Natural Language Processing with Deep Learning

CS224N/Ling284

Diyi Yang

Lecture 12: Question Answering

(slides based on John Hewitt, Danqi Chen, Nelson Liu)

Overview

1. What is question answering? (10 mins)

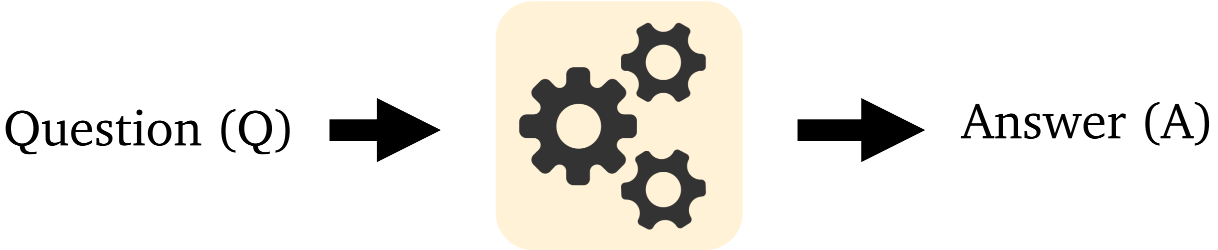
2. Reading comprehension (30 mins)
   - ❑ How to answer questions over **a single passage of text**

3. Open-domain (textual) question answering (20 mins)
   - ❑ How to answer questions over **a large collection of documents**

4. Retrieval-augmented generation for question answering (10 mins)

- Next lecture: Nicholas Carlini on *Privacy and Security in LLMs*
- Assignment 5 due on Feb 18th

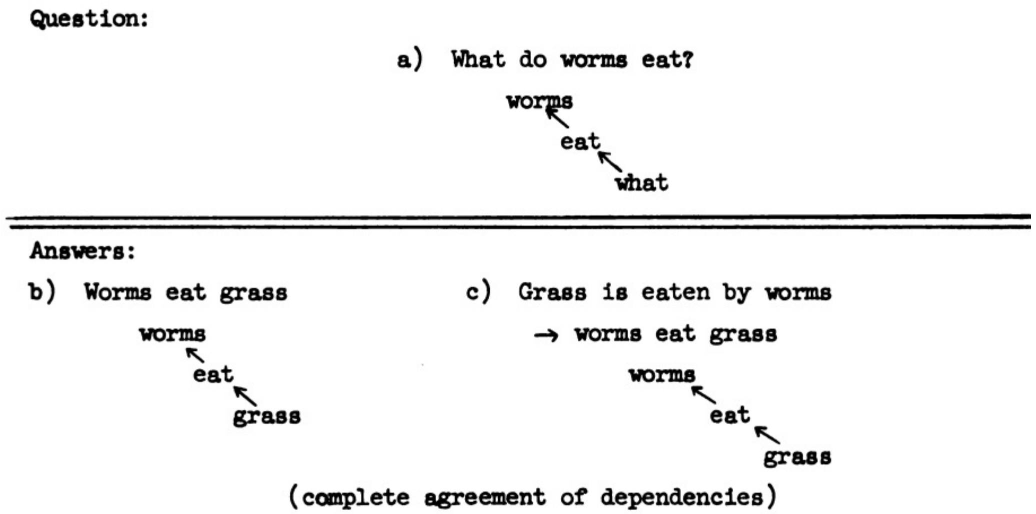
1. What is question answering?



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

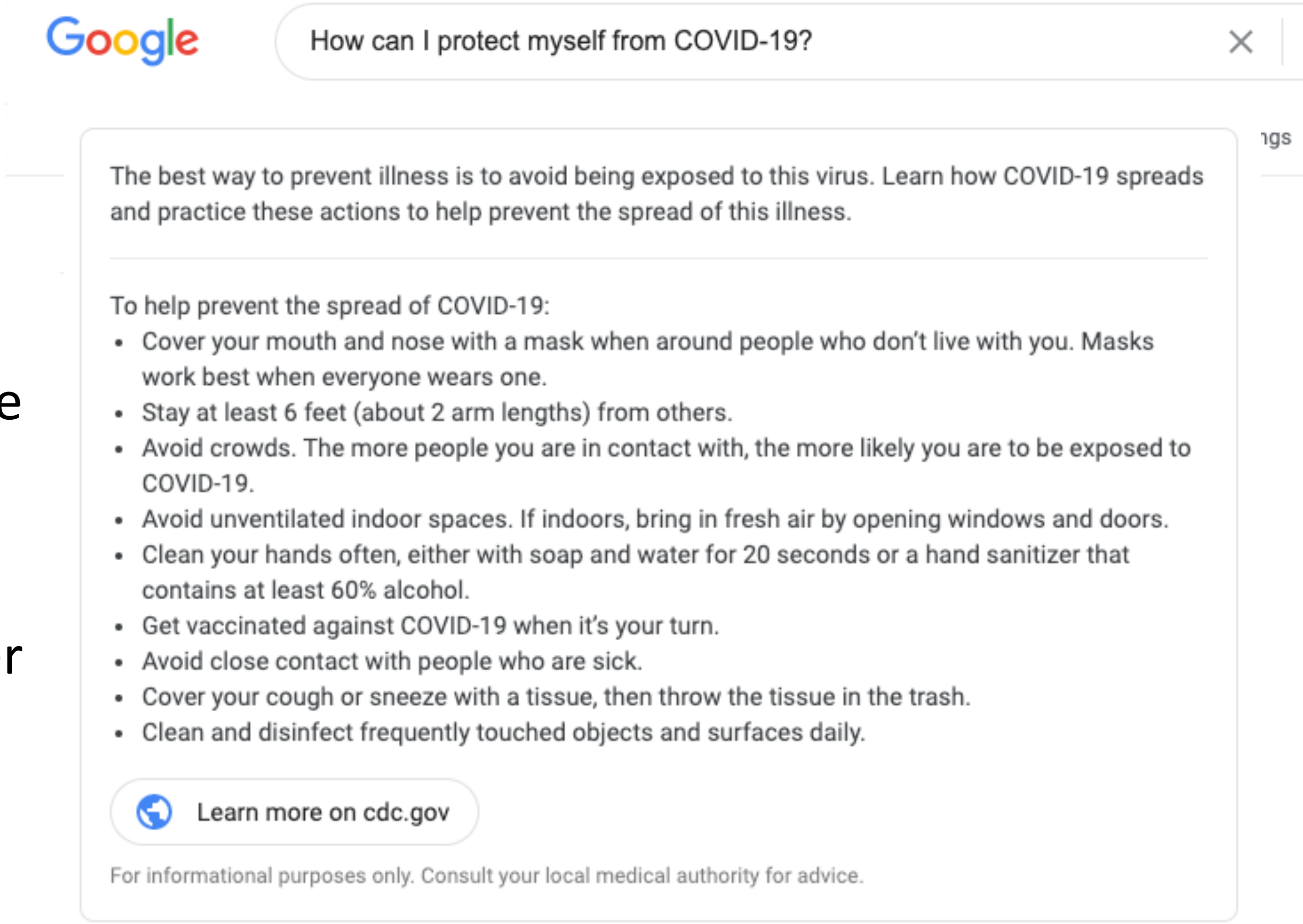
The earliest QA systems dated back to 1960s!

(Simmons et al., 1964)



