Sep 27, 2022

Lots of practical applications



How long was Einstein a professor at Princeton?

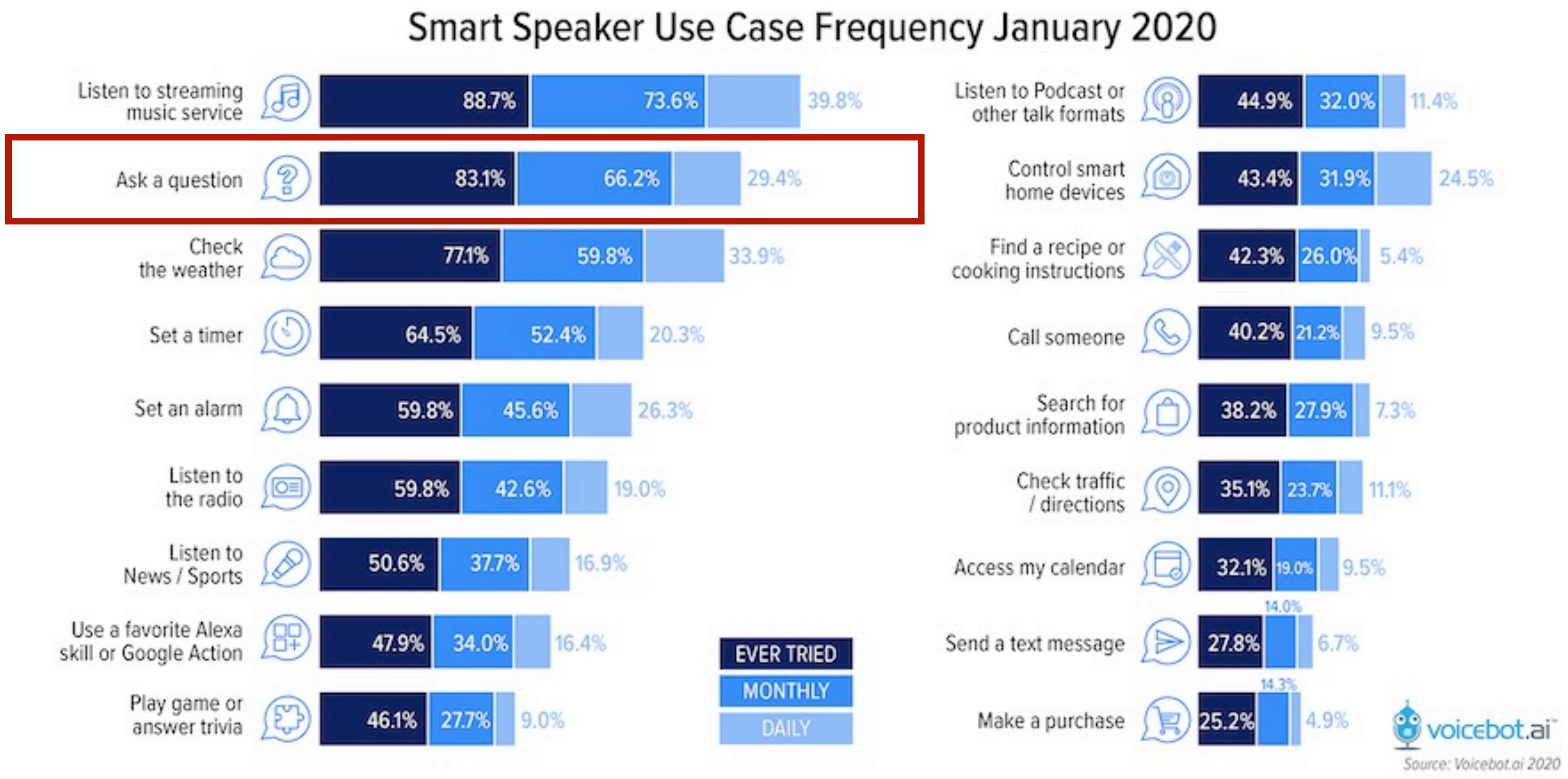
About 4,760,000 results (0.62 seconds)

Although Albert Einstein was **never on the faculty at Princeton**, he occupied an office in the University's mathematics building in the 1930s while waiting for construction of the Institute for Advanced Study, and his ideas have inspired generations of physicists and mathematicians at Princeton and around the world. Jan 7, 2016

https://www.princeton.edu › 2016/01/07 › einsteins-legacy

Einstein's legacy - Princeton University

Lots of practical applications



Source: Voicebot.ai 2020

IBM Watson beated Jeopardy champions

IBM Watson defeated two of Jeopardy's greatest champions in 2011

IBM Watson beated Jeopardy champions

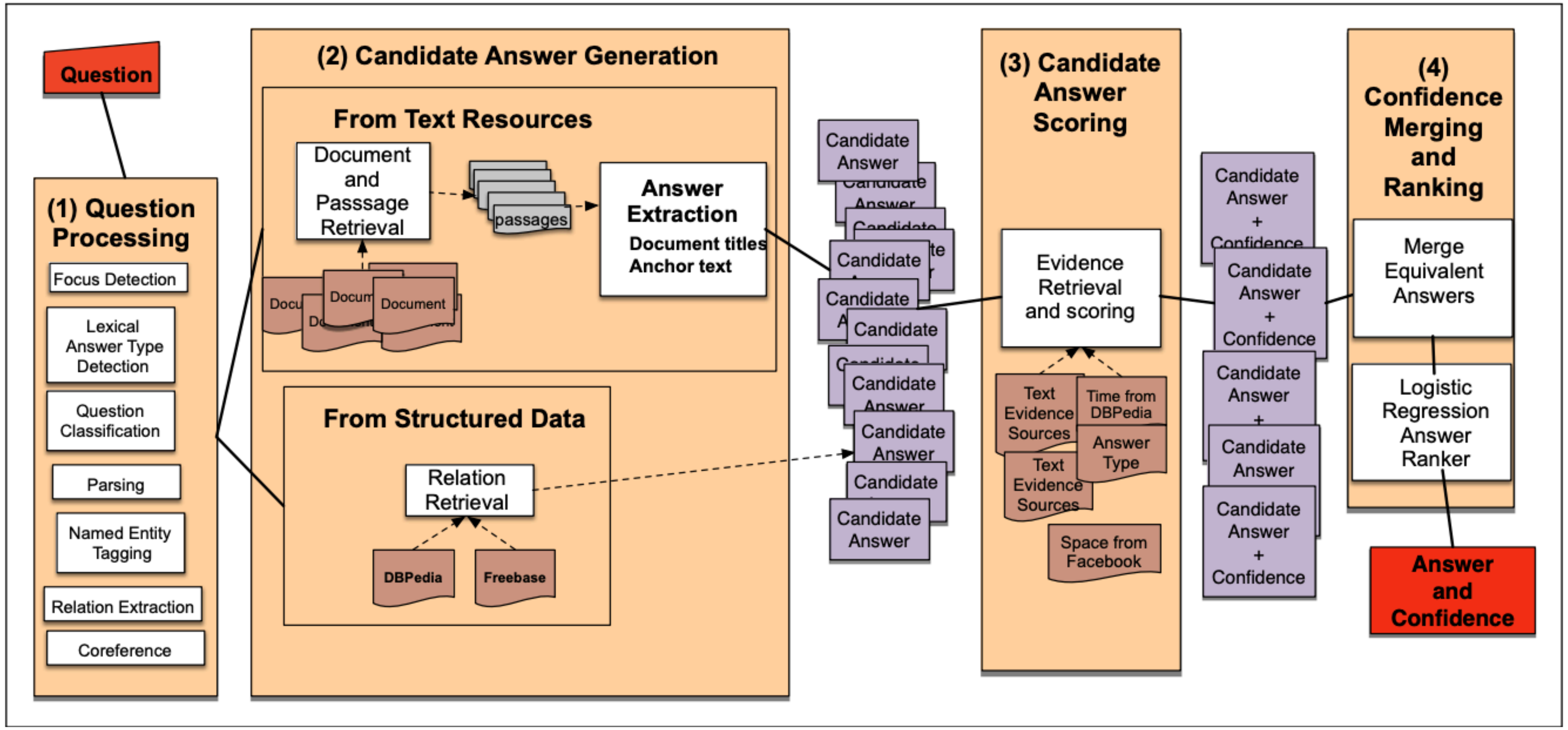
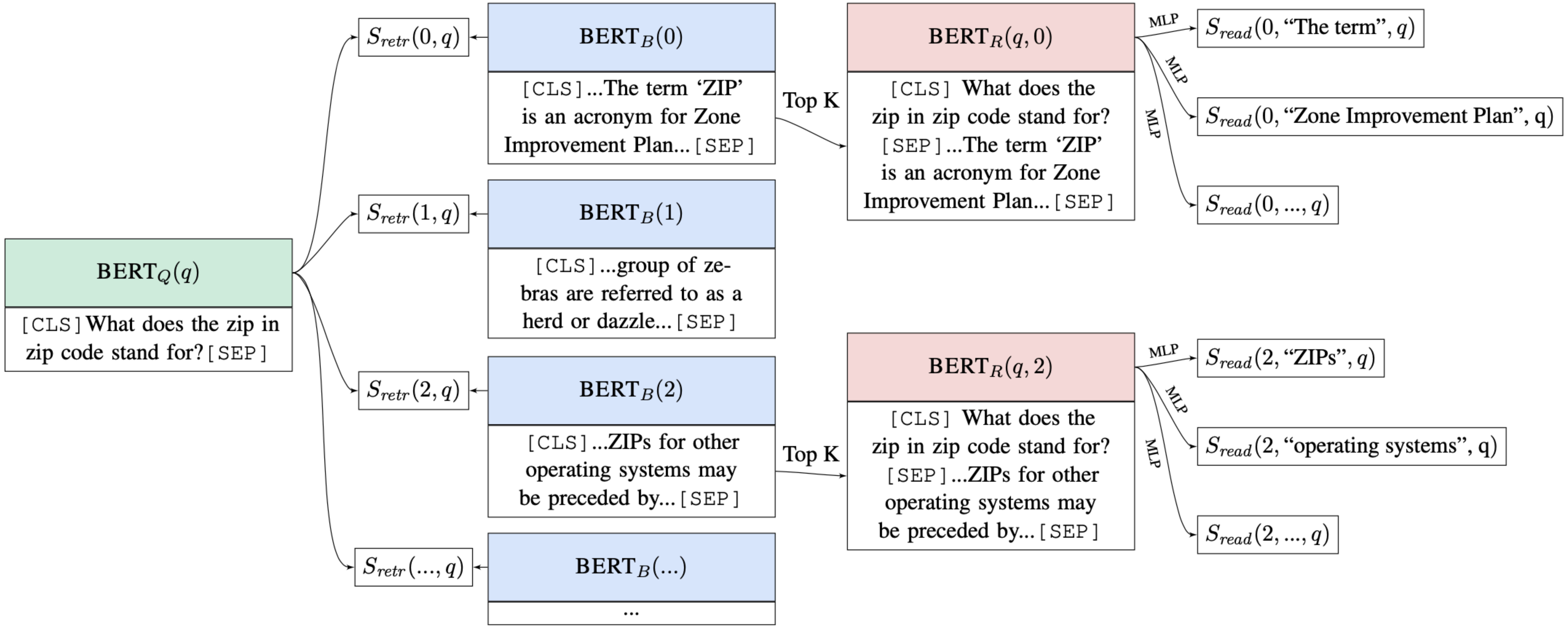


Image credit: J & M, edition 3

(1) Question processing, (2) Candidate answer generation, (3) Candidate answer scoring, and (4) Confidence merging and ranking.

Question answering in deep learning era



$\text{BERT}_Q(q)$
[CLS] What does the zip in zip code stand for? [SEP]
$S_{retr}(0,q)$
$S_{retr}(1,q)$
$S_{retr}(2,q)$
$S_{retr}(...,q)$
$\text{BERT}_B(0)$
[CLS] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
$\text{BERT}_B(1)$
[CLS] ...group of zebras are referred to as a herd or dazzle... [SEP]
$\text{BERT}_B(2)$
[CLS] ...ZIPs for other operating systems may be preceded by... [SEP]
$\text{BERT}_B(...)$
...
Top K
Top K
$\text{BERT}_R(q,0)$
[CLS] What does the zip in zip code stand for? [SEP] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
$\text{BERT}_R(q,2)$
[CLS] What does the zip in zip code stand for? [SEP] ...ZIPs for other operating systems may be preceded by... [SEP]
MLP
MLP
MLP
MLP
MLP
MLP
$S_{read}(0, \text{"The term"}, q)$
$S_{read}(0, \text{"Zone Improvement Plan"}, q)$
$S_{read}(0, ..., q)$
$S_{read}(2, \text{"ZIPs"}, q)$
$S_{read}(2, \text{"operating systems"}, q)$
$S_{read}(2, ..., q)$

Image credit: (Lee et al., 2019)

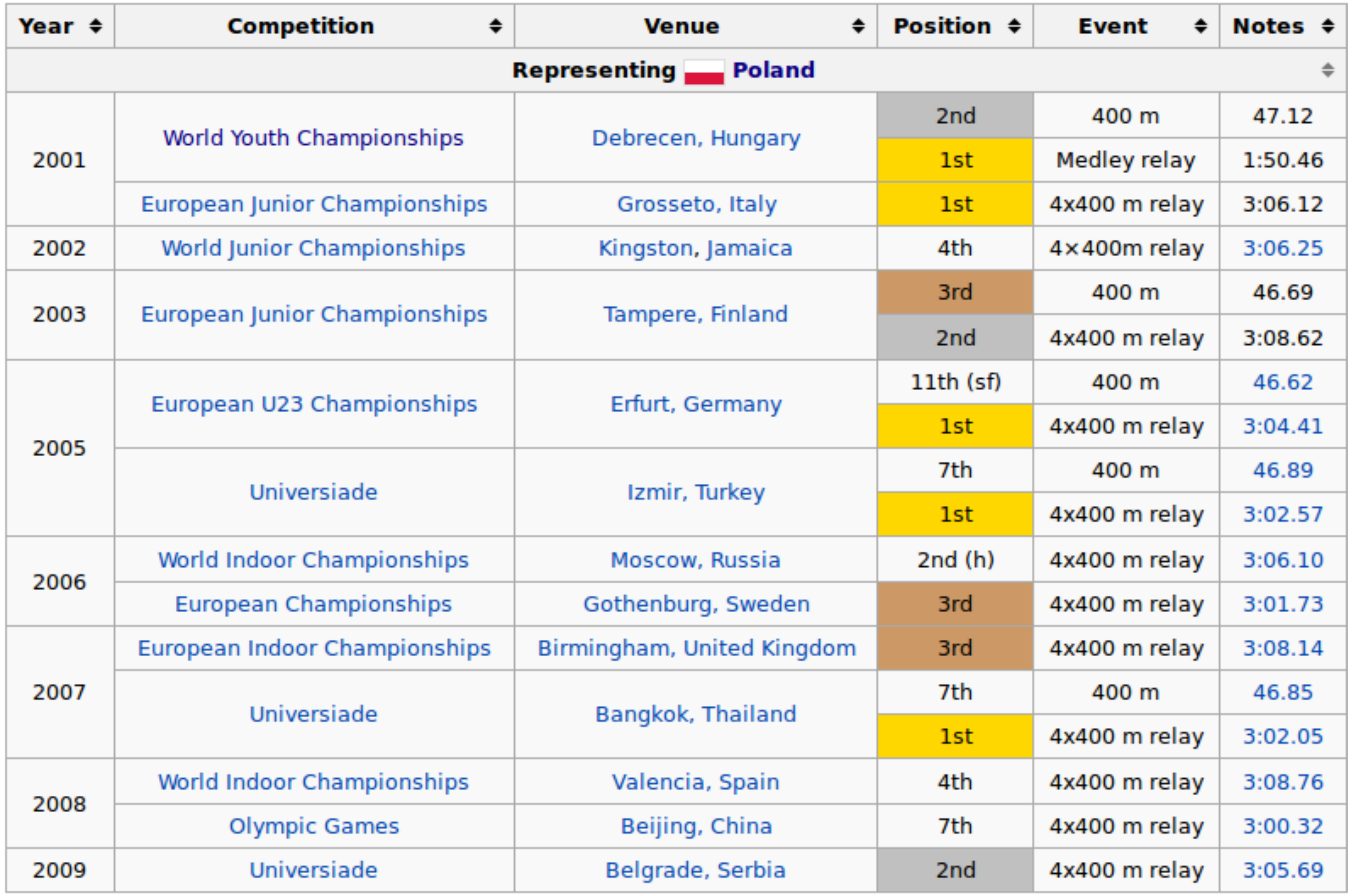
Almost all the state-of-the-art question answering systems are built on top of end-to-end training and pre-trained language models (e.g., BERT)!

Beyond textual QA problems

Today, we will mostly focus on how to answer questions based on **unstructured text**.