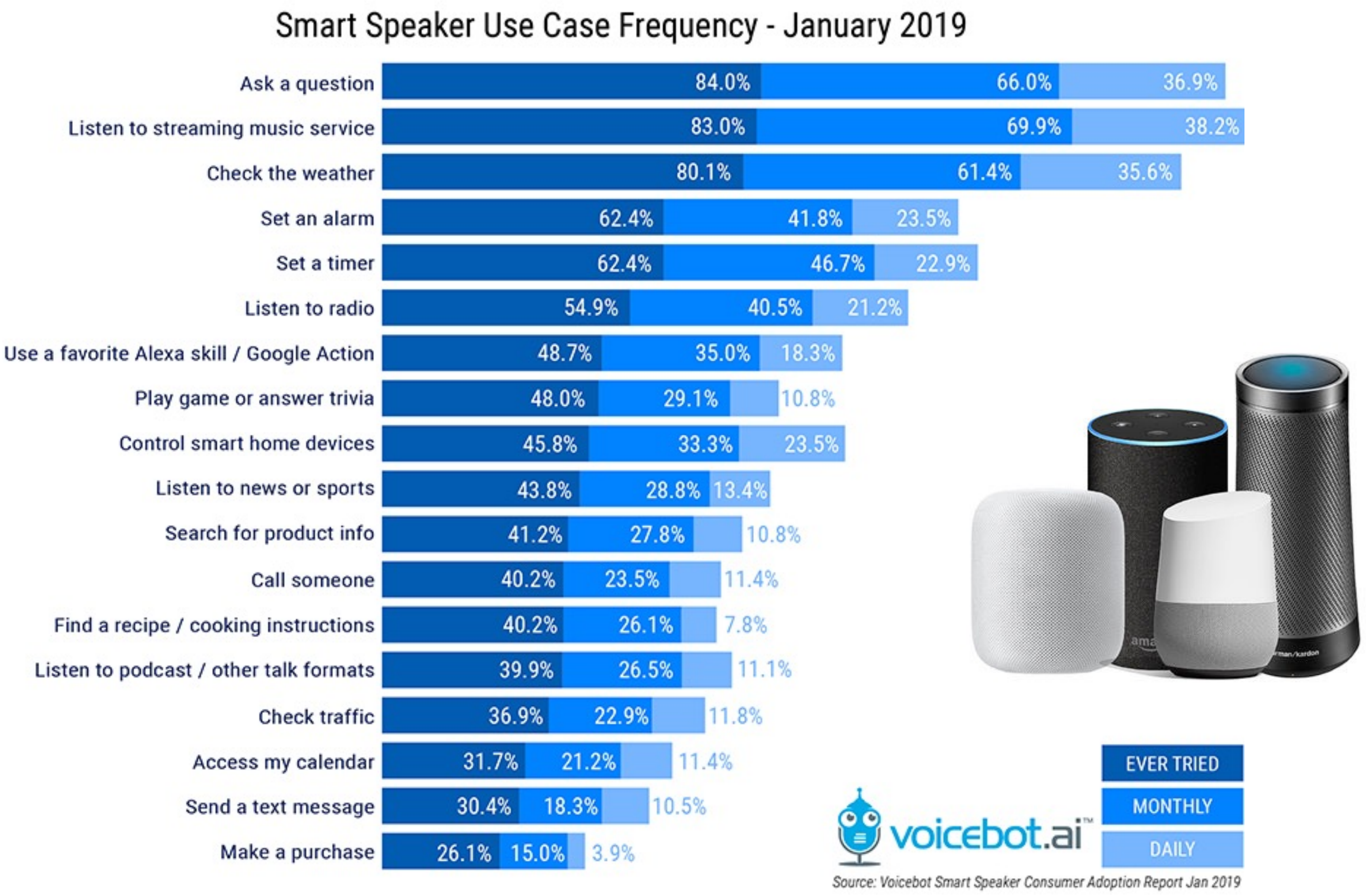
1. What is question answering?

- Lots of immediate applications: search engines, dialogue systems
- QA is an important testbed for evaluating how well compute systems understand human language
- "Since question can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**"



1. What is question answering?

- People ask a lot of questions to Digital Personal Assistants



Question answering taxonomy

- Factoid questions vs. Non-factoid questions
- **Answers**
  - A short span of text
  - A paragraph
  - Yes/No
  - A database entry
  - A list
- **Context**
  - A passage, a document, a large collection of documents
  - Knowledge base
  - Semi-structured tables
  - Images
  - The web

2011: IBM Watson beat Jeopardy champions

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

IBM Watson beat Jeopardy champions

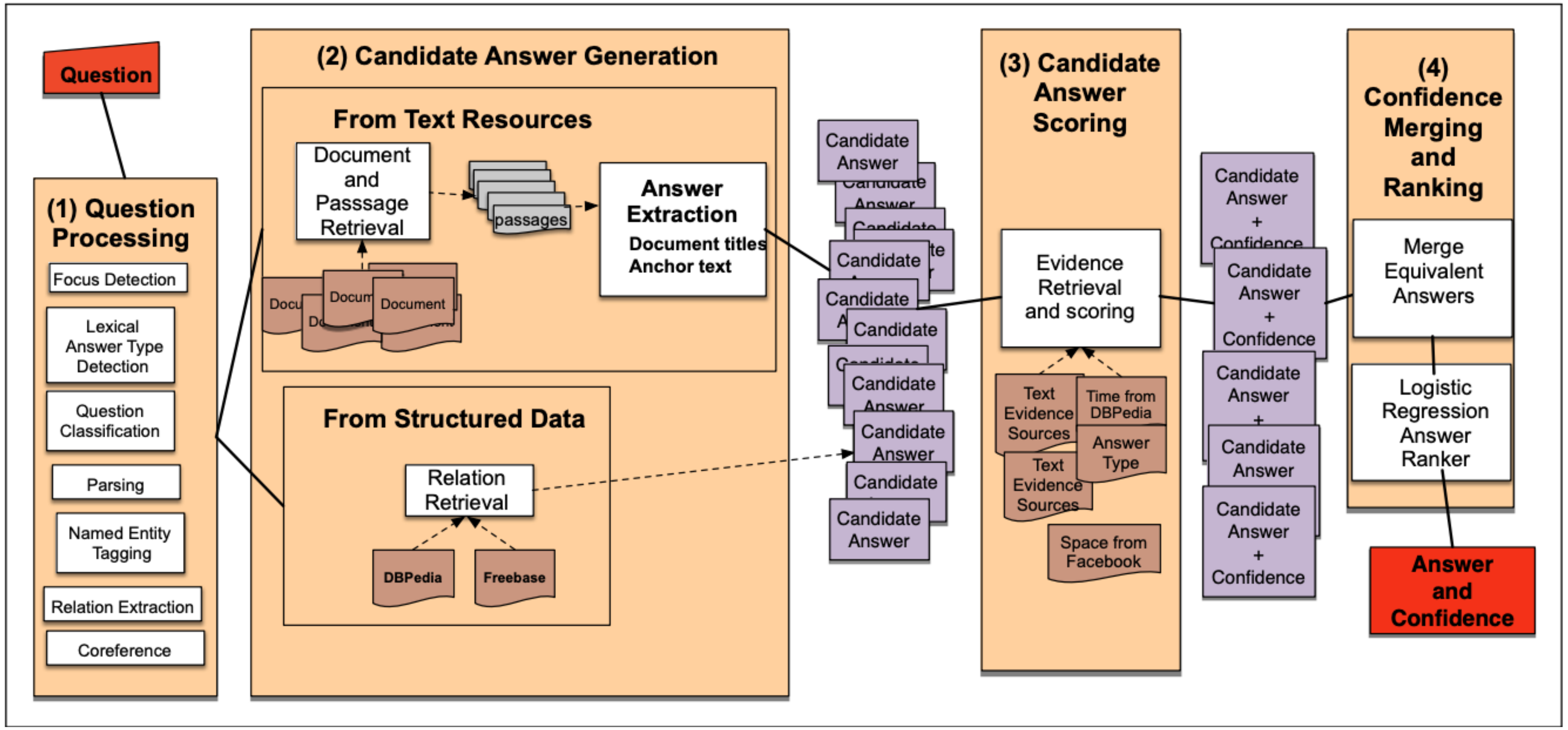


Image credit:
J & M, edition 3

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

Question answering in deep learning era

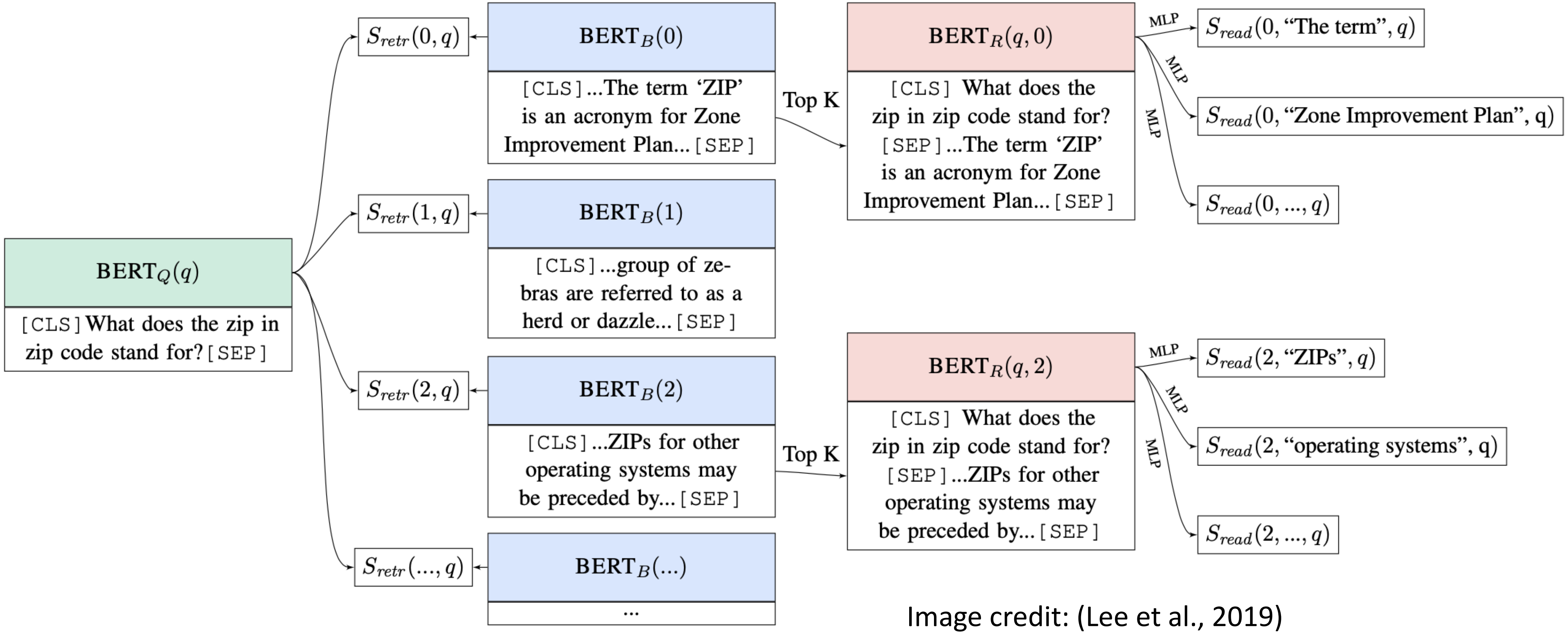


Image credit: (Lee et al., 2019)

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)!

Textual Question Answering (Reading Comprehension)

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

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

Textual Question Answering

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?

A) Fries

B) Pudding

C) James

D) Jane

2) What did James pull off of the shelves in the grocery store?

A) pudding

B) fries

C) food

D) splinters

(Richardson et al, 2013): MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Conversational Question Answering

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q: What are the candidates **running** for? A: Governor

Q: **Where?** A: Virginia

Q: Who is the democratic candidate? A: Terry McAuliffe

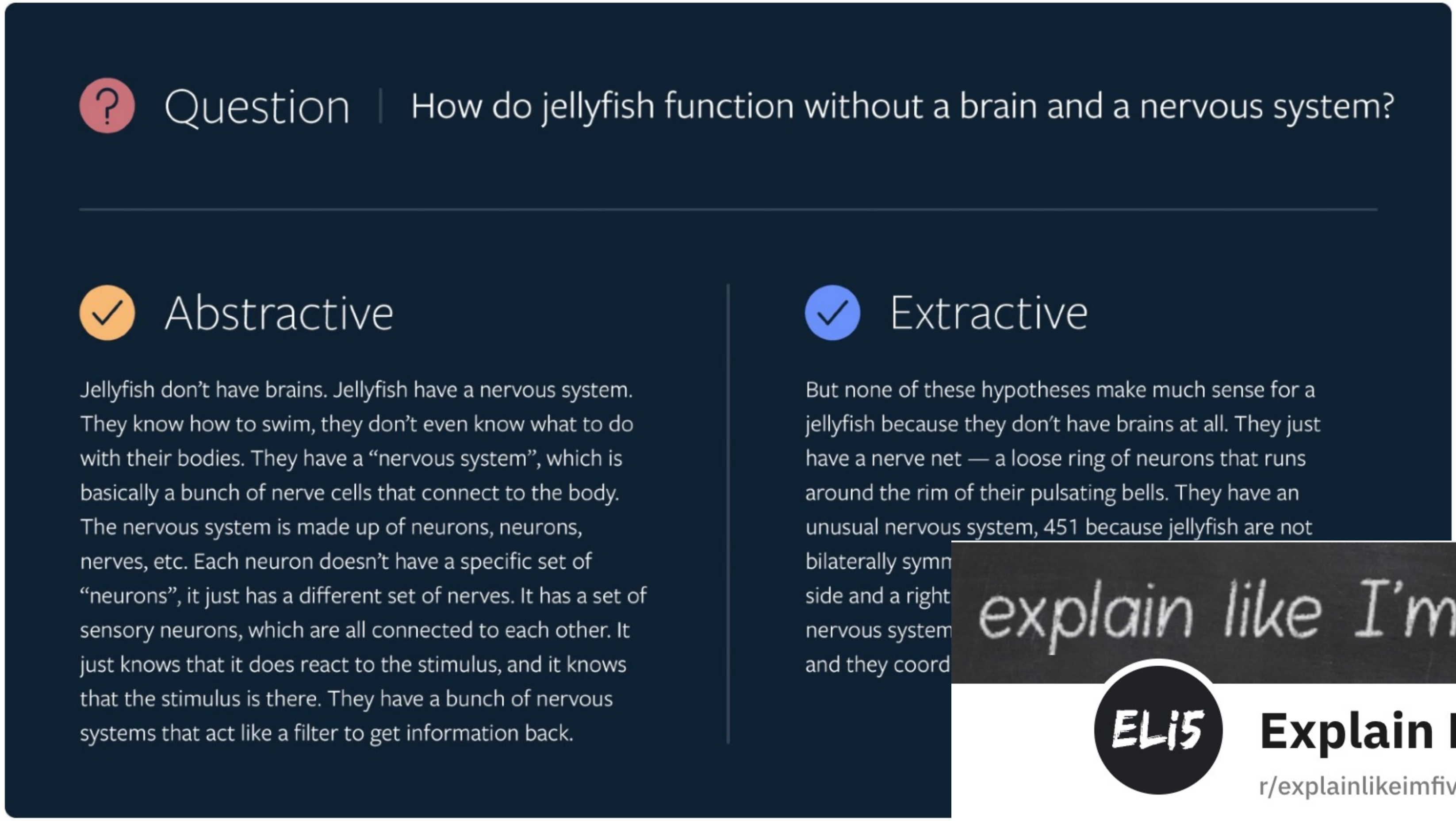
Q: Who is **his** opponent? A: Ken Cuccinelli

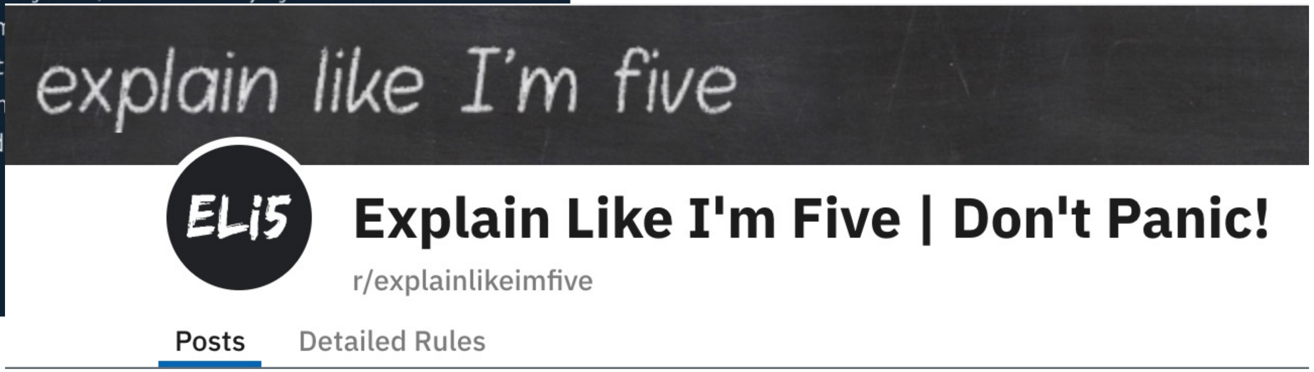
Q: What party does **he** belong to? A: Republican

Q: Which of **them** is winning?