Question answering over tables

(Pasupat and Liang, 2015):
WikiTableQuestions



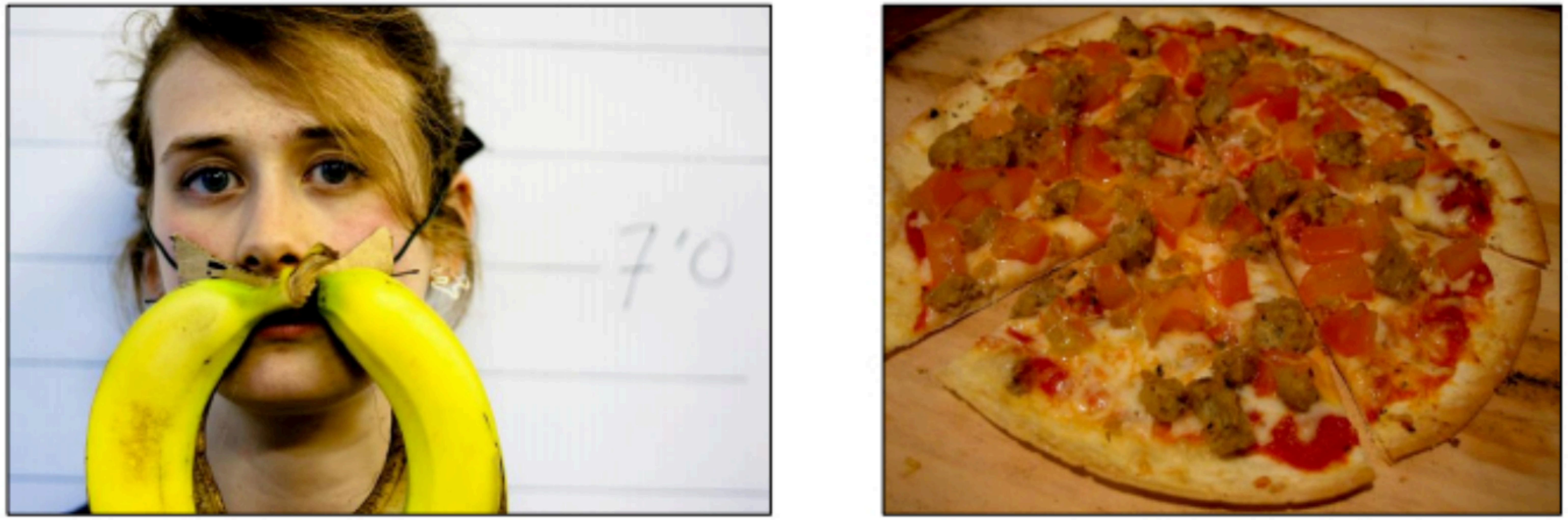
Year	Competition	Venue	Position	Event	Notes
Representing Poland					
2001	World Youth Championships	Debrecen, Hungary	2nd	400 m	47.12
			1st	Medley relay	1:50.46
	European Junior Championships	Grosseto, Italy	1st	4x400 m relay	3:06.12
2002	World Junior Championships	Kingston, Jamaica	4th	4×400m relay	3:06.25
2003	European Junior Championships	Tampere, Finland	3rd	400 m	46.69
			2nd	4x400 m relay	3:08.62
2005	European U23 Championships	Erfurt, Germany	11th (sf)	400 m	46.62
			1st	4x400 m relay	3:04.41
	Universiade	Izmir, Turkey	7th	400 m	46.89
			1st	4x400 m relay	3:02.57
2006	World Indoor Championships	Moscow, Russia	2nd (h)	4x400 m relay	3:06.10
	European Championships	Gothenburg, Sweden	3rd	4x400 m relay	3:01.73
2007	European Indoor Championships	Birmingham, United Kingdom	3rd	4x400 m relay	3:08.14
	Universiade	Bangkok, Thailand	7th	400 m	46.85
			1st	4x400 m relay	3:02.05
2008	World Indoor Championships	Valencia, Spain	4th	4x400 m relay	3:08.76
	Olympic Games	Beijing, China	7th	4x400 m relay	3:00.32
2009	Universiade	Belgrade, Serbia	2nd	4x400 m relay	3:05.69

In what city did Piotr's last 1st place finish occur?

Beyond textual QA problems

Today, we will mostly focus on how to answer questions based on **unstructured text**.

Visual QA



What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

(Antol et al., 2015): Visual Question Answering

2. Reading comprehension

Reading comprehension = comprehend a passage of text and answer questions about its content $(P, Q) \longrightarrow A$

> Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

Q: What language did Tesla study while in school?

A: German

2. Reading comprehension

Reading comprehension: building systems to comprehend a passage of text and answer questions about its content $(P, Q) \longrightarrow A$

Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

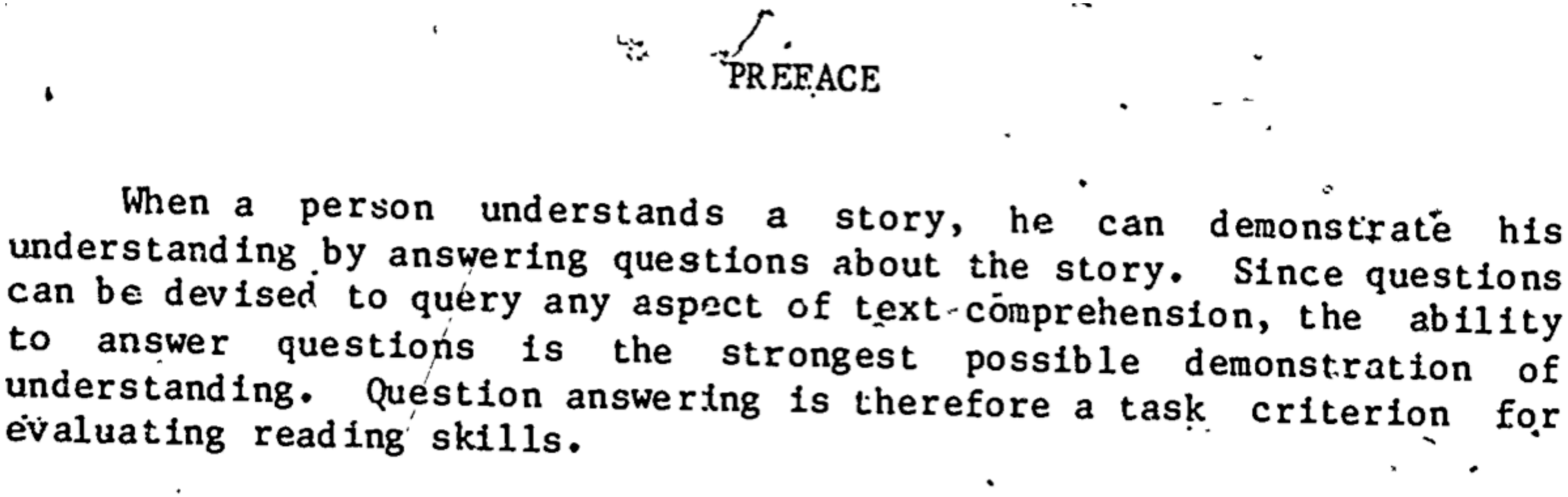
Q: Which linguistic minority is larger, Hindi or Malayalam?

A: Hindi

Why do we care about this problem?

- Useful for many practical applications
- Viewed as an important testbed for evaluating how well computer systems understand human language

Wendy Lehnert 1977. “The Process of Question Answering”



PREFACE

When a person understands a story, he can demonstrate his understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding. Question answering is therefore a task criterion for evaluating reading skills.

Why do we care about this problem?

- Useful for many practical applications
- Viewed as an important testbed for evaluating how well computer systems understand human language
- Many other NLP tasks can be reduced to a reading comprehension problem:

Information extraction

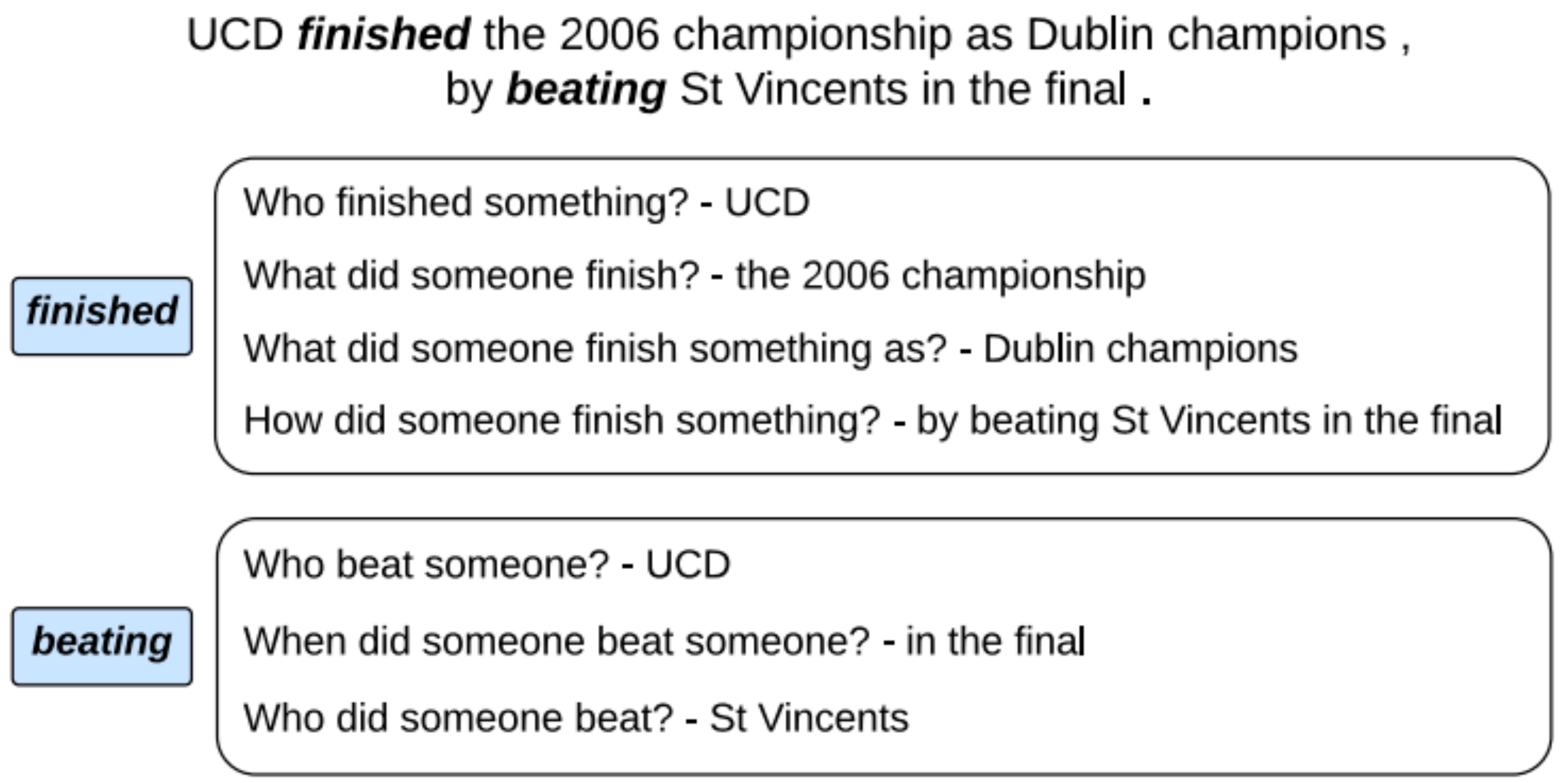
(Barack Obama, educated_at, ?)

Question: Where did Barack Obama graduate from?

Passage: Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.

(Levy et al., 2017)

Semantic role labeling



UCD ***finished*** the 2006 championship as Dublin champions ,
by ***beating*** St Vincents in the final .

Who finished something? - UCD
What did someone finish? - the 2006 championship
What did someone finish something as? - Dublin champions
How did someone finish something? - by beating St Vincents in the final

Who beat someone? - UCD
When did someone beat someone? - in the final
Who did someone beat? - St Vincents

(He et al., 2015)

Stanford question answering dataset (SQuAD)

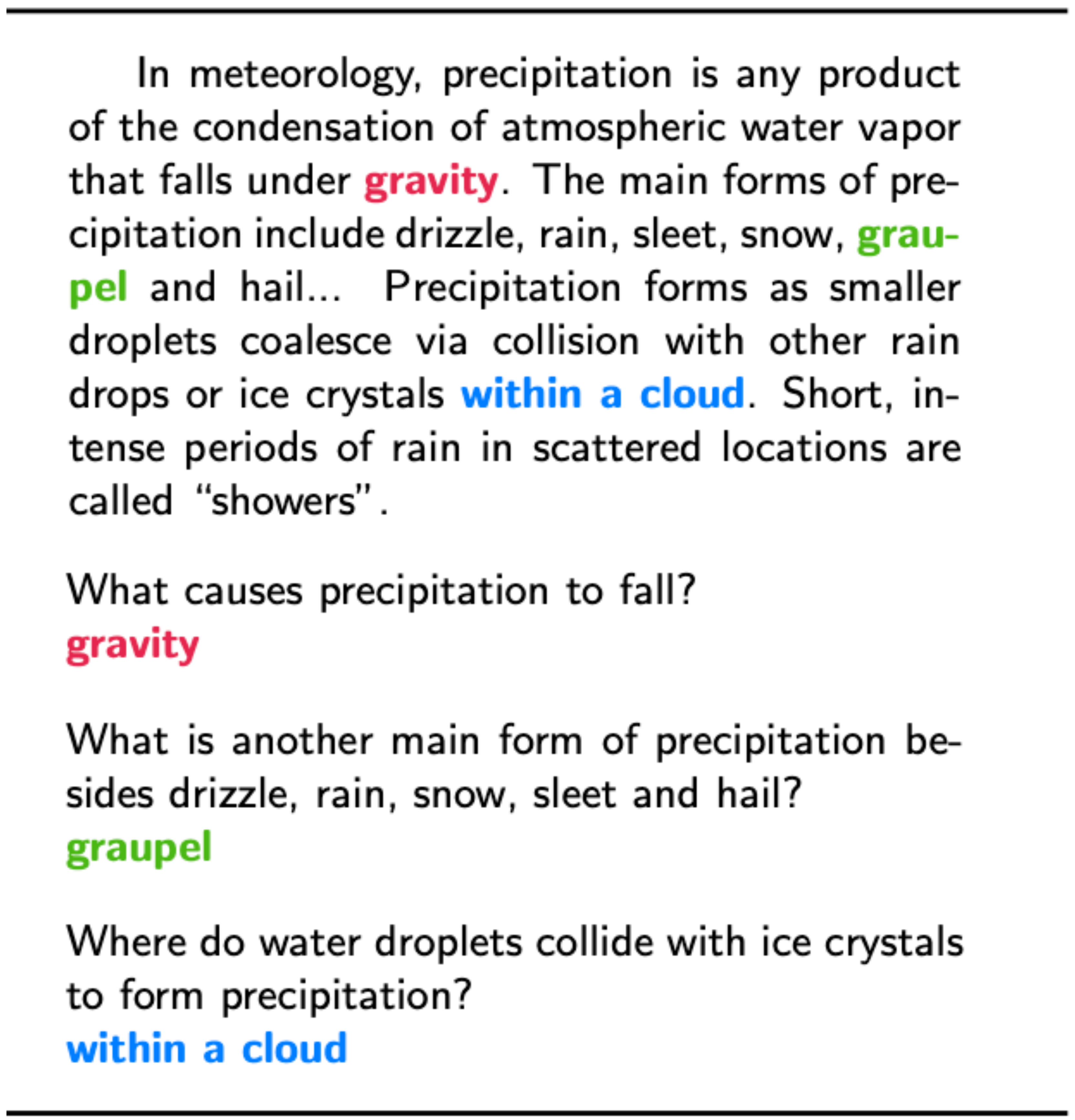
- 100k annotated (passage, question, answer) triples

  Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension!

- Passages are selected from English Wikipedia, usually 100~150 words.
- Questions are crowd-sourced.
- Each answer is a short segment of text (or span) in the passage.

  This is a limitation— not all the questions can be answered in this way!

- SQuAD still remains the most popular reading comprehension dataset; it is “almost solved” today and the state-of-the-art exceeds the estimated human performance.



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

(Rajpurkar et al., 2016): SQuAD: 100,000+ Questions for Machine Comprehension

Stanford question answering dataset (SQuAD)

- **Evaluation**: exact match (0 or 1) and F1 (partial credit).
- For development and testing sets, 3 gold answers are collected
- We compare the predicted answer to *each* gold answer and take max scores. Finally, we take the average of all the examples for both exact match and F1.
- Estimated human performance: EM = 82.3, F1 = 91.2

Q: What did Tesla do in December 1878?

A: {left Graz, left Graz, left Graz and severed all relations with his family}

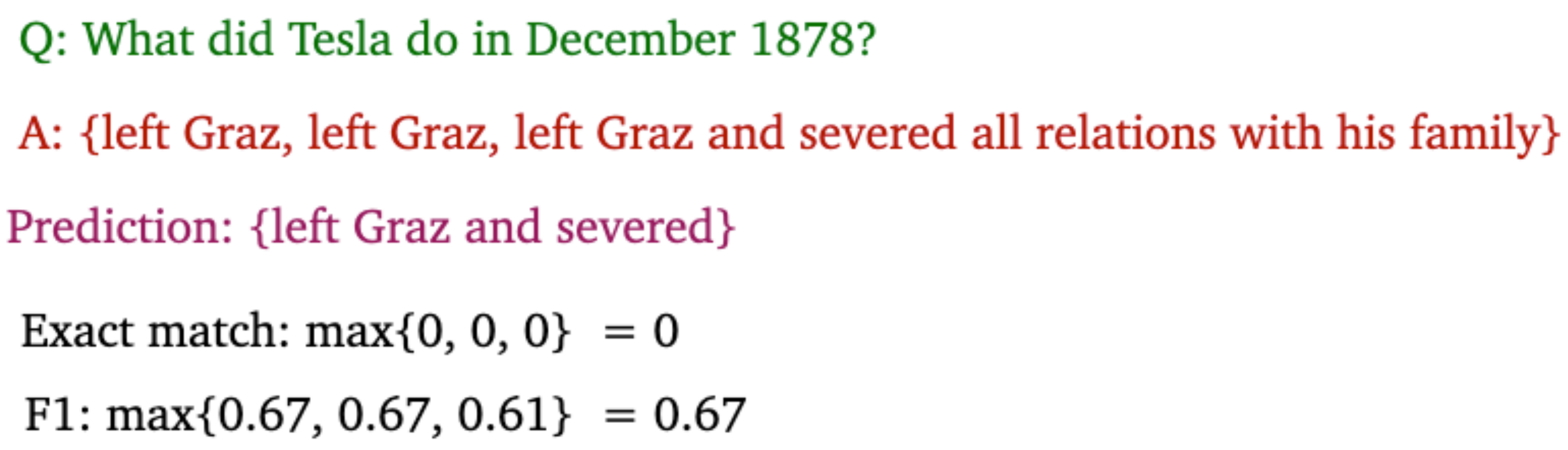
Prediction: {left Graz and severed}

Exact match: max{0, 0, 0} = 0

F1: max{0.67, 0.67, 0.61} = 0.67

Evaluation metrics for QA

- If the answer is a short segment of text (e.g., SQuAD),



Q: What did Tesla do in December 1878?