(Reddy et al, 2019): CoQA: A Conversational Question Answering Challenge

Long-form Question Answering

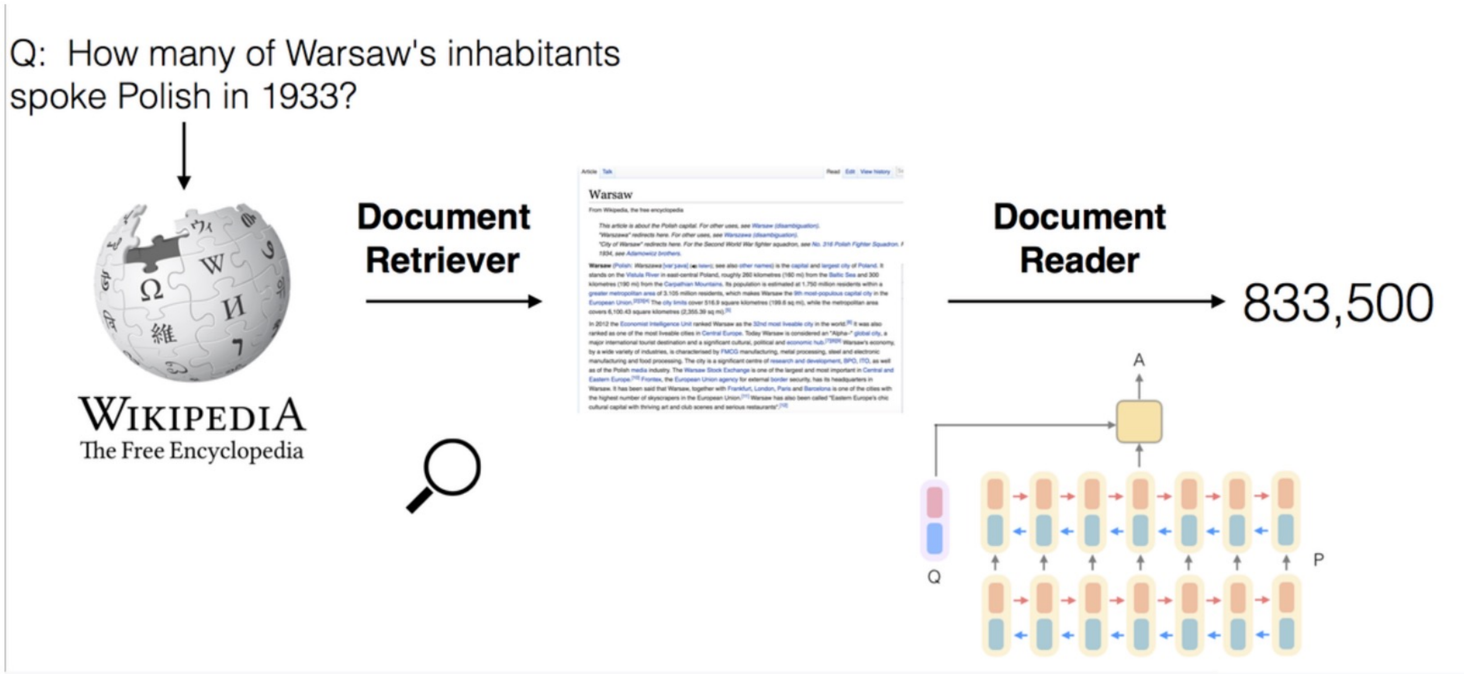




https://ai.facebook.com/blog/longform-qa/
(Fan et al, 2019): ELI5: Long Form Question Answering

Open-domain Question Answering

DrQA



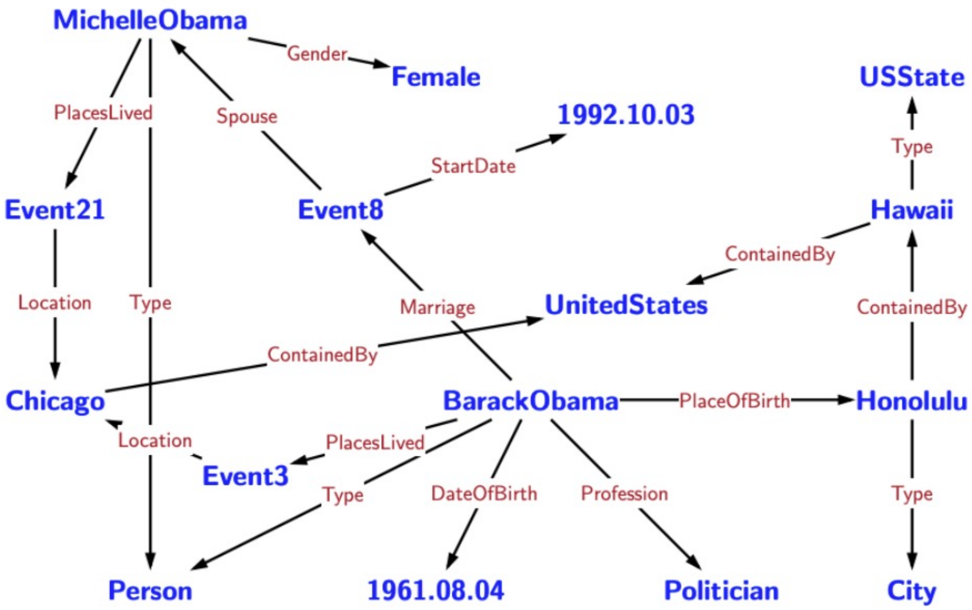

```
>>> process('What is the answer to life, the universe, and everything?')

Top Predictions:
+------+--------+---------------------------------------------------+--------------+-----------+
| Rank | Answer |                        Doc                        | Answer Score | Doc Score |
+------+--------+---------------------------------------------------+--------------+-----------+
|  1   |   42   | Phrases from The Hitchhiker's Guide to the Galaxy |    47242     |   141.26  |
+------+--------+---------------------------------------------------+--------------+-----------+
```

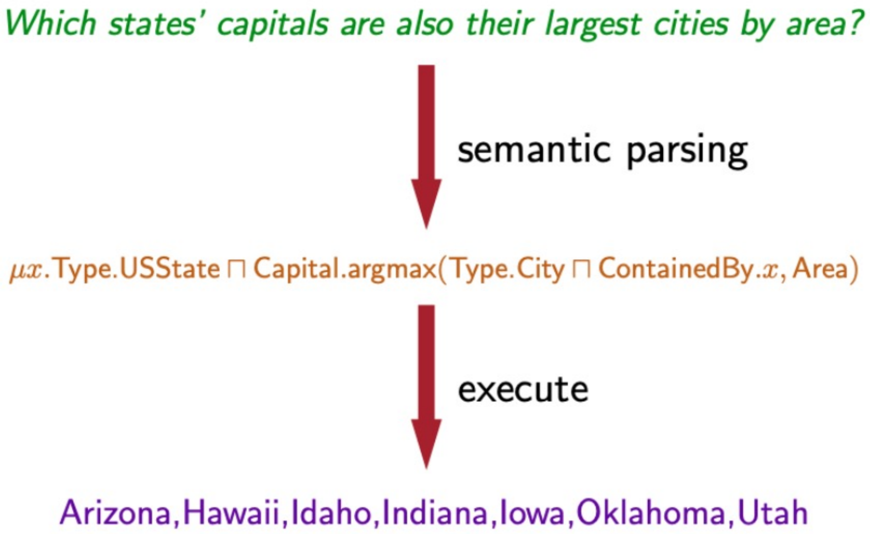
(Chen et al, 2017): Reading Wikipedia to Answer Open-Domain Questions

Knowledge Base Question Answering

Freebase

100M entities (nodes) 1B assertions (edges)





(Berant et al, 2013): Semantic Parsing on Freebase from Question-Answer Pairs

Table-based Question Answering

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x = Greece held its last Summer Olympics in which year?

y = 2004

(Pasupat and Liang, 2015): Compositional Semantic Parsing on Semi-Structured Tables.

Visual Question Answering

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

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

GPT-4
As of Feb 15th, 2024

how many slices of pizza are there? is this a vegetarian pizza?

ChatGPT
There are 8 slices of pizza in the image. As for whether it's a vegetarian pizza, it appears to have toppings that could be vegetables, such as pieces that look like diced tomatoes and possibly green peppers. However, without a clearer image or a list of ingredients, it's impossible to be certain if there are no meat products on the pizza. Vegetarian pizzas typically do not include meat, but they can include a variety of vegetable toppings like onions, peppers, mushrooms, olives, and more. If there is any meat or meat-based toppings visible, then it would not be a vegetarian pizza.

(Antol et al, 2015): Visual Question Answering

Why do we care about this problem?

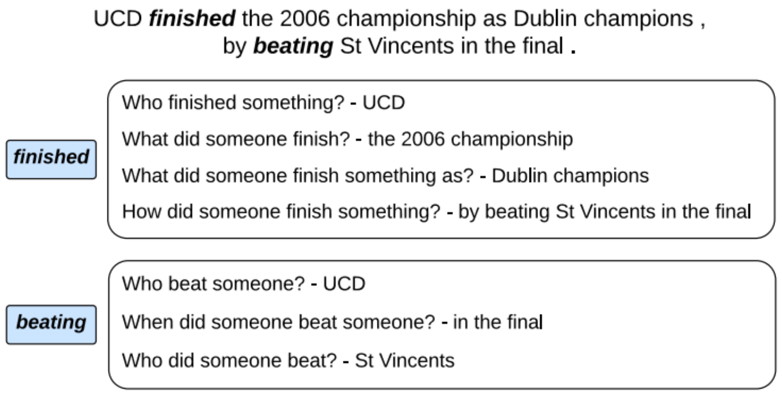
- Reading comprehension is an important testbed for evaluating how well computer systems understand human language
  - Wendy Lehnert 1977: “Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding.”
- 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



(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!

Stanford Question Answering Dataset (SQuAD)

- SQuAD was **for years the most popular** reading comprehension dataset; it is "almost solved" today (though the underlying task is not,) 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 of Text

SQuAD Evaluation

- **Evaluation:** exact match (0 or 1) and F1 (partial credit).
- For development and testing sets, 3 gold answers are collected, because there could be multiple plausible answers.
- We compare the predicted answer to each gold answer *(a, an, the, punctuations are removed)* and take max scores. Finally, we take the average of all the examples for both exact match and F1.

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Q: Rather than taxation, what are private schools largely funded by?

A: {tuition, charging their students tuition, tuition}

Other question answering datasets

- **TriviaQA:** Questions and answers by trivia enthusiasts. Independently collected web paragraphs that contain the answer and seem to discuss question, but no human verification that paragraph supports answer to question
- **Natural Questions:** Question drawn from frequently asked Google search questions. Answers from Wikipedia paragraphs. Answer can be substring, yes, no, or *NOT_PRESENT*. Verified by human annotation.
- **HotpotQA.** Constructed questions to be answered from the whole of Wikipedia which involve getting information from two pages to answer a multistep query:
  - Q: Which novel by the author of "Armada" will be adapted as a feature film by Steven Spielberg? A: *Ready Player One*

Conventional feature-based methods for reading comprehension

- Generate a list of candidate answers $(a_1, a_2, \dots, a_M)$
- Define a feature vector $\phi(p, q, a_i) \in R^d$:
  - Word/bigram features
  - Parse tree matches
  - Dependency labels, length, part-of-speech tags
- Apply a multi-class logistic regression model

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

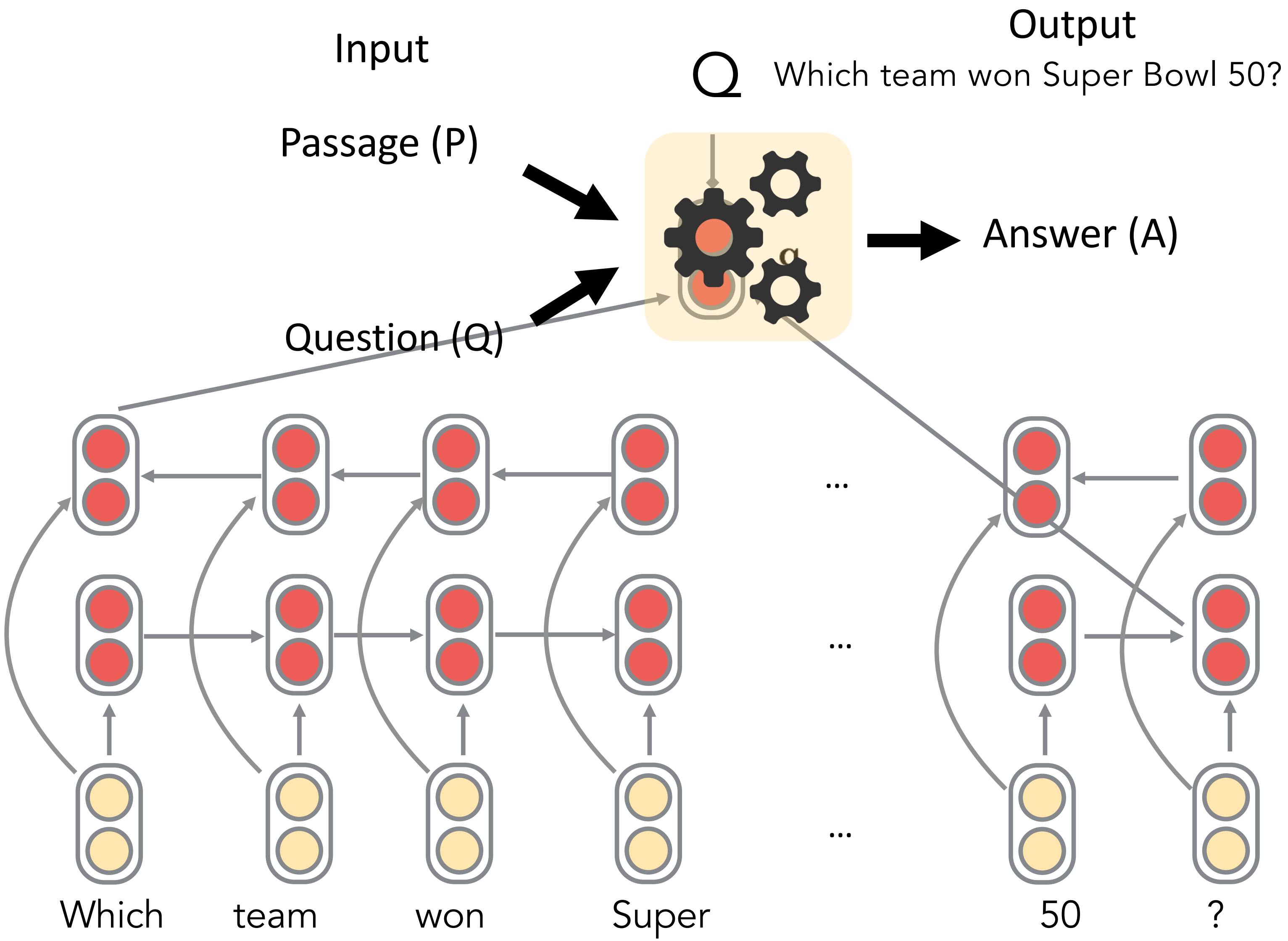
Neural models for reading comprehension

- How can we build a neural model to solve SQuAD?

- 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≤ start ≤ end ≤N
  - N~100, M ~15

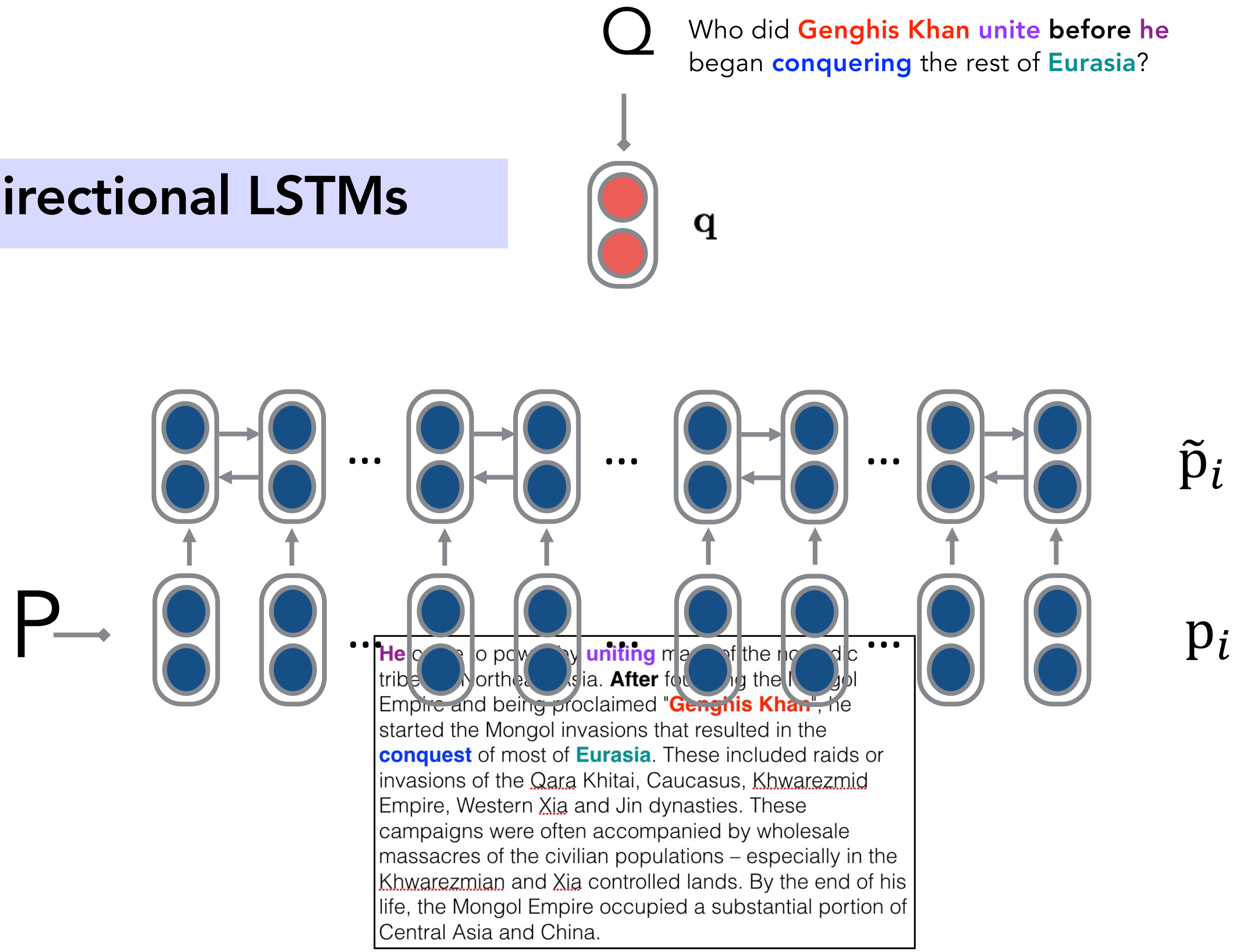
- Stanford Attentive Reader [Chen, Bolton, & Manning 2016] [Chen, Fisch, Weston & Bordes 2017]
  - Demonstrated a minimal, highly successful architecture for reading comprehension and question answering
  - Became known as the Stanford Attentive Reader