A: {left Graz, left Graz, left Graz and severed all relations with his family}

Prediction: {left Graz and severed}

Exact match: max{0, 0, 0} = 0

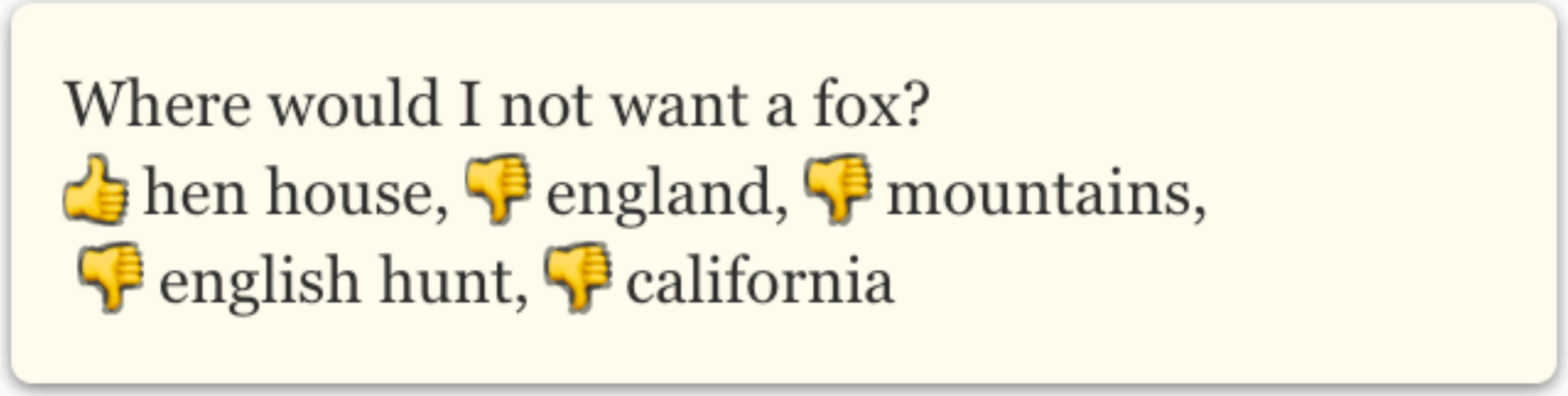
F1: max{0.67, 0.67, 0.61} = 0.67

- Many QA tasks are still hard to evaluate automatically

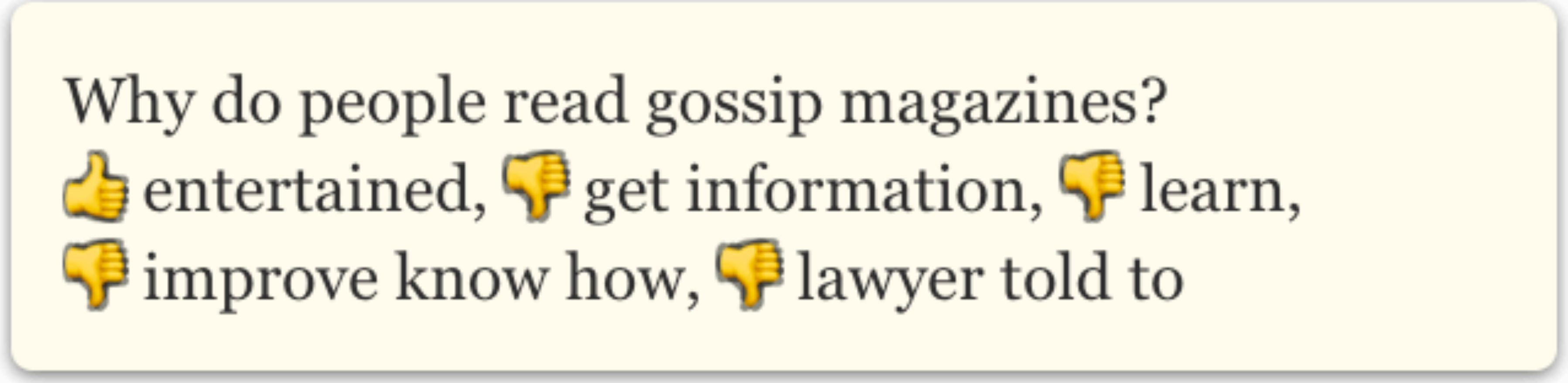
“Why is the sky blue?”

“How to make ramen eggs?”

- Multiple-choice QA tasks:



Where would I not want a fox?
👍 hen house, 👎 england, 👎 mountains,
👎 english hunt, 👎 california



Why do people read gossip magazines?
👍 entertained, 👎 get information, 👎 learn,
👎 improve know how, 👎 lawyer told to

https://www.tau-nlp.sites.tau.ac.il/commonsenseqa

Neural models for reading comprehension

- Problem formulation
    - Input: $C = (c_1, c_2, \ldots, c_N), Q = (q_1, q_2, \ldots, q_M), c_i, q_i \in V$
    - Output: $1 \leq$ start $\leq$ end $\leq N$

N∼100, M ∼15

answer is a span in the passage

- 2016: Stanford researchers built a logistic regression model (e.g., word matching, POS tags, parse trees)

F1 = 51.0%

- 2016-2018: A family of LSTM-based models with attention

Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), Match-LSTM (Wang et al., 2017), BiDFA (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017)..

F1: 60-80%

- 2019-current: Fine-tuning BERT-like models for reading comprehension

F1: 90%-95%

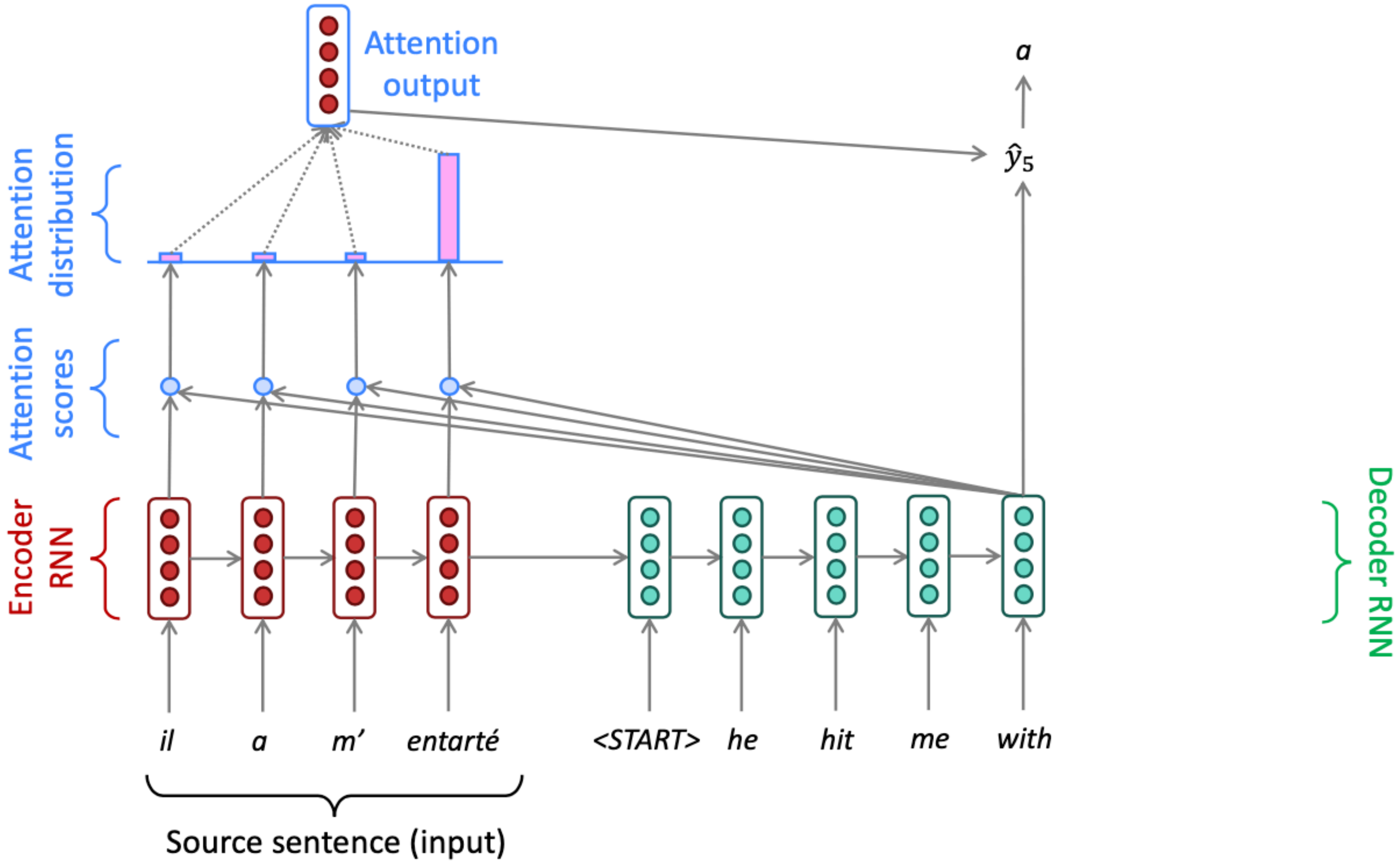
Seq2seq model with attention

- Instead of source and target sentences, we also have two sequences: passage and question

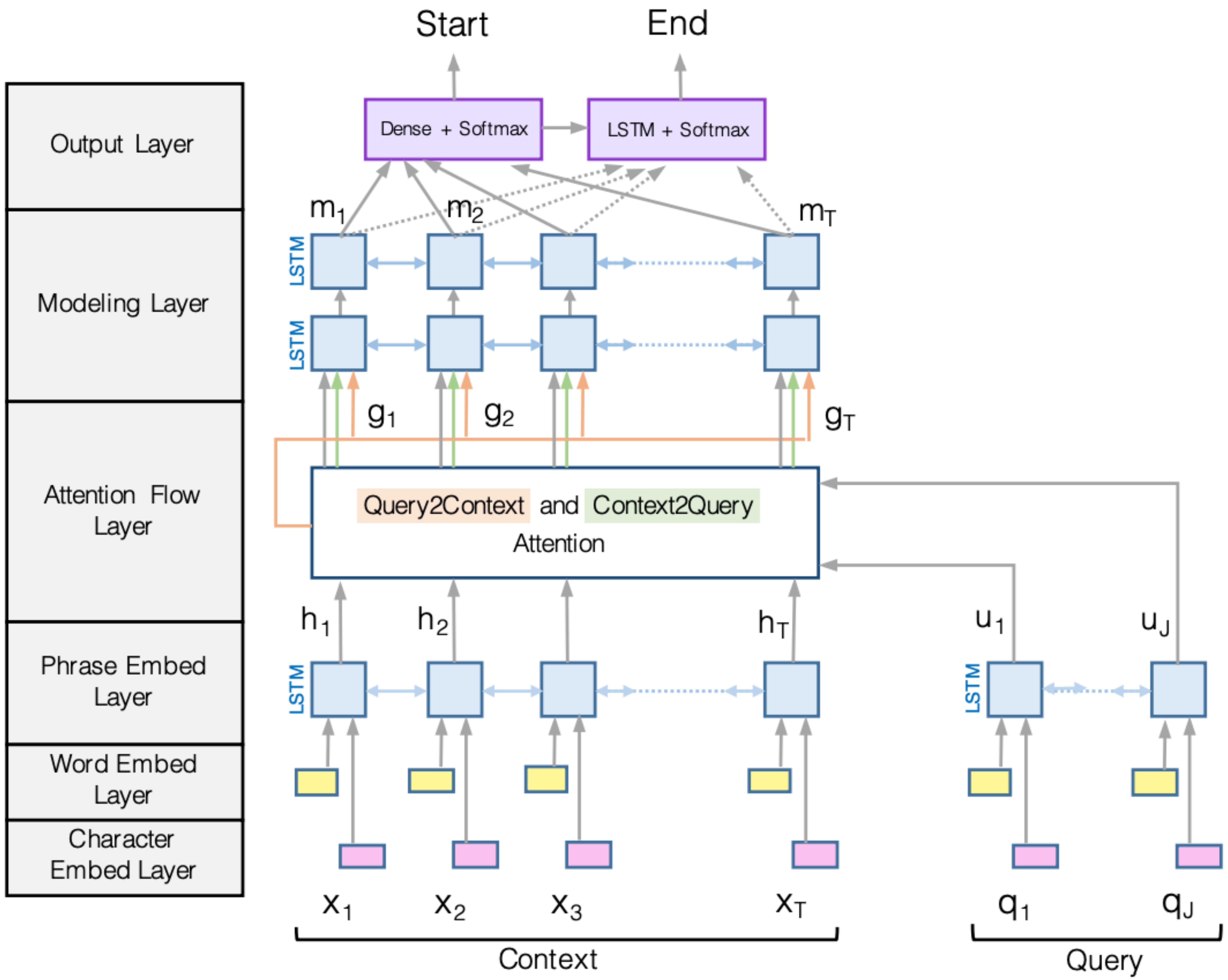
- We need to model which words in the passage are most relevant to the question (and which question words)

  Attention is the key ingredient here!

- We don't need an autoregressive decoder to generate the target sentence word-by-word. Instead, we just need to train two classifiers to predict the start and end positions of the answer!

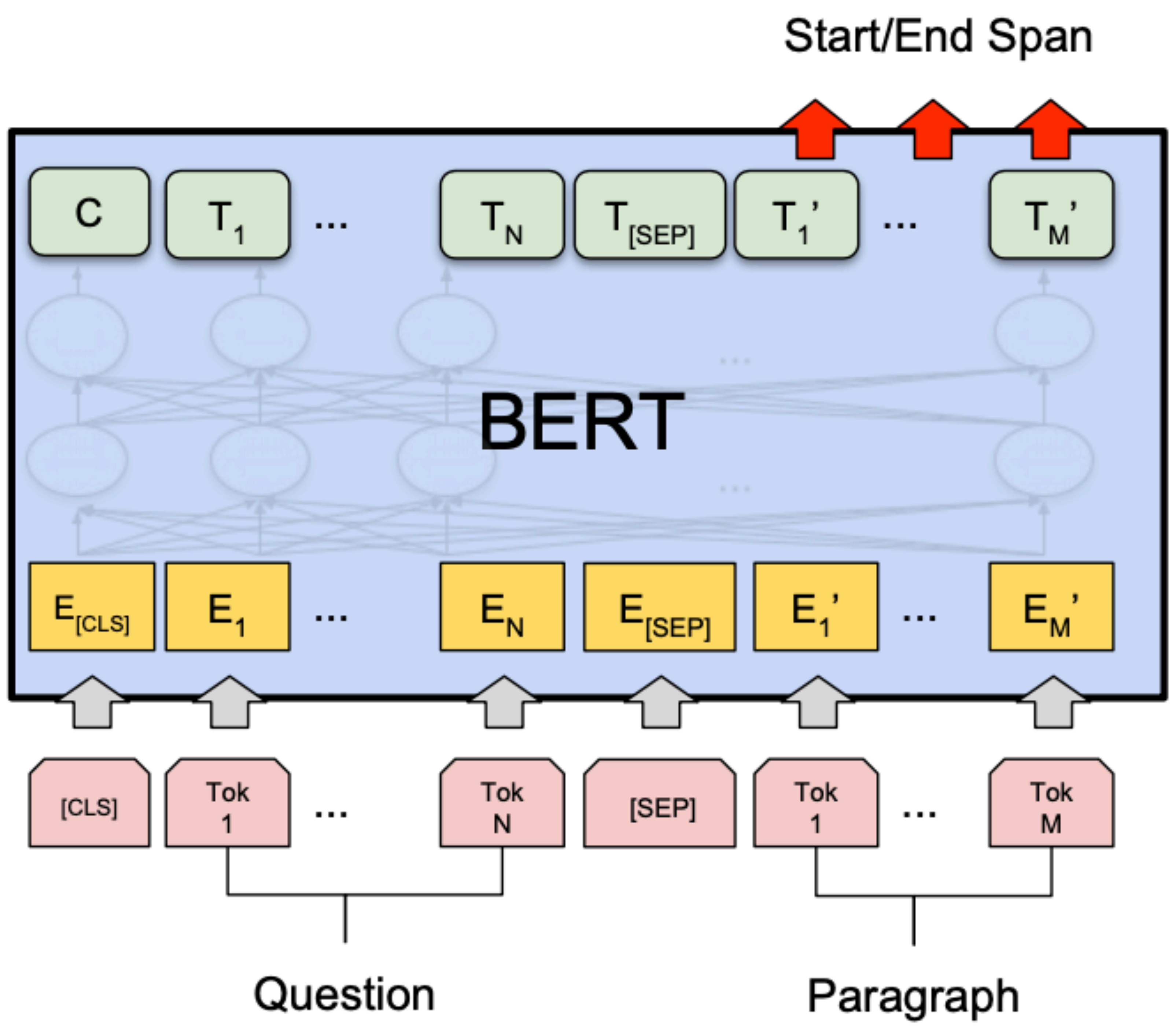


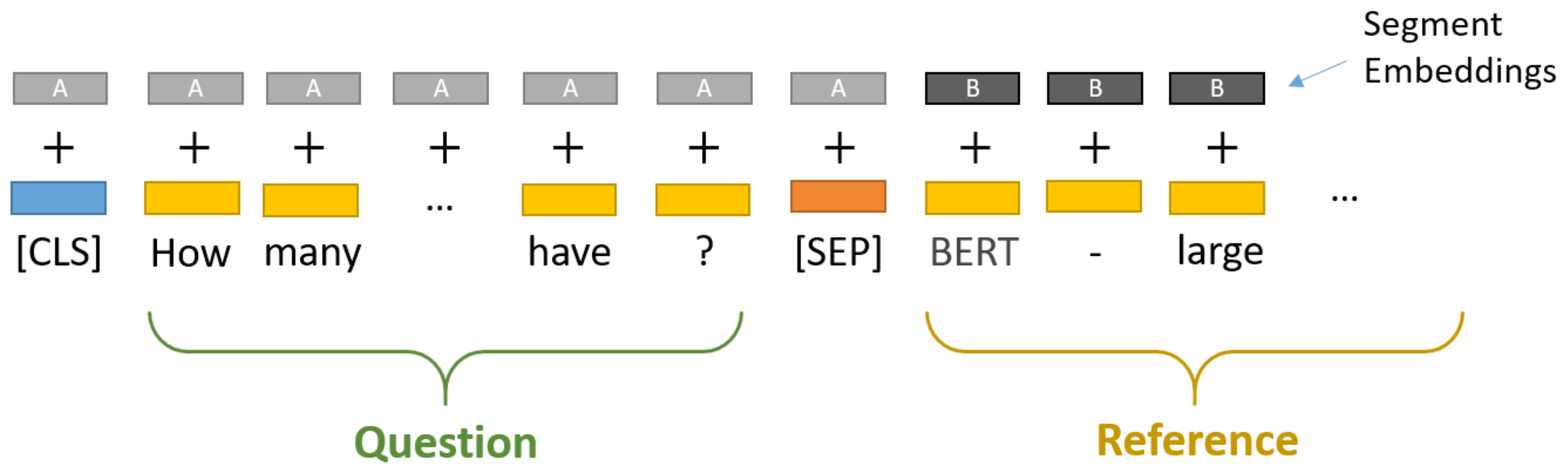
LSTM-based models with attention



(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension

BERT for reading comprehension





Question: How many parameters does BERT-large have?