Stanford Attentive Reader



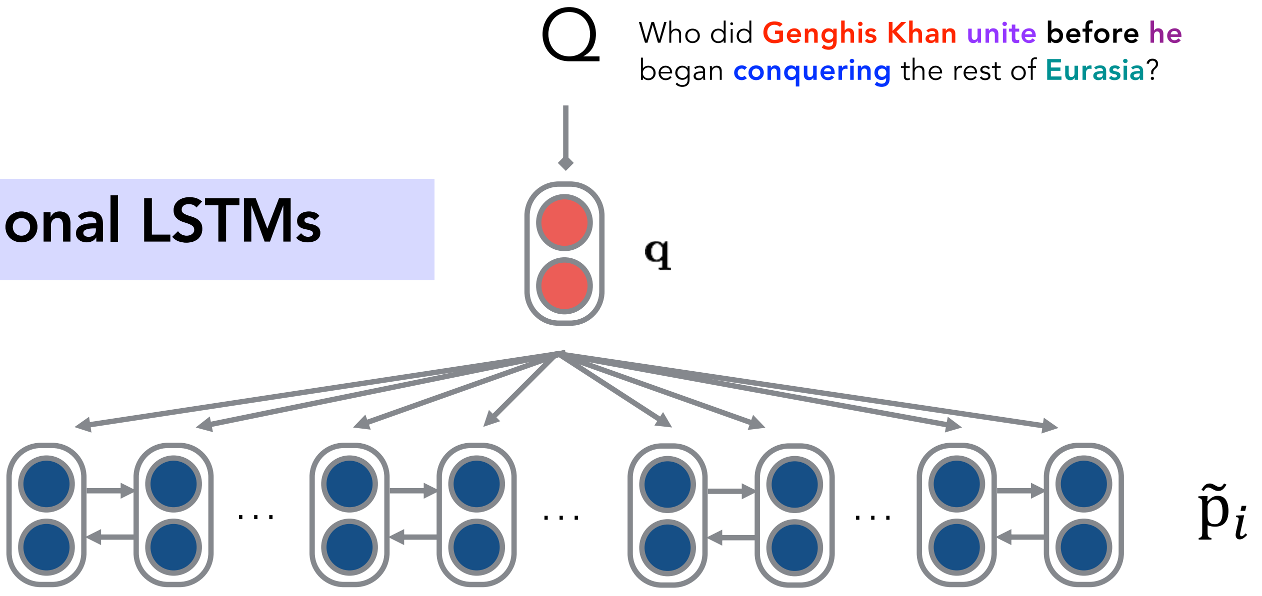
Stanford Attentive Reader

Bidirectional LSTMs



Stanford Attentive Reader

Bidirectional LSTMs



He came to power by uniting many of the nomadic tribes of Northeast Asia. After founding the Mongol Empire and being proclaimed "Genghis Khan", he started the Mongol invasions that resulted in the conquest of most of Eurasia. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmian and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.

Attention

$$\alpha_i = \boxed{\operatorname{softmax}_i (\mathbf{q}^\top \mathbf{W} \ldots}$$

→ predict **start** token

Attention

$$\alpha'_i = \operatorname{softmax}_i(\mathbf{q}^\top \mathbf{W}'_s \tilde{\mathbf{p}}_i)$$

→ predict **end** token

SQuAD 1.1 Results (single model, c. Feb 2017)

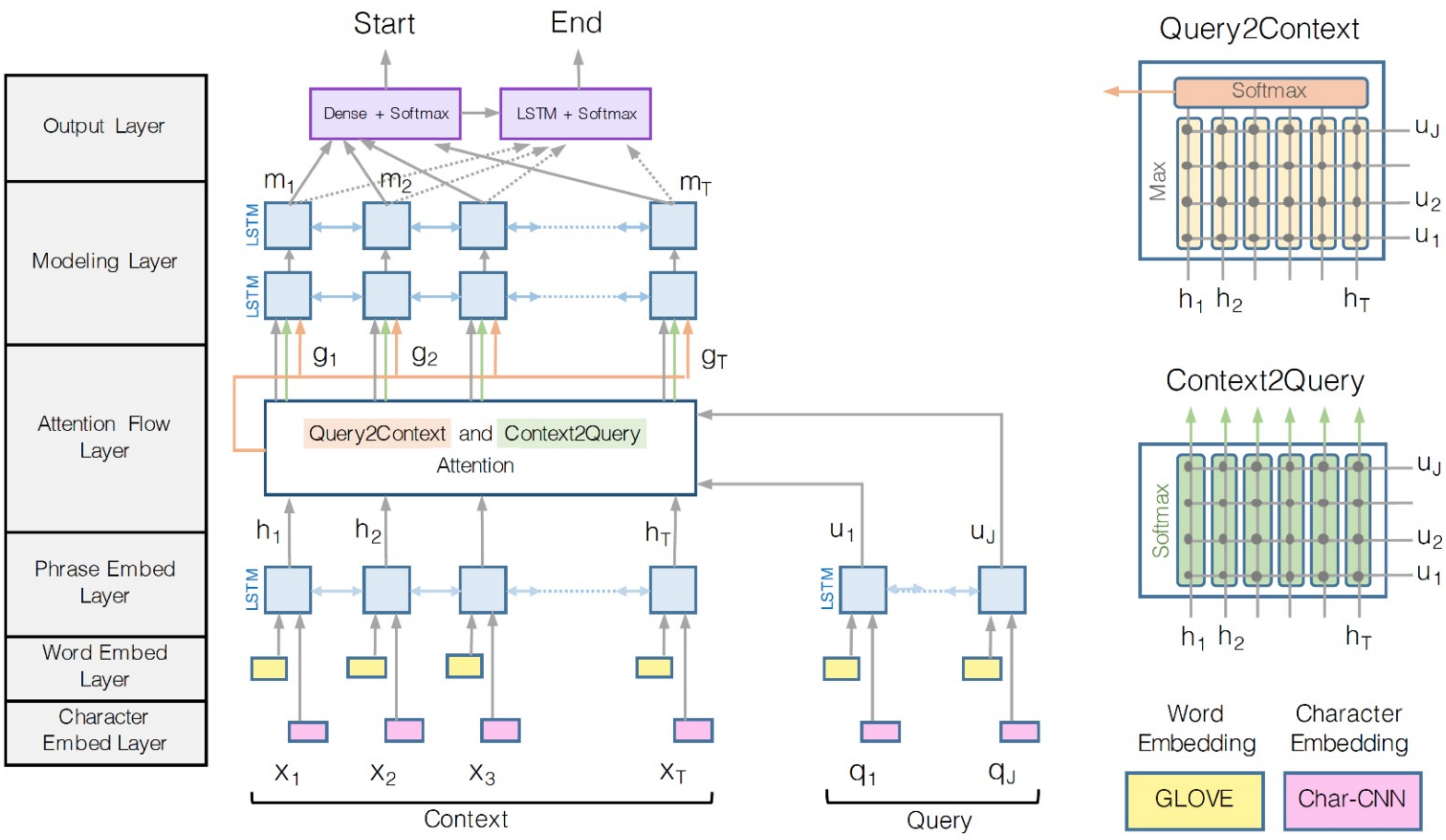
	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]	79.7
Google Brain / CMU (Feb 2018)	88.0
Human performance	91.2
Pretrained + Finetuned Models circa 2021	>93.0

BiDAF: the Bidirectional Attention Flow model

- Encode the question using word/char embeddings; pass on an biLSTM encoder
- Encode the passage similarly
- Passage-to-question and question-to-passage attention
- Modeling layer: another BiLSTM layer
- Output layer: two classifiers for predicting start and end points
- The entire model can be trained in an end-to-end way

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

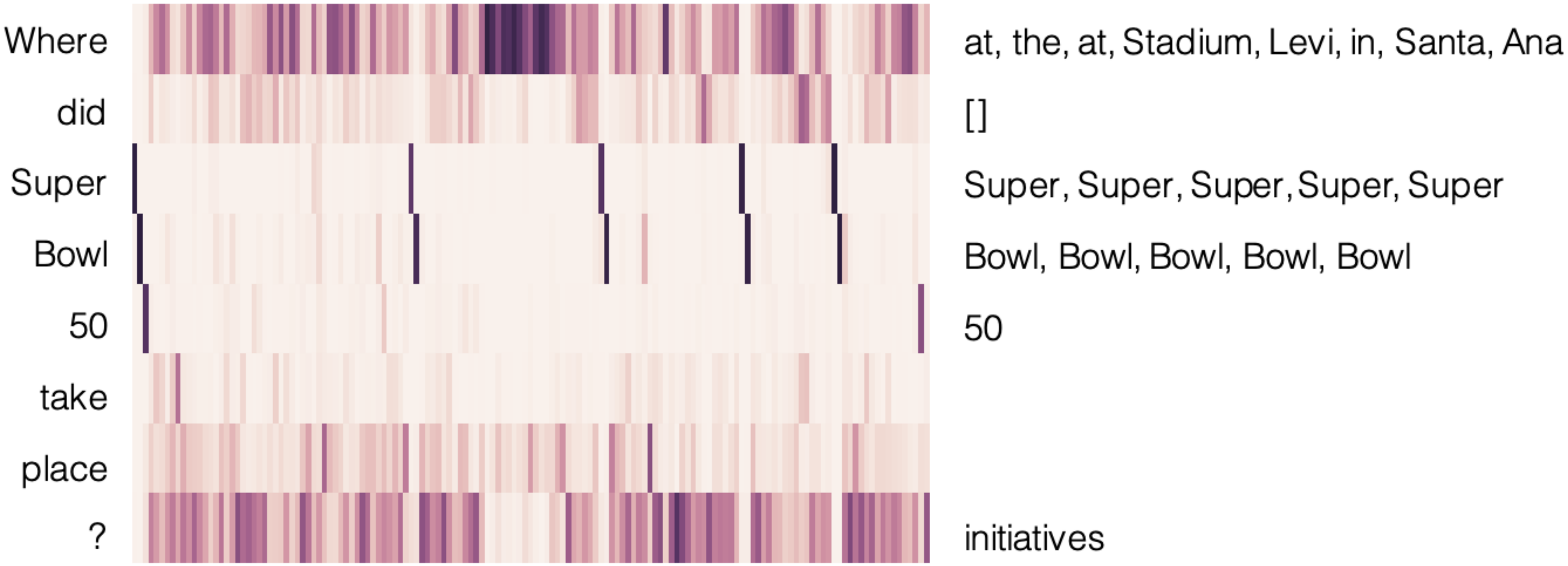
BiDAF: the Bidirectional Attention Flow model



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

BiDAF: attention visualization

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season . The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title . The game was played on February 7 , 2016 , **at Levi 's Stadium in the San Francisco Bay Area at Santa Clara , California** . As this was the 50th Super Bowl , the league emphasized the " golden anniversary " with various gold-themed initiatives , as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as " Super Bowl L "), so that the logo could prominently feature the Arabic numerals 50 .



BiDAF: results on SQuAD

	Single Model		Ensemble	
	EM	F1	EM	F1
Logistic Regression Baseline[a]	40.4	51.0	-	-
Dynamic Chunk Reader[b]	62.5	71.0	-	-
Fine-Grained Gating[c]	62.5	73.3	-	-
Match-LSTM[d]	64.7	73.7	67.9	77.0
Multi-Perspective Matching[e]	65.5	75.1	68.2	77.2
Dynamic Coattention Networks[f]	66.2	75.9	71.6	80.4
R-Net[g]	**68.4**	**77.5**	72.1	79.7
BiDAF (Ours)	68.0	77.3	**73.3**	**81.1**

(a) Results on the SQuAD test set

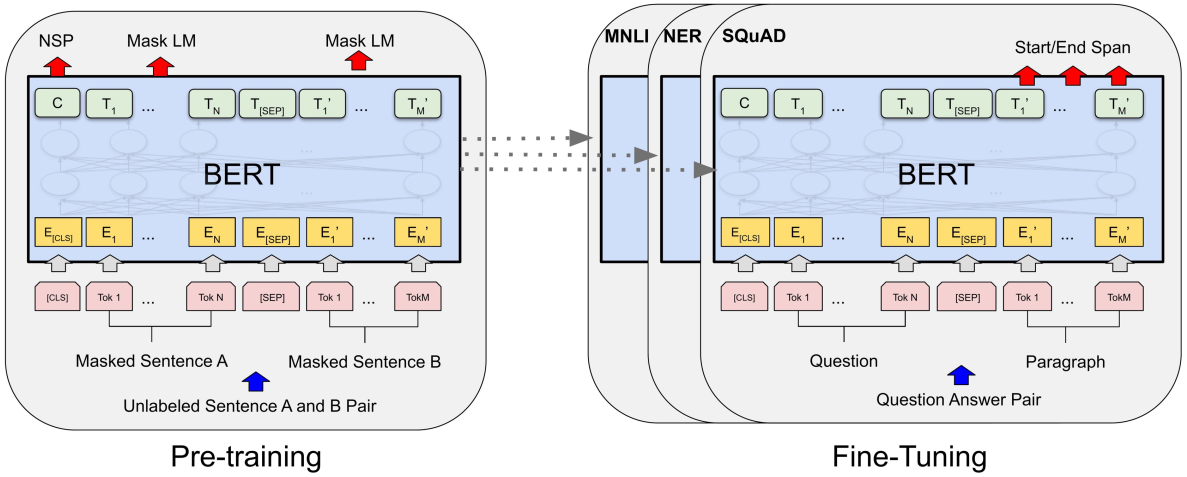
	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BiDAF (single)	67.7	77.3
BiDAF (ensemble)	72.6	80.7

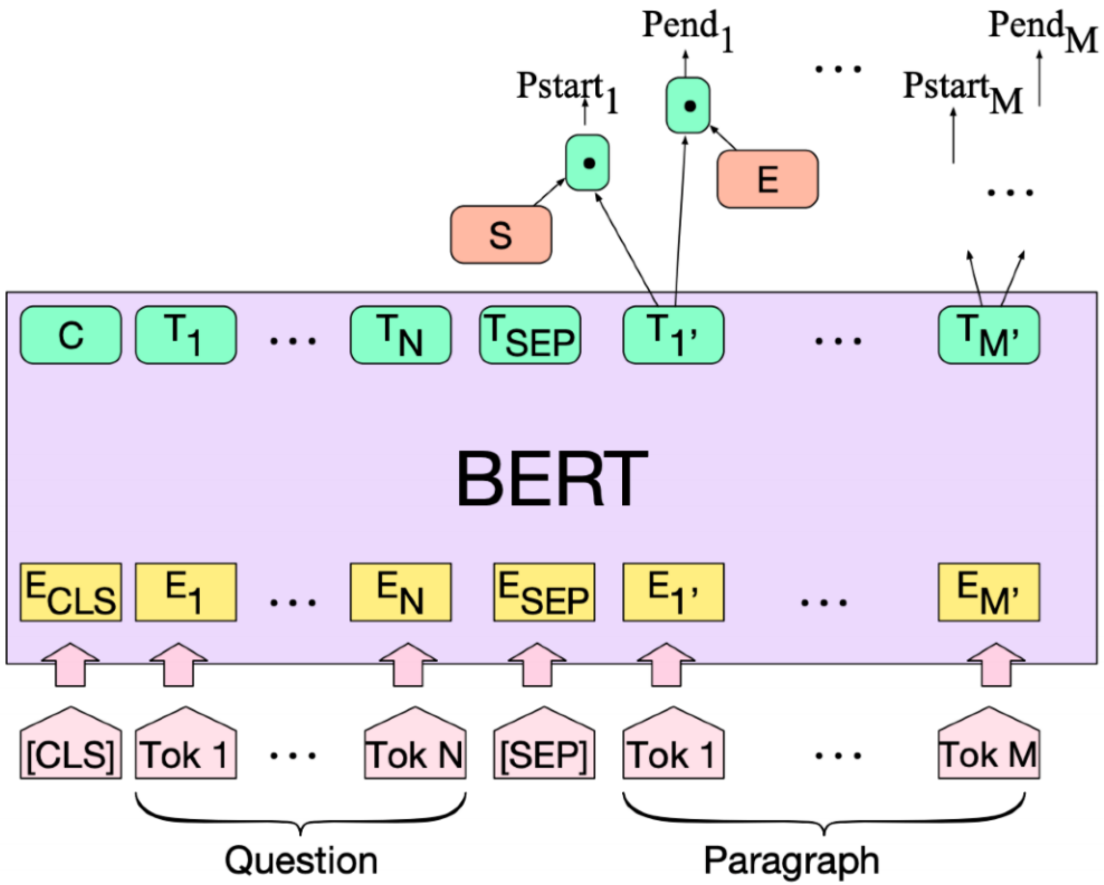
(b) Ablations on the SQuAD dev set

BERT for reading comprehension

- BERT is pre-trained on two training objectives:
  - Masked language model (MLM)
  - Next sentence prediction (NSP)

- $\text{BERT}_{\text{base}}$ has 12 layers and 110M parameters, $\text{BERT}_{\text{large}}$ has 24 layers and 330M parameters