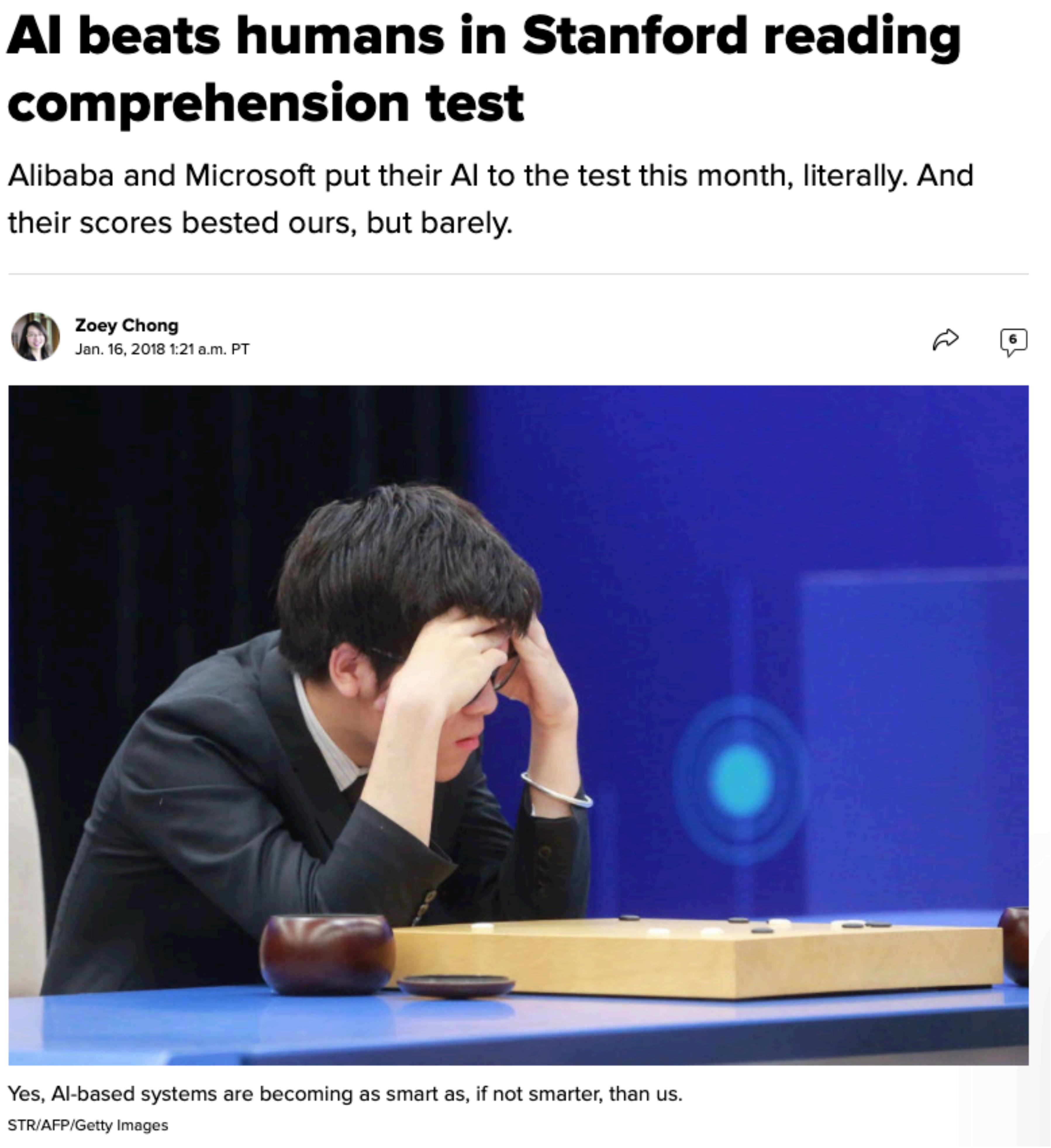
Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

Image credit: https://mccormickml.com/

BERT for reading comprehension

	F1
Human performance	91.2*
BiDAF	77.3
BERT-base	88.5
BERT-large	90.9
XLNet	94.5
RoBERTa	94.6
ALBERT	94.8

Is reading comprehension solved?

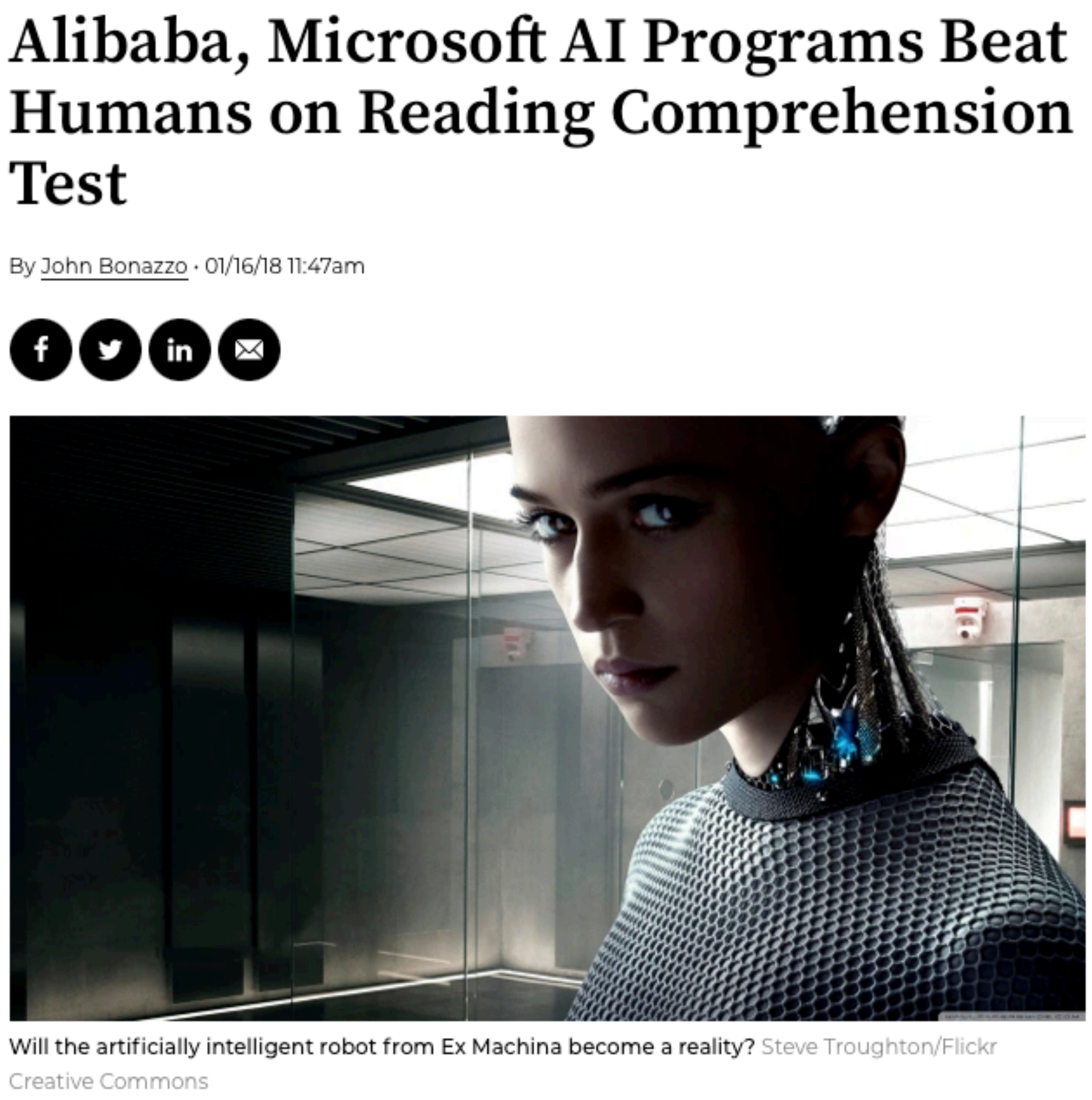


AI beats humans in Stanford reading comprehension test

Alibaba and Microsoft put their AI to the test this month, literally. And their scores bested ours, but barely.

Zoey Chong
Jan. 16, 2018 1:21 a.m. PT

Yes, AI-based systems are becoming as smart as, if not smarter, than us.
STR/AFP/Getty Images

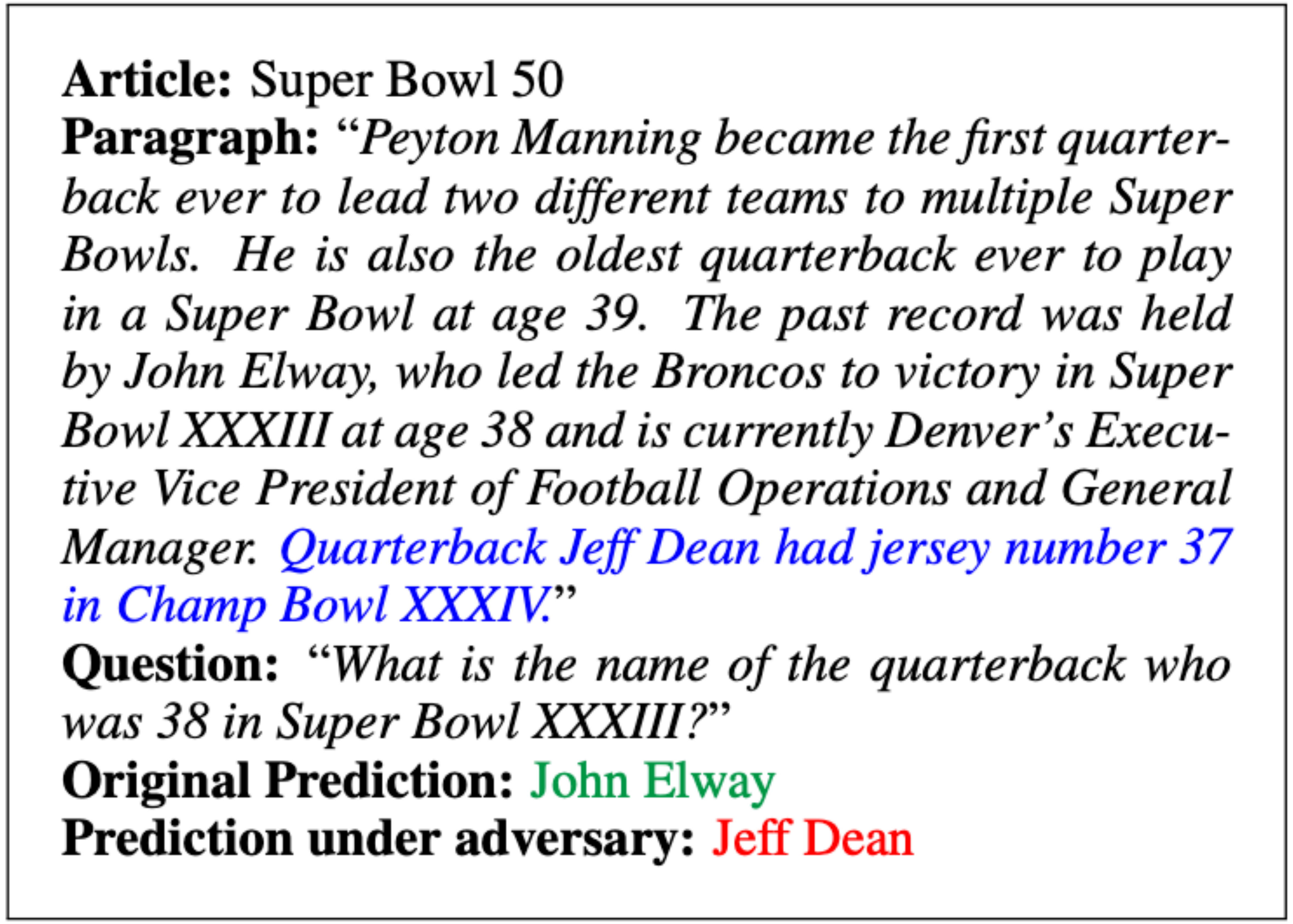


Alibaba, Microsoft AI Programs Beat Humans on Reading Comprehension Test

By John Bonazzo · 01/16/18 11:47am

Will the artificially intelligent robot from Ex Machina become a reality? Steve Troughton/Flickr Creative Commons

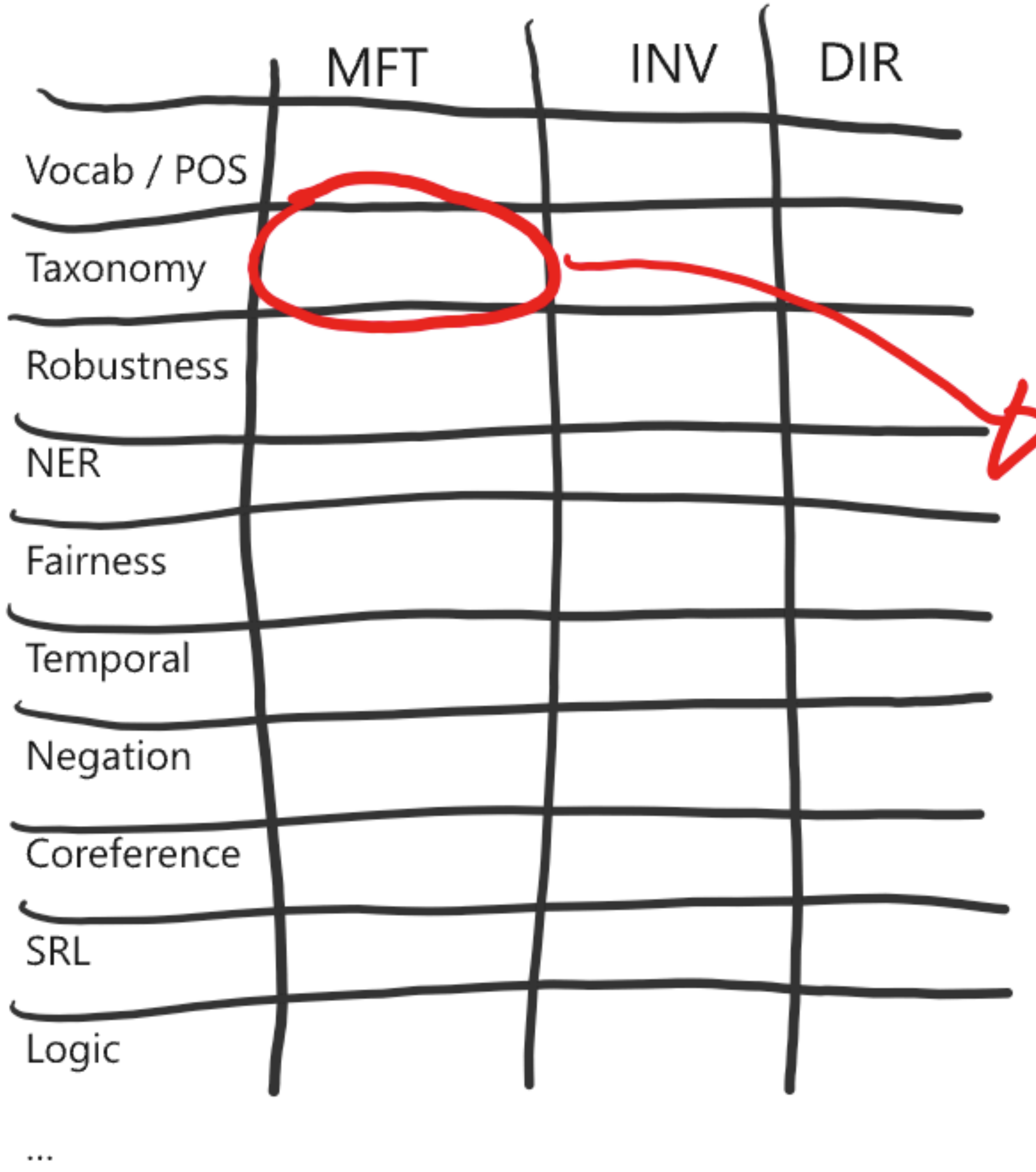
Is reading comprehension solved?



Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

(Jia and Liang, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems

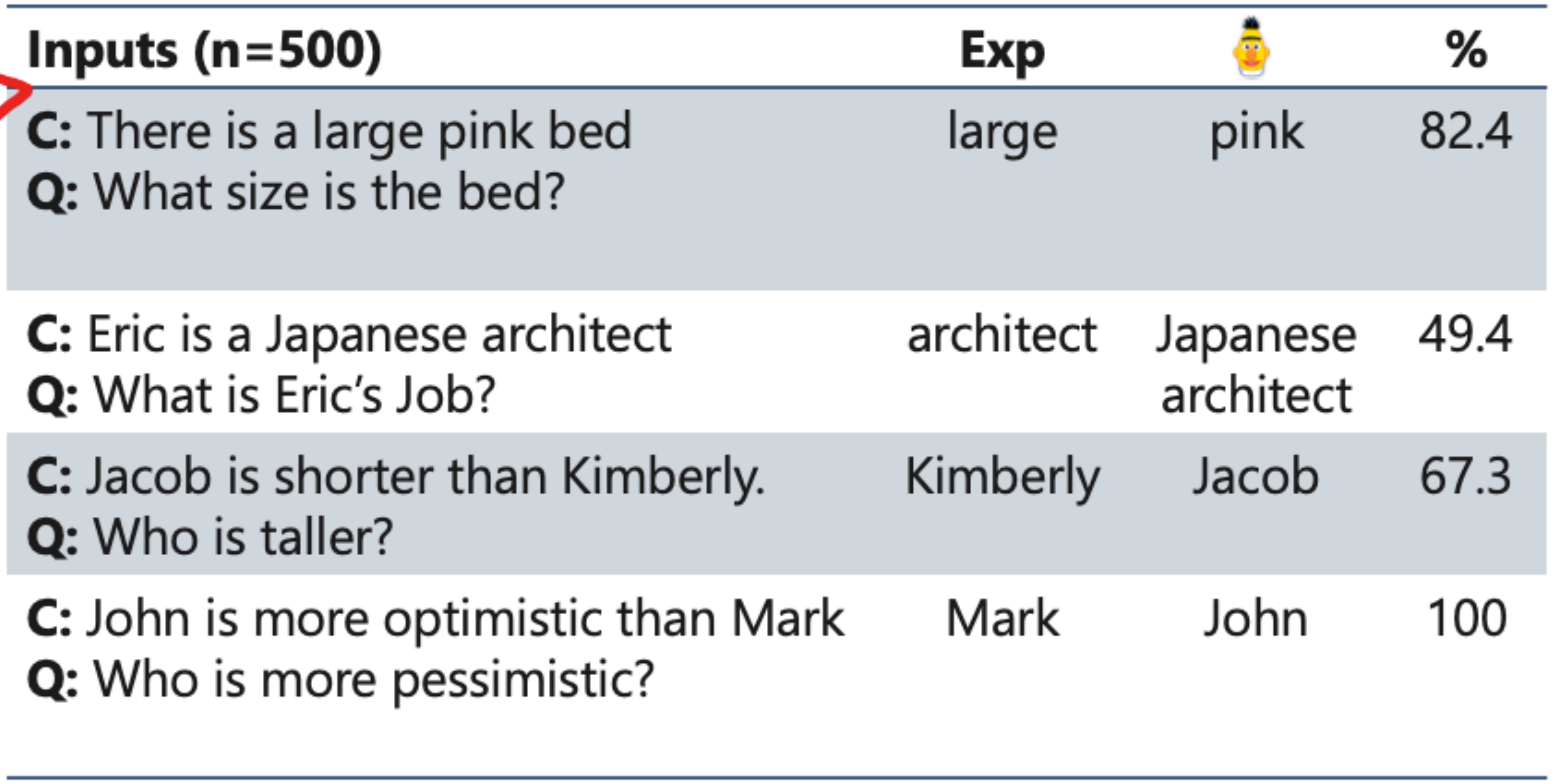
Is reading comprehension solved?



	MFT	INV	DIR
Vocab / POS			
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Negation			
Coreference			
SRL			
Logic			

...

Basic properties

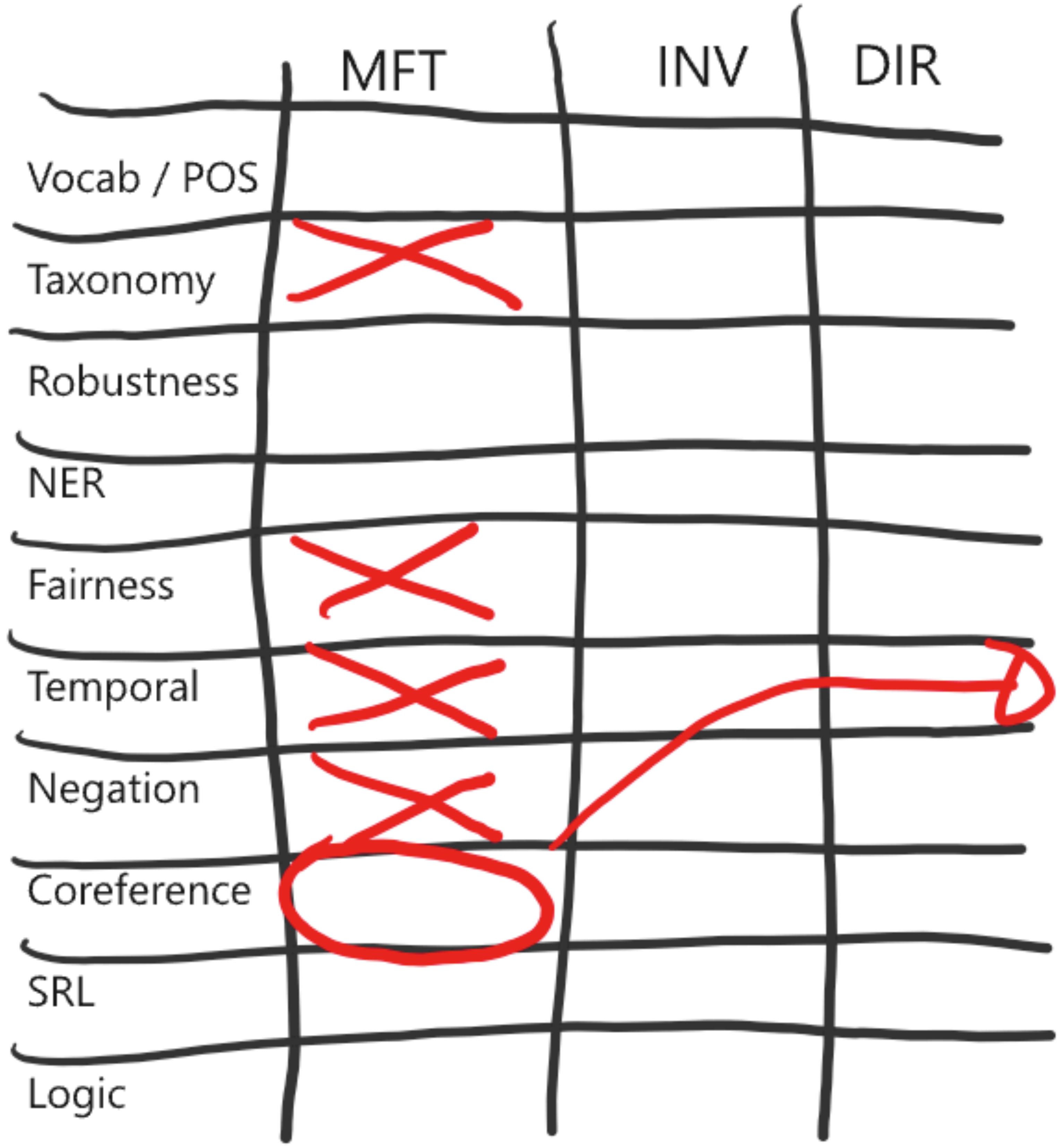


Inputs (n=500)	Exp		%
C: There is a large pink bed **Q:** What size is the bed?	large	pink	82.4
C: Eric is a Japanese architect **Q:** What is Eric's Job?	architect	Japanese architect	49.4
C: Jacob is shorter than Kimberly. **Q:** Who is taller?	Kimberly	Jacob	67.3
C: John is more optimistic than Mark **Q:** Who is more pessimistic?	Mark	John	100

Slide credit: Marco Tulio Ribeiro

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Is reading comprehension solved?



	MFT	INV	DIR
Vocab / POS			
Taxonomy	✗		
Robustness			
NER			
Fairness	✗		
Temporal	✗		
Negation	✗		
Coreference			
SRL			
Logic			
...			

Simple coreference

Inputs (n=500)	Exp		%
C: Melissa and Antonio are friends. He is a journalist, she is an adviser. **Q:** Who is a journalist?	Antonio	Melissa	100
C: Kimberly and Jennifer are friends. The former is a teacher. **Q:** Who is a teacher?	Kimberly	Jennifer	100

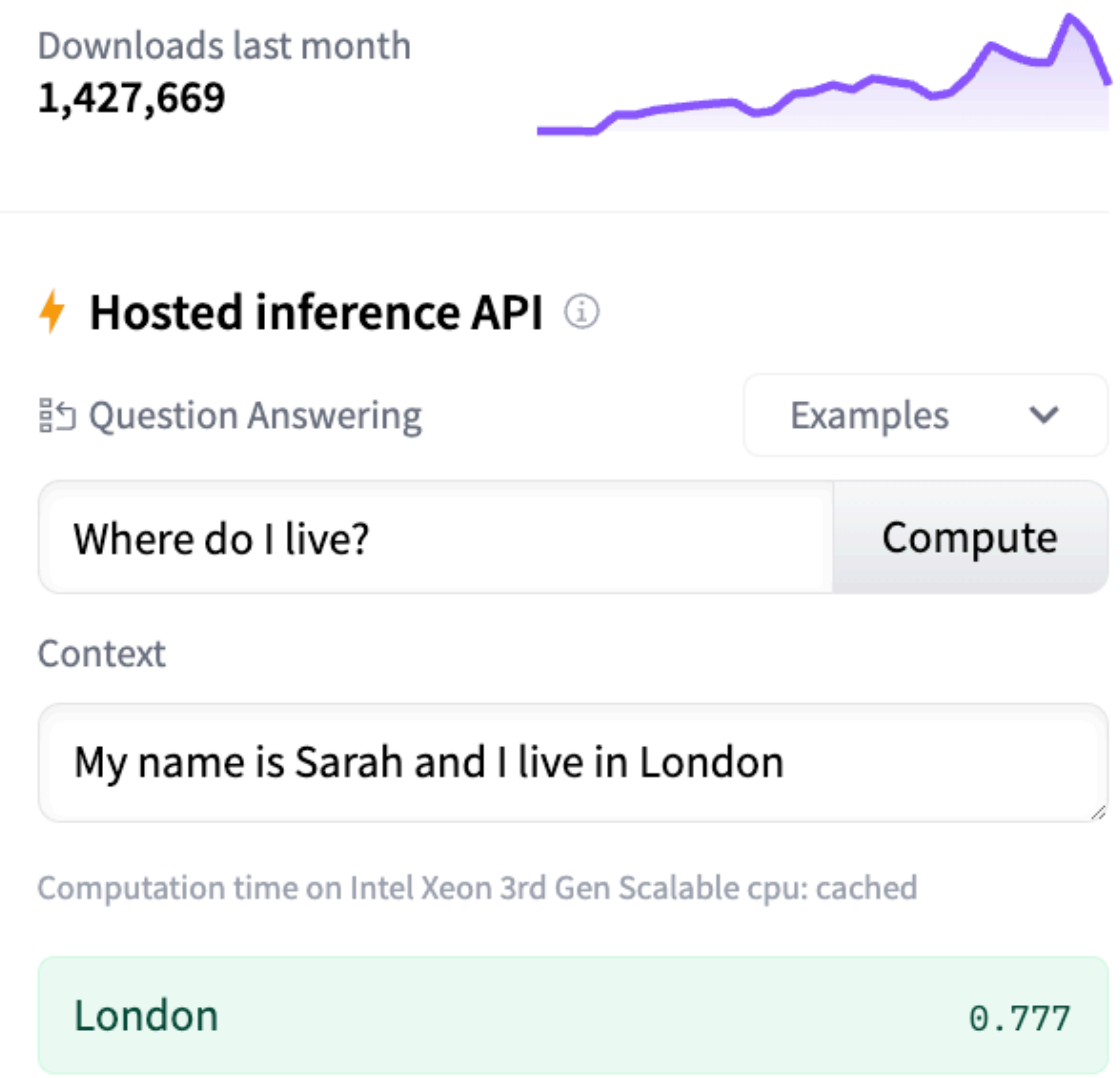
Slide credit: Marco Tulio Ribeiro

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Is reading comprehension solved?

Try out models yourself!

https://huggingface.co/deepset/roberta-base-squad2



Downloads last month
1,427,669

Hosted inference API

Question Answering

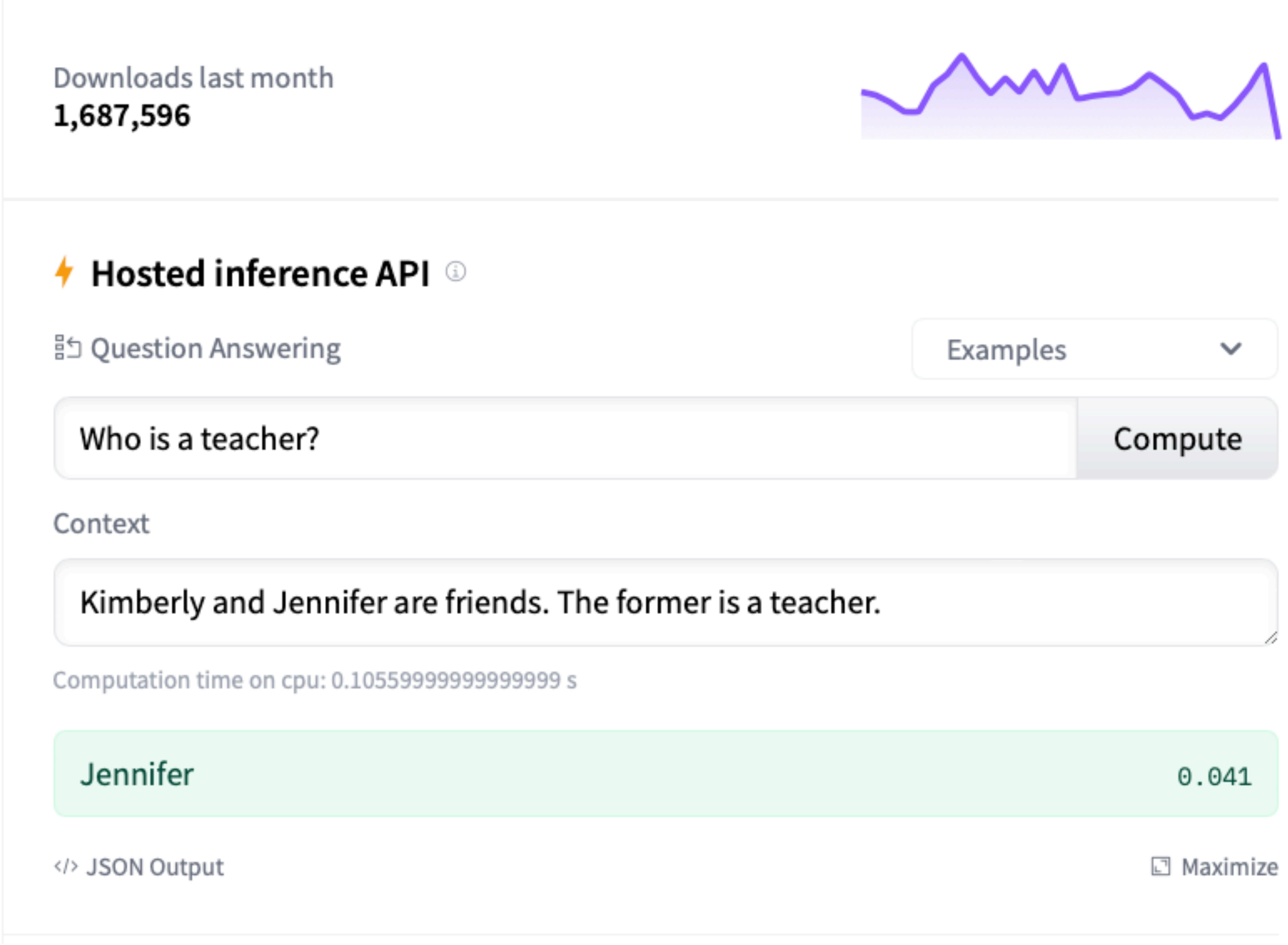
Where do I live?

Context

My name is Sarah and I live in London

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

London 0.777



Downloads last month
1,687,596

Hosted inference API

Question Answering

Who is a teacher?

Context

Kimberly and Jennifer are friends. The former is a teacher.

Computation time on cpu: 0.10559999999999999 s

Jennifer 0.041

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

More and more reading comprehension tasks

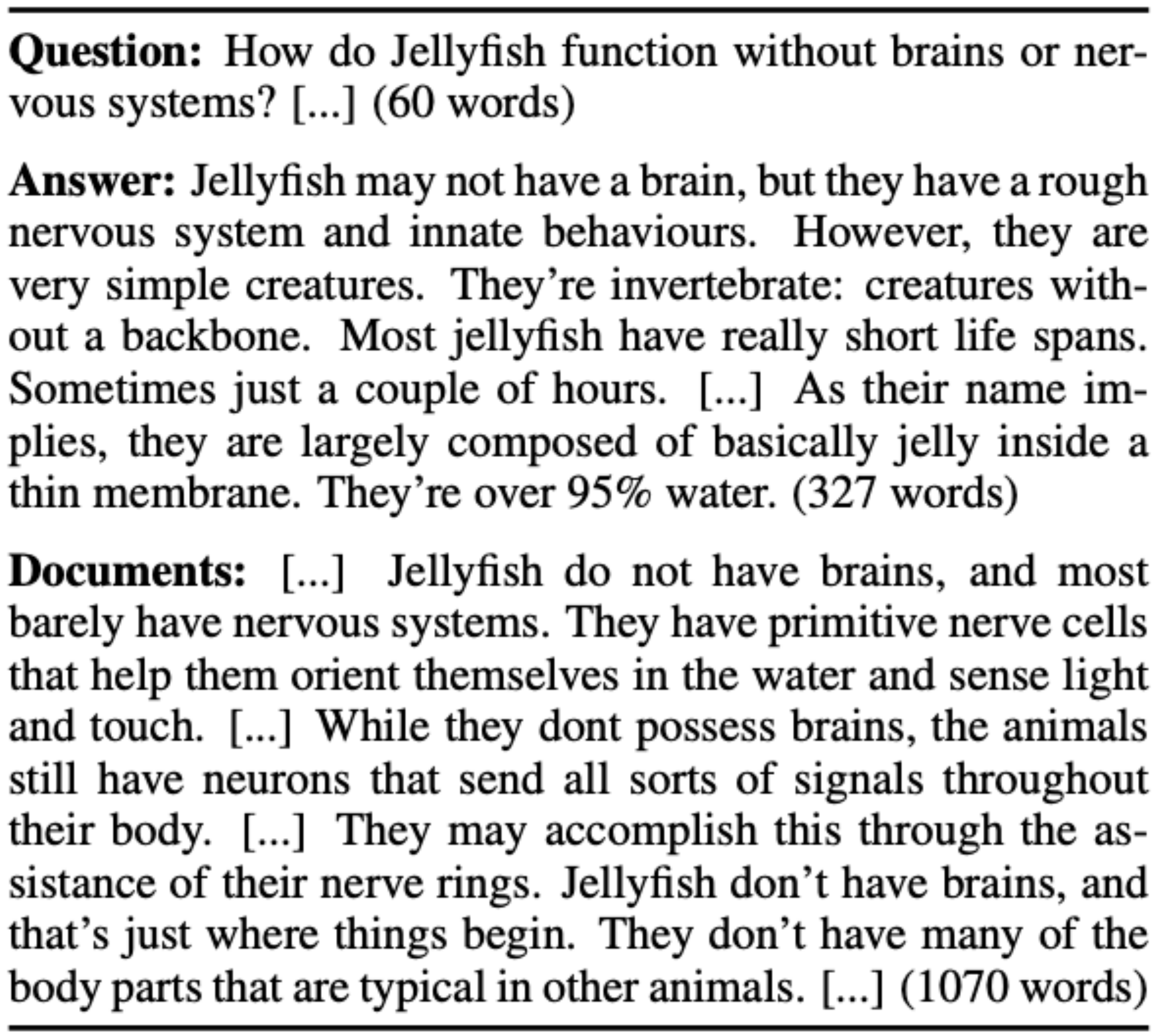
Questions that require discrete reasoning

Q: Where did Charles travel to first, Castile or Barcelona?

In 1517, the seventeen-year-old King sailed to Castile, where he was formally recognised as King of Castile. There, his Flemish court provoked much scandal, … In May 1518, Charles traveled to Barcelona in Aragon, where he would remain for nearly two years.

DROP (Dua et al., 2019)

Questions that require long answers



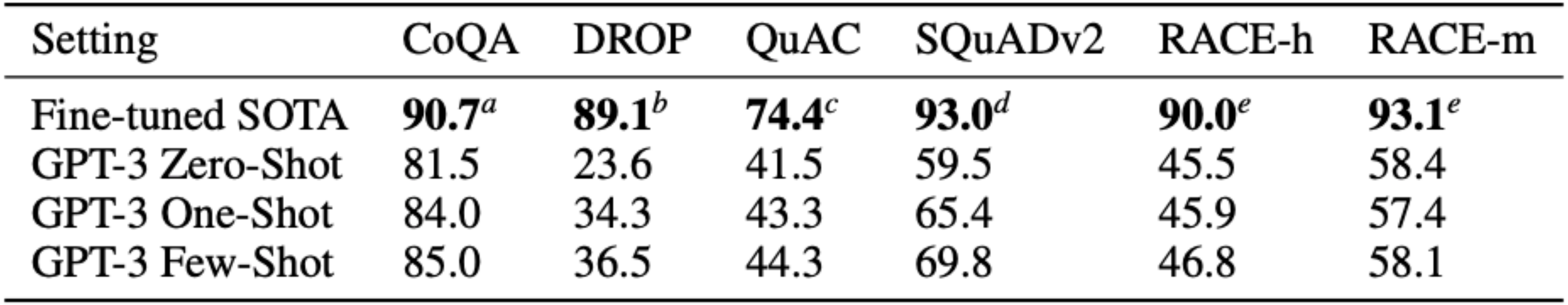
Question: How do Jellyfish function without brains or nervous systems? [...] (60 words)

Answer: Jellyfish may not have a brain, but they have a rough nervous system and innate behaviours. However, they are very simple creatures. They're invertebrate: creatures without a backbone. Most jellyfish have really short life spans. Sometimes just a couple of hours. [...] As their name implies, they are largely composed of basically jelly inside a thin membrane. They're over 95% water. (327 words)

Documents: [...] Jellyfish do not have brains, and most barely have nervous systems. They have primitive nerve cells that help them orient themselves in the water and sense light and touch. [...] While they dont possess brains, the animals still have neurons that send all sorts of signals throughout their body. [...] They may accomplish this through the assistance of their nerve rings. Jellyfish don't have brains, and that's just where things begin. They don't have many of the body parts that are typical in other animals. [...] (1070 words)

ELI5 (Fan et al., 2019)

How does GPT-3 perform on reading comprehension tasks?



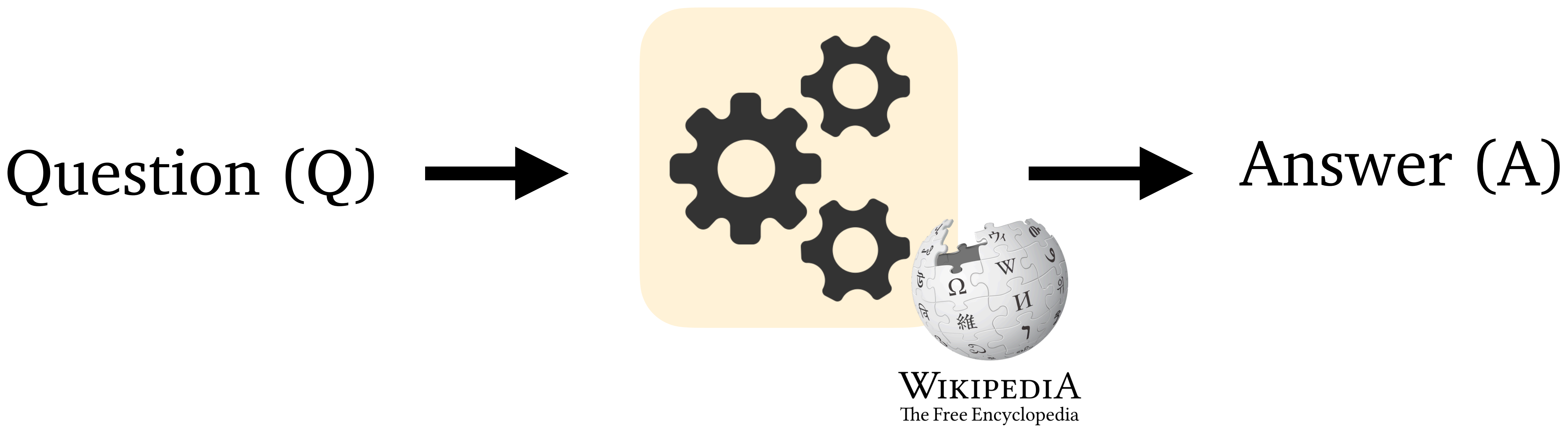
Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

The few-shot performance of GPT-3 falls far behind:

- Context length is long - you can only pack a few examples in the window
- No fine-tuning

(Brown et al., 2020): Language Models are Few-Shot Learners

3. Open-domain question answering



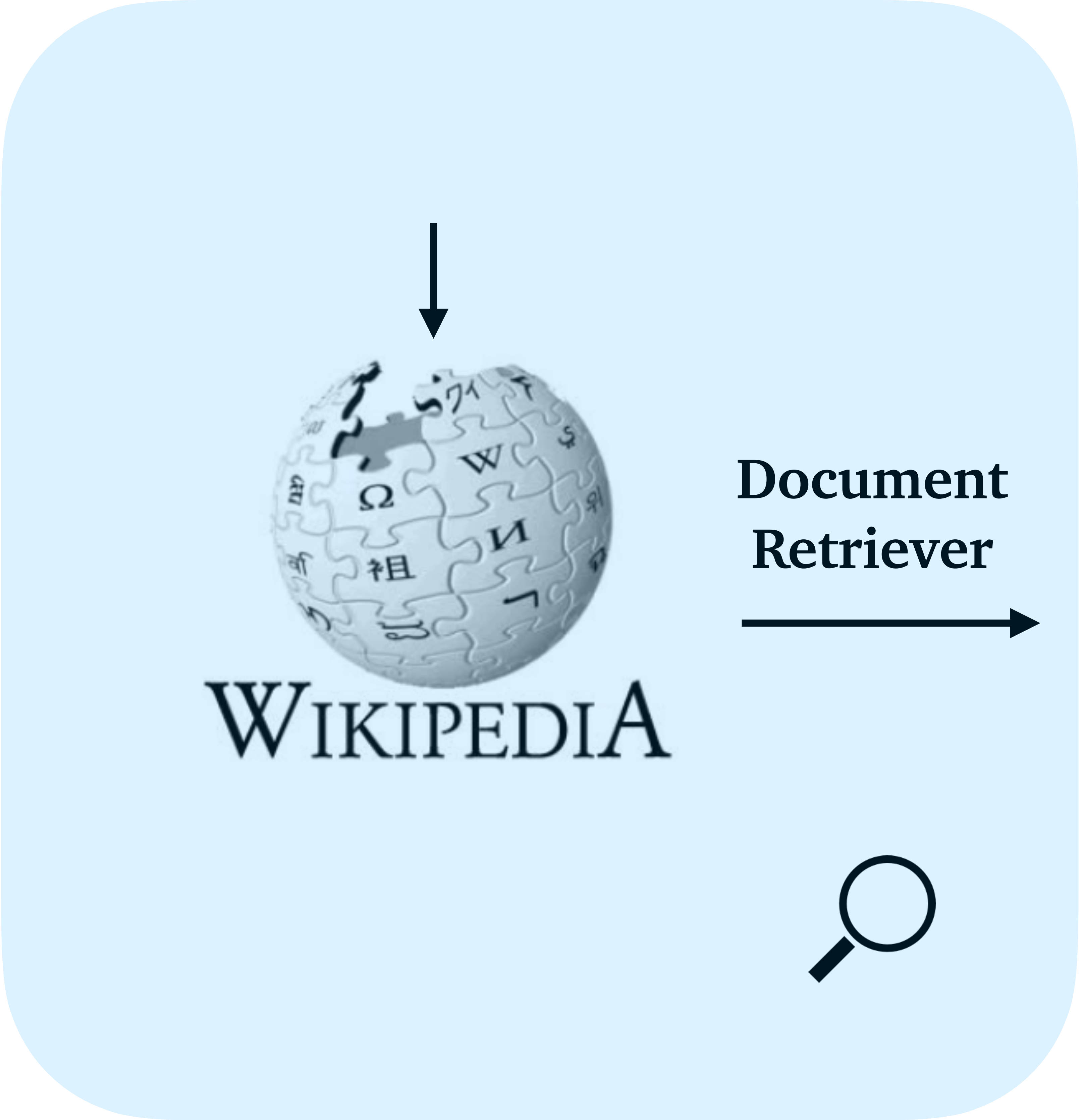
Question (Q)

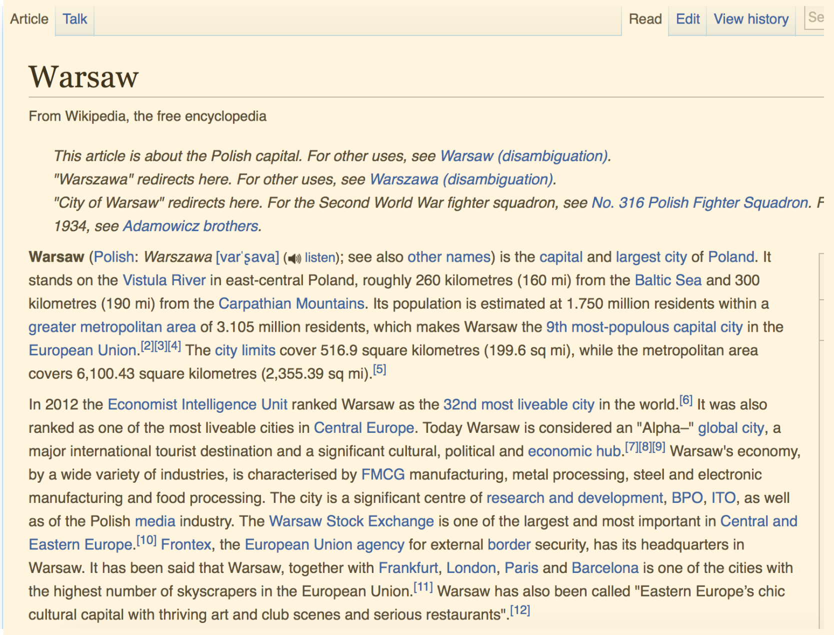
Answer (A)

- Different from reading comprehension, we don’t assume a given passage.
- Instead, we only have access to a large collection of documents (e.g., Wikipedia). We don’t know where the answer is located, and the goal is to return the answer for any open-domain questions.
- A much more challenging but practical problem!

*In contrast to **closed-domain** systems that deal with questions under a specific domain (medicine, technical support)..*

Retriever-reader framework





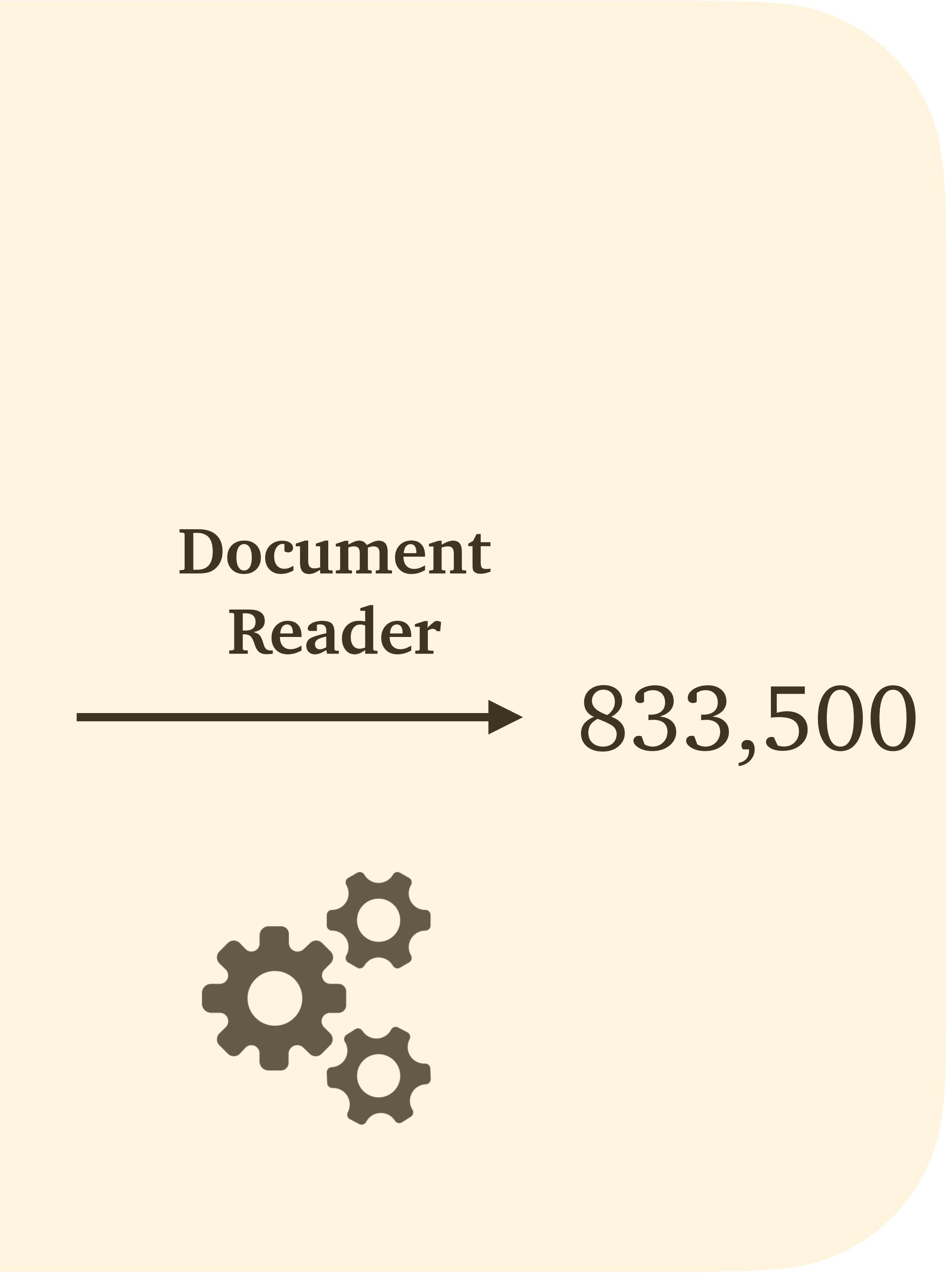
Warsaw

From Wikipedia, the free encyclopedia

This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F
1934, see Adamowicz brothers.

Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

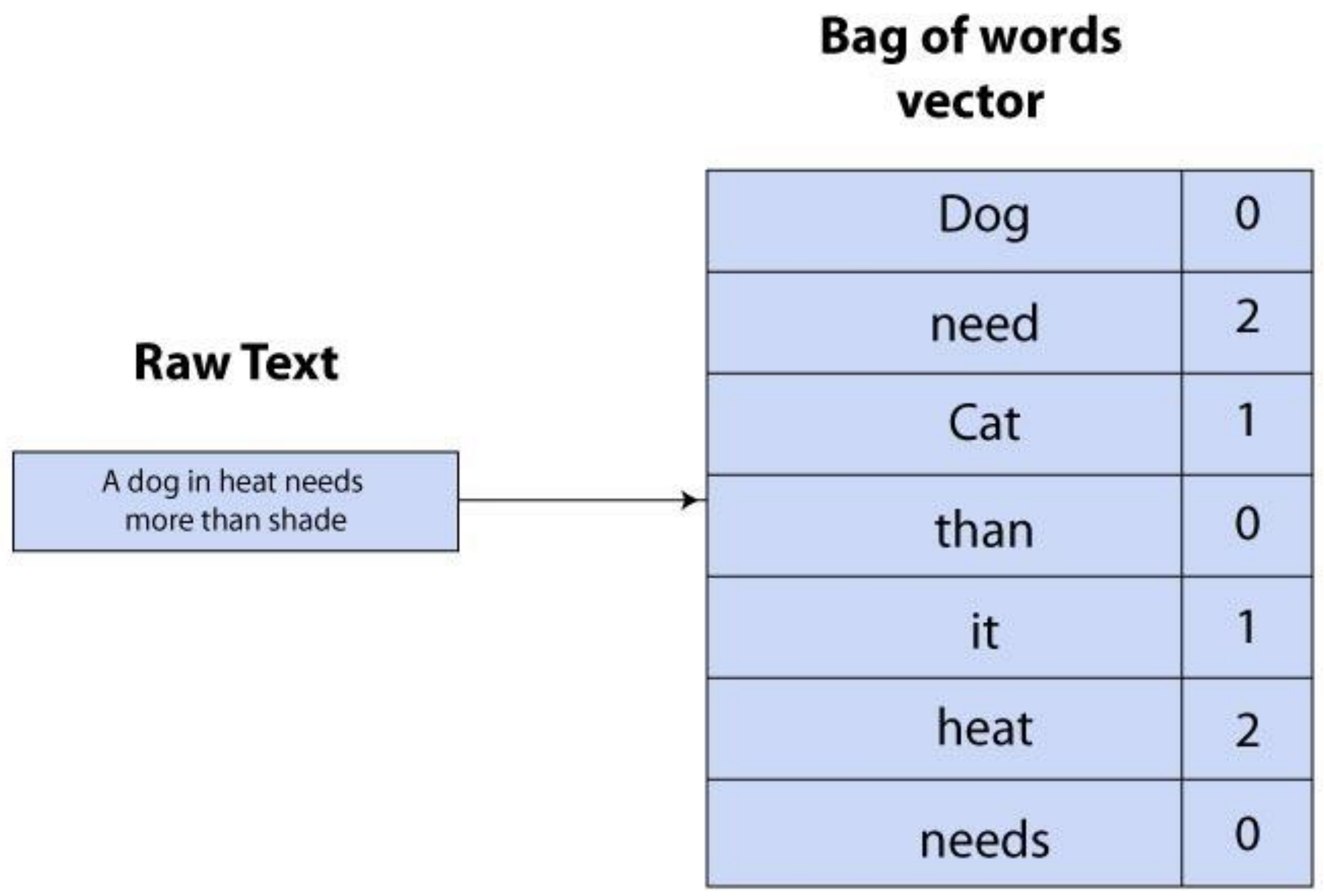


https://github.com/facebookresearch/DrQA

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Retriever-reader framework

- Input: a large collection of documents $\mathscr{D} = D_1, D_2, \ldots, D_N$ and Q
- Output: an answer string A
- Retriever: $f(\mathscr{D}, Q) \longrightarrow P_1, \ldots, P_K$ K is pre-defined (e.g., 100)
- Reader: $g(Q, \{P_1, \ldots, P_K\}) \longrightarrow A$ A reading comprehension problem!



Raw Text

A dog in heat needs more than shade

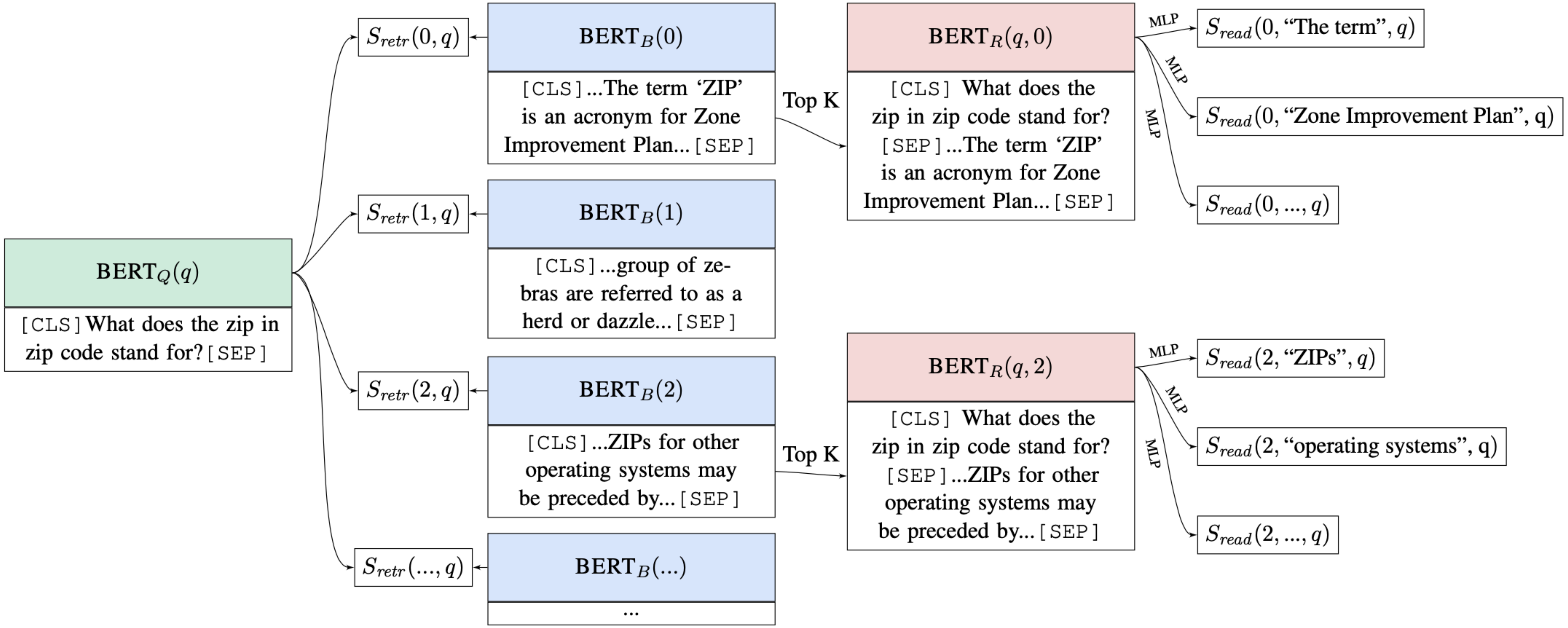
Bag of words vector

Dog	0
need	2
Cat	1
than	0
it	1
heat	2
needs	0

- In DrQA, the retriever is implemented as TF-IDF matching
- Classical information retrieval pipeline with word matching

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

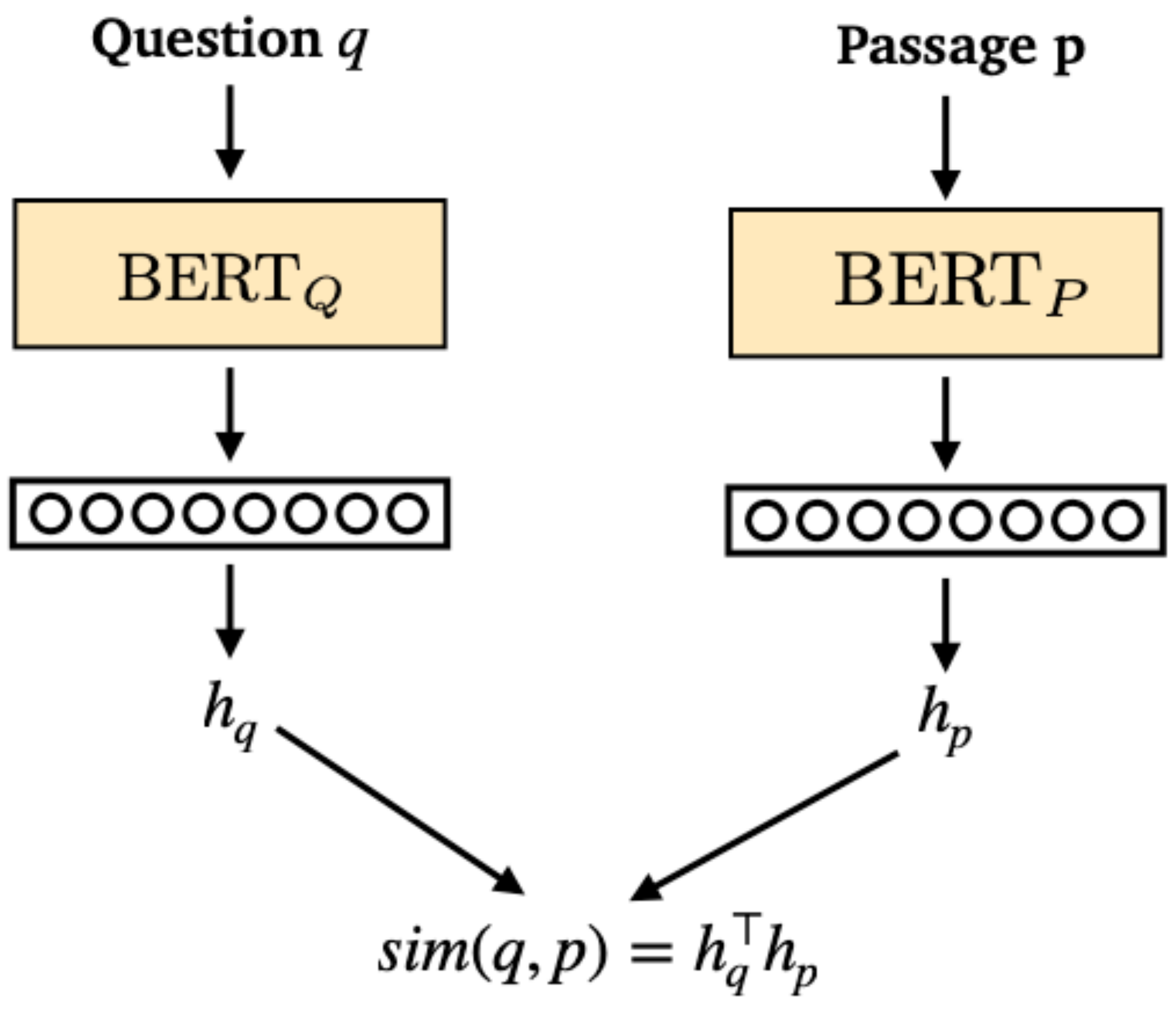
Joint training of retriever and reader

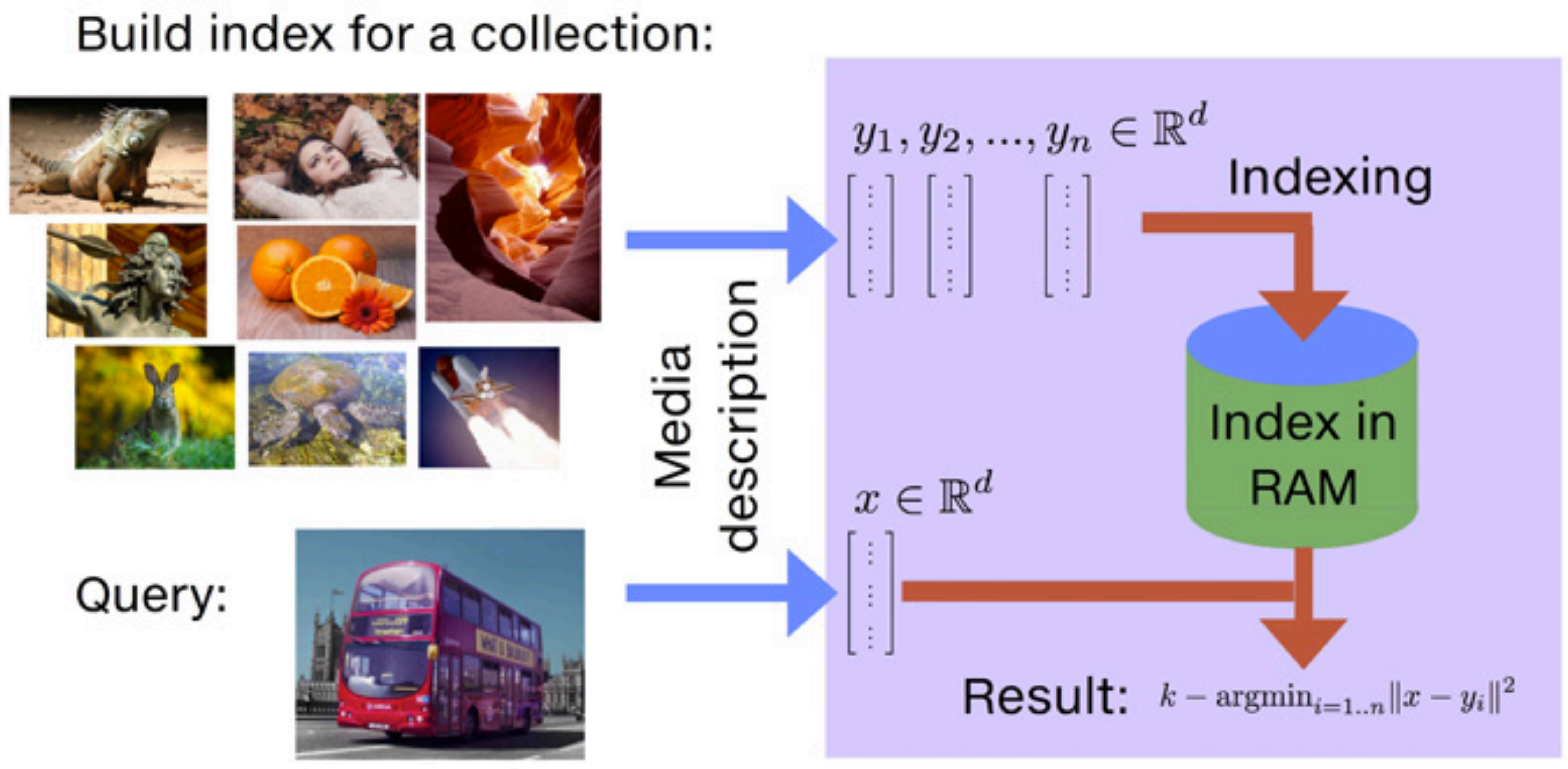


$\text{BERT}_Q(q)$
[CLS] What does the zip in zip code stand for? [SEP]
$S_{retr}(0,q)$
$S_{retr}(1,q)$
$S_{retr}(2,q)$
$S_{retr}(...,q)$
$\text{BERT}_B(0)$
[CLS] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
$\text{BERT}_B(1)$
[CLS] ...group of zebras are referred to as a herd or dazzle... [SEP]
$\text{BERT}_B(2)$
[CLS] ...ZIPs for other operating systems may be preceded by... [SEP]
$\text{BERT}_B(...)$
...
Top K
Top K
$\text{BERT}_R(q,0)$
[CLS] What does the zip in zip code stand for? [SEP] ...The term 'ZIP' is an acronym for Zone Improvement Plan... [SEP]
$\text{BERT}_R(q,2)$
[CLS] What does the zip in zip code stand for? [SEP] ...ZIPs for other operating systems may be preceded by... [SEP]
MLP
MLP
MLP
MLP
MLP
MLP
$S_{read}(0, \text{"The term"}, q)$
$S_{read}(0, \text{"Zone Improvement Plan"}, \text{q})$
$S_{read}(0, ..., q)$
$S_{read}(2, \text{"ZIPs"}, q)$
$S_{read}(2, \text{"operating systems"}, \text{q})$
$S_{read}(2, ..., q)$

Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering

Dense passage retrieval



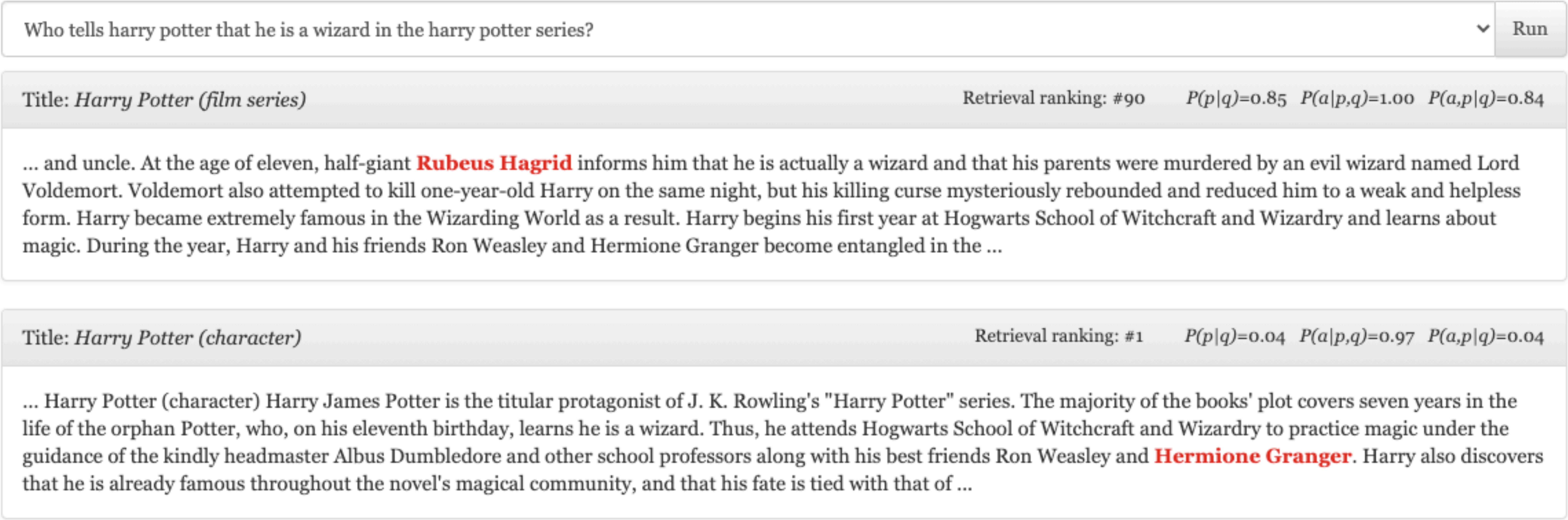

Build index for a collection:
$y_1, y_2, ..., y_n \in \mathbb{R}^d$
Indexing
Index in RAM
Media description
Query:
$x \in \mathbb{R}^d$
Result: $k - \text{argmin}_{i=1..n} \|x - y_i\|^2$

(Johnson et al., 2017)

Training a retriever using **1000 question-answer pairs** beat BM25 (on Natural Questions)!

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Dense passage retrieval



Who tells harry potter that he is a wizard in the harry potter series?

Title: *Harry Potter (film series)* Retrieval ranking: #90 $P(p|q)$=0.85 $P(a|p,q)$=1.00 $P(a,p|q)$=0.84

... and uncle. At the age of eleven, half-giant **Rubeus Hagrid** informs him that he is actually a wizard and that his parents were murdered by an evil wizard named Lord Voldemort. Voldemort also attempted to kill one-year-old Harry on the same night, but his killing curse mysteriously rebounded and reduced him to a weak and helpless form. Harry became extremely famous in the Wizarding World as a result. Harry begins his first year at Hogwarts School of Witchcraft and Wizardry and learns about magic. During the year, Harry and his friends Ron Weasley and Hermione Granger become entangled in the ...

Title: *Harry Potter (character)* Retrieval ranking: #1 $P(p|q)$=0.04 $P(a|p,q)$=0.97 $P(a,p|q)$=0.04

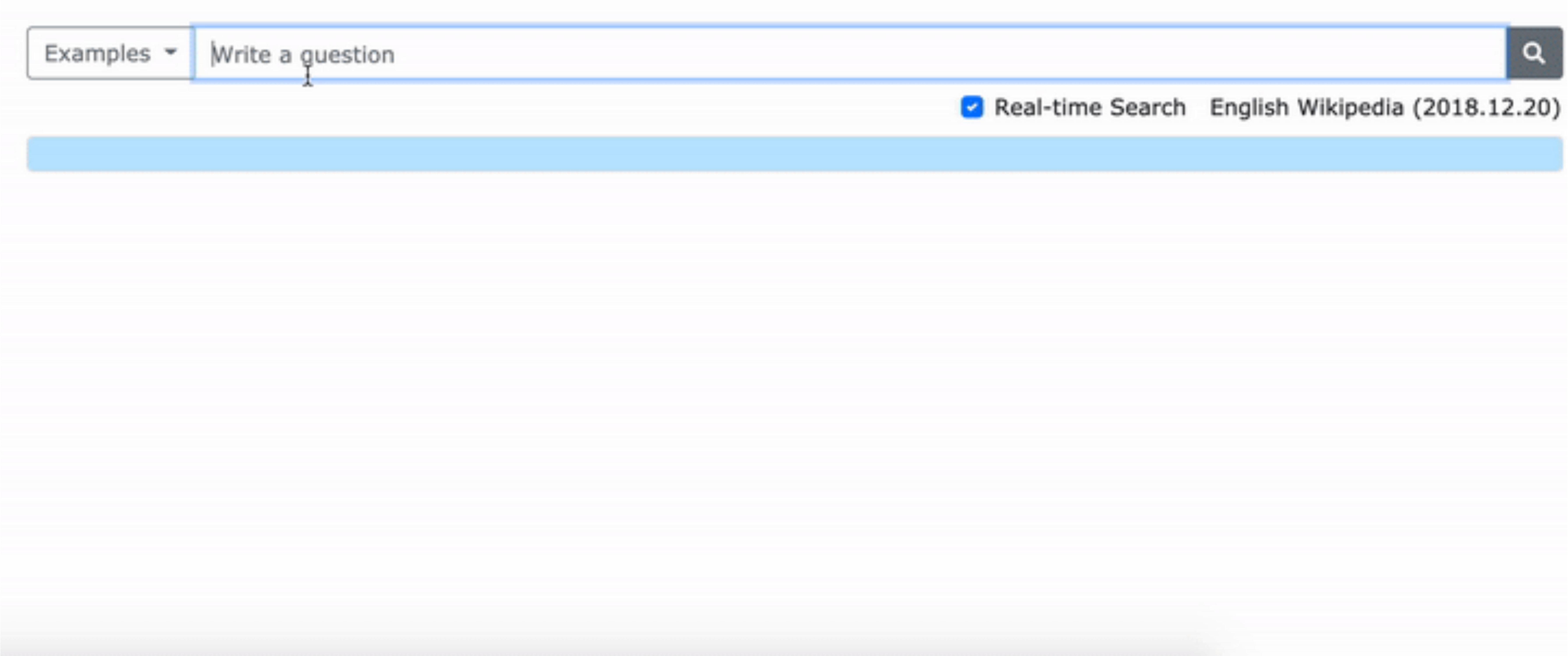
... Harry Potter (character) Harry James Potter is the titular protagonist of J. K. Rowling's "Harry Potter" series. The majority of the books' plot covers seven years in the life of the orphan Potter, who, on his eleventh birthday, learns he is a wizard. Thus, he attends Hogwarts School of Witchcraft and Wizardry to practice magic under the guidance of the kindly headmaster Albus Dumbledore and other school professors along with his best friends Ron Weasley and **Hermione Granger**. Harry also discovers that he is already famous throughout the novel's magical community, and that his fate is tied with that of ...

http://qa.cs.washington.edu:2020/

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

DensePhrases

You can retrieve answers from 60-billion phrases directly!



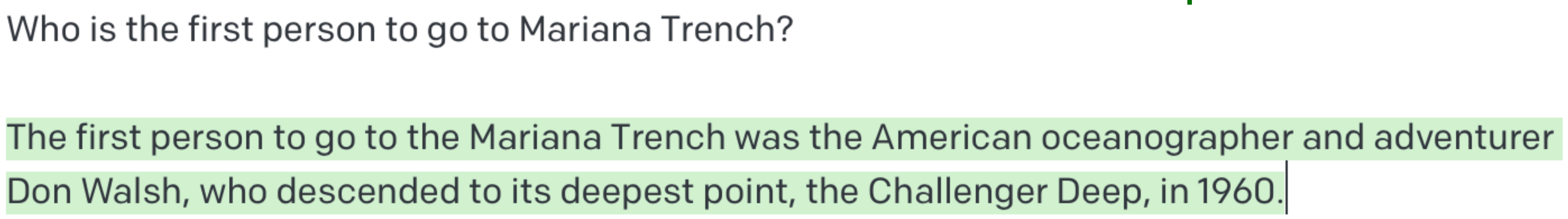
☑ Real-time Search English Wikipedia (2018.12.20)

https://github.com/princeton-nlp/DensePhrases

Lee et al., 2021. Learning Dense Representations of Phrases at Scale

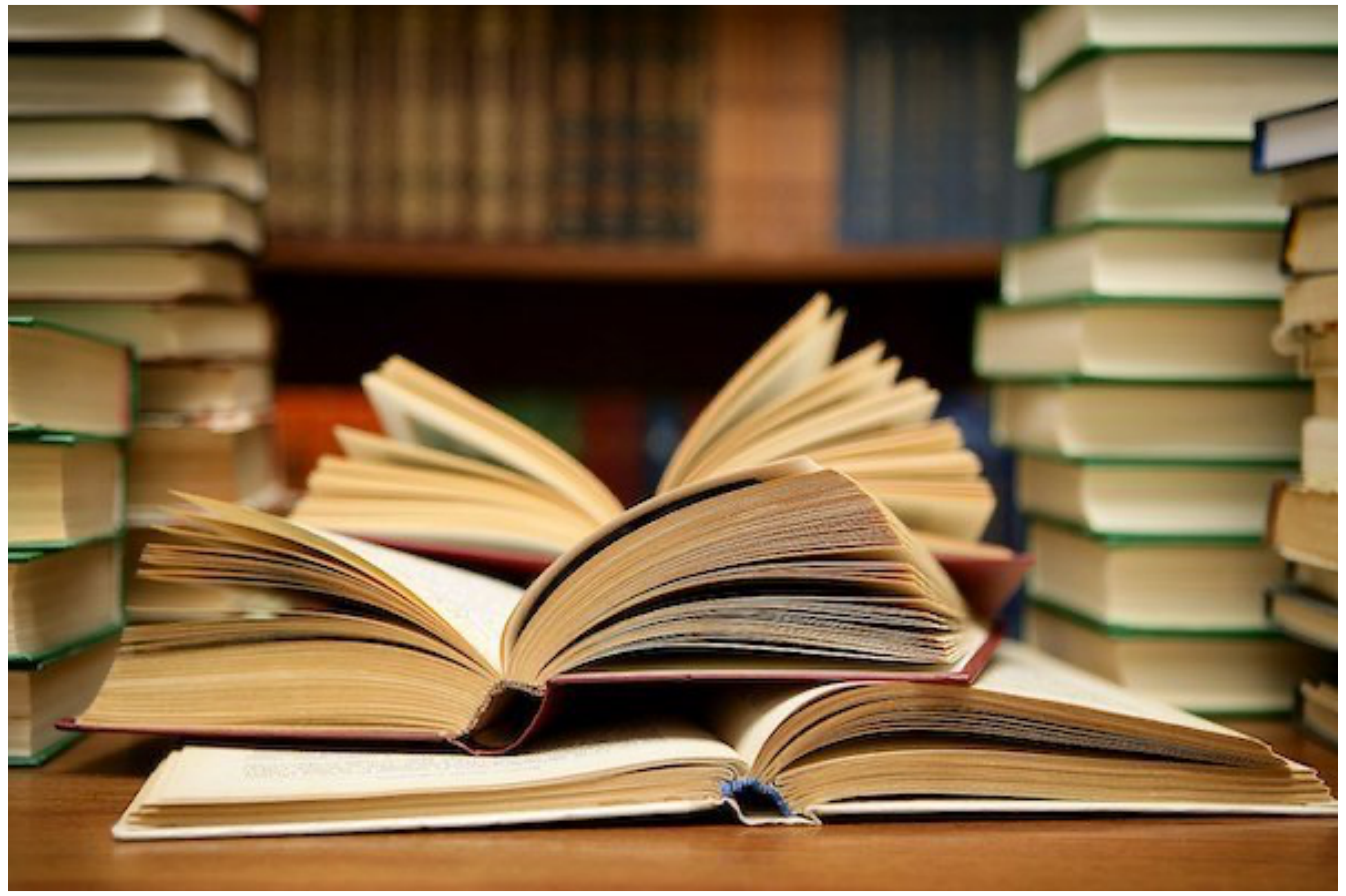
What about GPT-3?

GPT-3 needs to memorize all information in its internal parameters without explicit retrieval!



Who is the first person to go to Mariana Trench?

The first person to go to the Mariana Trench was the American oceanographer and adventurer Don Walsh, who descended to its deepest point, the Challenger Deep, in 1960.

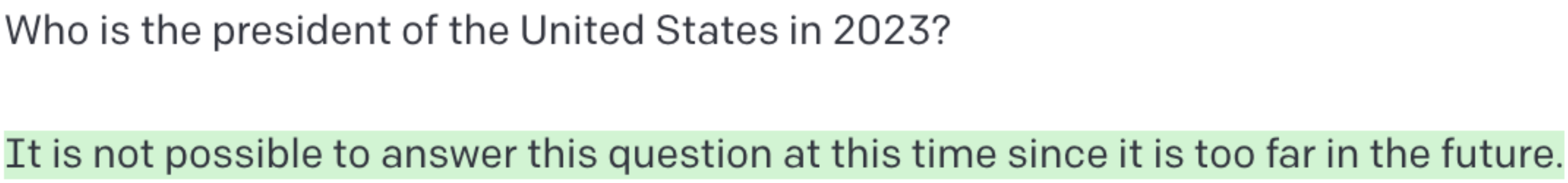


Open-book QA

MAN, THESE CLOSED BOOK EXAMS SUCK! WHO COULD POSSIBLY REMEMBER ALL THIS CRAP?
DONE AND DONE!
CARTOONSTOCK
Search ID: cszn7
SATZ

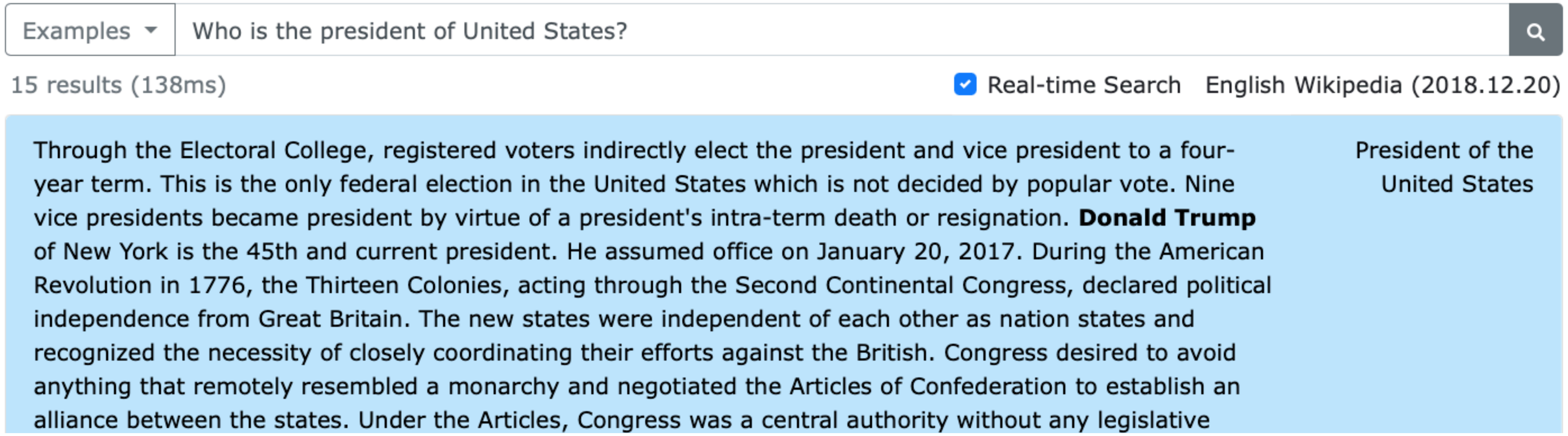
Closed-booK QA

Open-book QA vs closed-book QA



Who is the president of the United States in 2023?

It is not possible to answer this question at this time since it is too far in the future.



Examples
Who is the president of United States?
15 results (138ms)
☑ Real-time Search
English Wikipedia (2018.12.20)

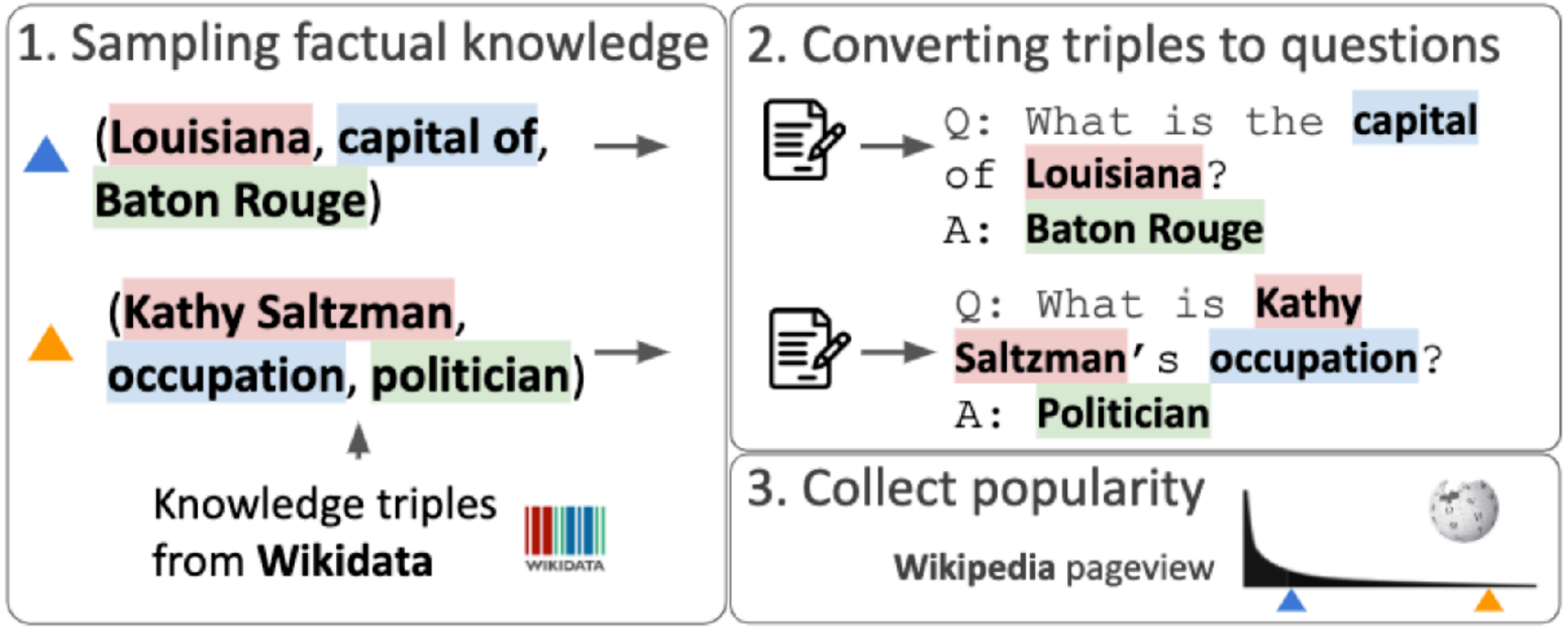
Through the Electoral College, registered voters indirectly elect the president and vice president to a four-year term. This is the only federal election in the United States which is not decided by popular vote. Nine vice presidents became president by virtue of a president's intra-term death or resignation. **Donald Trump** of New York is the 45th and current president. He assumed office on January 20, 2017. During the American Revolution in 1776, the Thirteen Colonies, acting through the Second Continental Congress, declared political independence from Great Britain. The new states were independent of each other as nation states and recognized the necessity of closely coordinating their efforts against the British. Congress desired to avoid anything that remotely resembled a monarchy and negotiated the Articles of Confederation to establish an alliance between the states. Under the Articles, Congress was a central authority without any legislative

President of the United States

Wikipedia 2018.12.20 snapshot

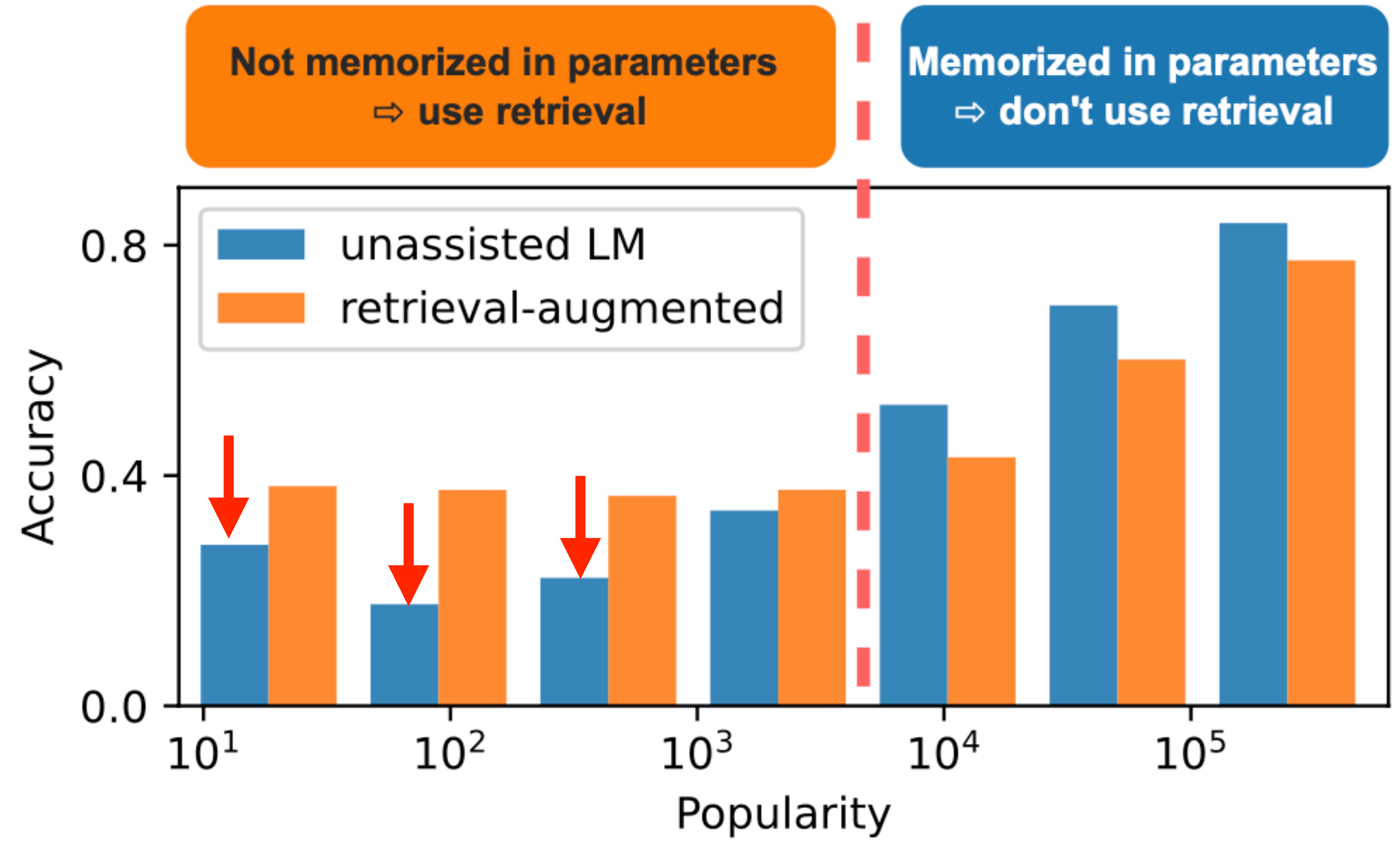
- We can store the entire Wikipedia corpus in 30Gb (GPT-3 needs at least 700Gb!)
- More importantly, we can control what text we put in this database and update it easily

Open-book QA vs closed-book QA


1. Sampling factual knowledge
(Louisiana, capital of, Baton Rouge)
(Kathy Saltzman, occupation, politician)
Knowledge triples from Wikidata
WIKIDATA
2. Converting triples to questions
Q: What is the capital of Louisiana?
A: Baton Rouge
Q: What is Kathy Saltzman's occupation?
A: Politician
3. Collect popularity
Wikipedia pageview

What is Kathy Saltzman's occupation?

GPT-3 davinci-003: 20%-30% accuracy



Mallen et al., 2022: When Not to Trust Language Models: Investigating Effectiveness and Limitations of Parametric and Non-Parametric Memories