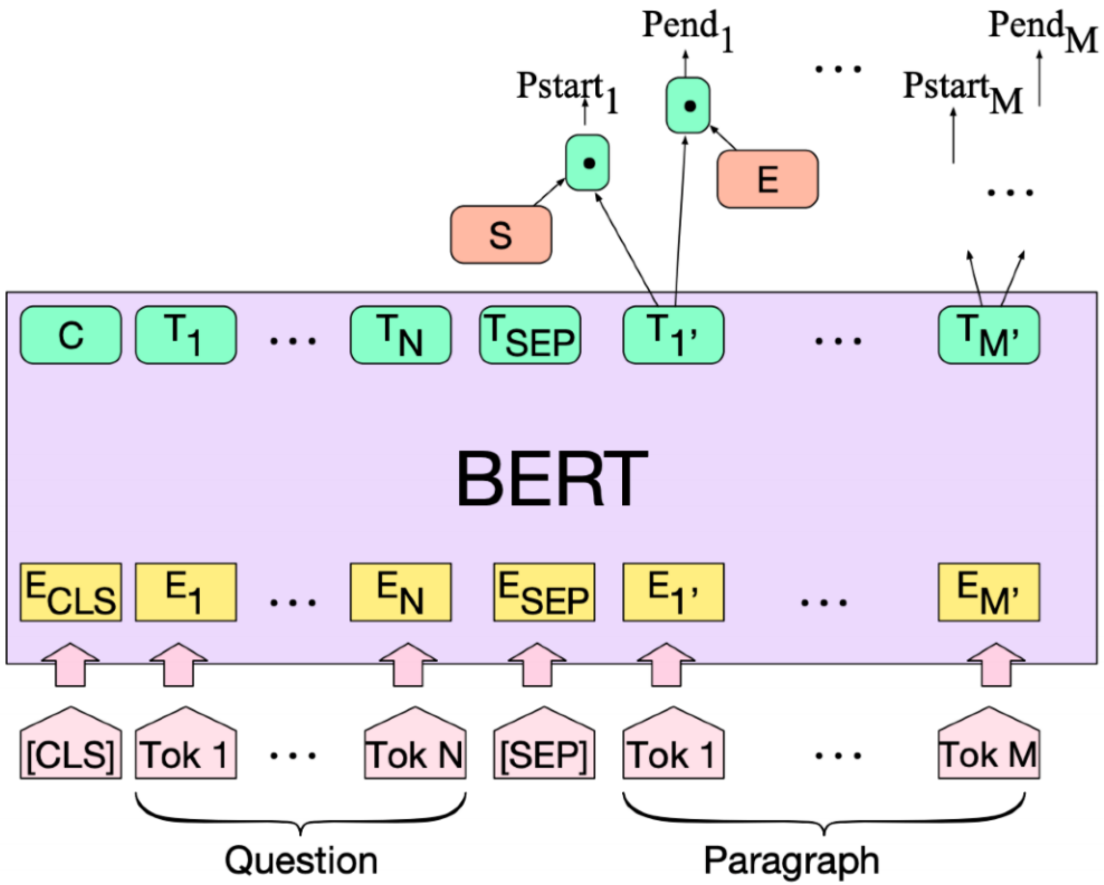
BERT for reading comprehension

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^{\intercal}\mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^{\intercal}\mathbf{h}_i)$$

where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT



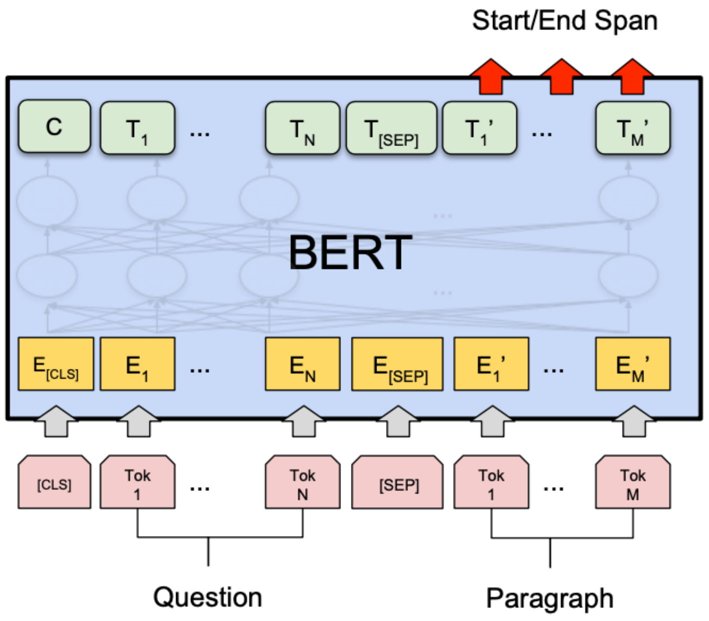
BERT for reading comprehension

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

- All the BERT parameters (e.g., 110M) as well as the newly introduced parameters (e.g., 768 x 2 = 1536) are optimized together for h_{start}, h_{end} (e.g., 768x2=1536) are optimized together for $\mathcal{L}$
- It works amazing well. Stronger pre-trained language models can lead to even better performance and SQuAD becomes a standard dataset for testing pre-trained models.

	F1	EM
Human performance	91.2*	82.3*
BiDAF	77.3	67.7
BERT-base	88.5	80.8
BERT-large	90.9	84.1
XLNet	94.5	89.0
RoBERTa	94.6	88.9
ALBERT	94.8	89.3

(dev set, except for human performance)



Comparisons between BiDAF and BERT models

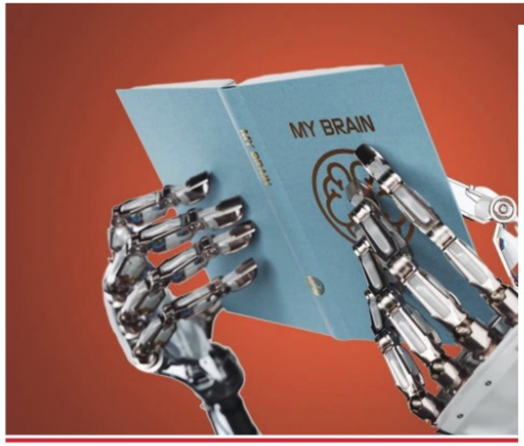
- BERT model has many many more parameters (110M or 330M)
  BiDAF has ~2.5M parameters.
- BiDAF is built on top of several bidirectional LSTMs while BERT is built on top of Transformers (no recurrence architecture and easier to parallelize).
- BERT is pre-trained while BiDAF is only built on top of GloVe (and all the remaining parameters need to be learned from the supervision datasets).

Pre-training is clearly a game changer but it is expensive..

Is reading comprehension solved?

AI systems are beating humans in reading comprehension

By Associated Press January 24, 2018 | 2:25pm



Artificial Intelligence Jan 15, 2018

AI Beats Humans at Reading Comprehension, but It Still Doesn't Truly Comprehend Language

AI Beat Humans at Reading! Maybe Not

Microsoft and Alibaba claimed software could read like a human. There's more to the story than that.

Is reading comprehension solved?

- We have already surpassed human performance on SQuAD. Does it mean that reading comprehension is already solved?

- The current systems still perform poorly on adversarial examples or examples from out-of-domain distributions (Jia and Liang, 2017)

- Systems trained on one dataset can't generalize to other datasets (Sen and Saffari, 2020)

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

		Evaluated on				
		SQuAD	TriviaQA	NQ	QuAC	NewsQA
Fine-tuned on	SQuAD	**75.6**	46.7	48.7	20.2	41.1
	TriviaQA	49.8	**58.7**	42.1	20.4	10.5
	NQ	53.5	46.3	**73.5**	21.6	24.7
	QuAC	39.4	33.1	33.8	**33.3**	13.8
	NewsQA	52.1	38.4	41.7	20.4	**60.1**

Is reading comprehension solved?

- BERT model trained on SQuAD

Failure rate

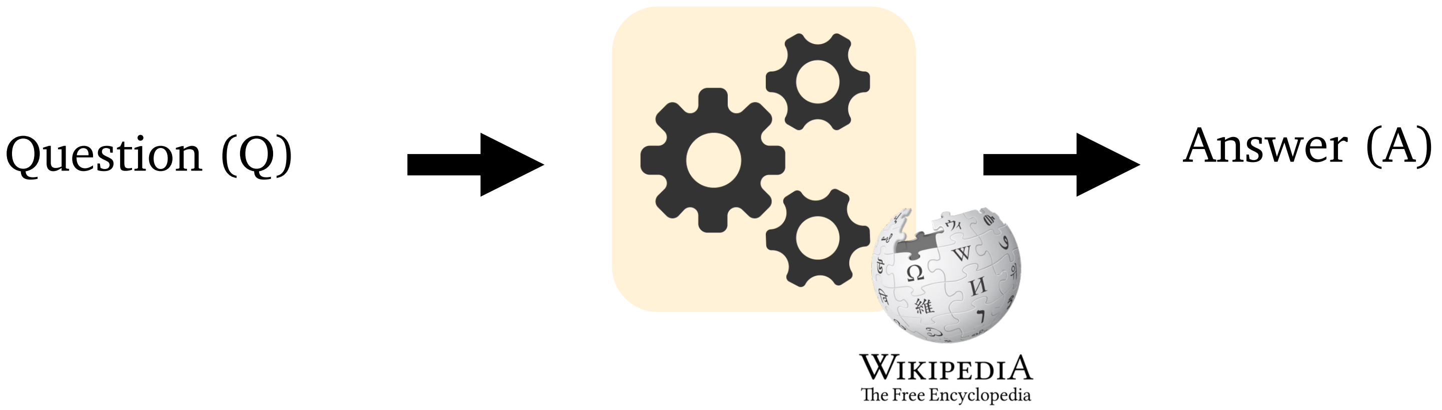
	Test	Failure rate	Example
Temporal	***MFT:*** change in one person only	41.5	**C:** Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. **Q:** Who is a model? **A:** Abigail Model: Abigail were writers, but there was a change in Abigail
Temporal	***MFT:*** Understanding before/after, last/first	82.9	**C:** Logan became a farmer before Danielle did. **Q:** Who became a farmer last? **A:** Danielle Model: Logan
Neg.	***MFT:*** Context has negation	67.5	**C:** Aaron is not a writer. Rebecca is. **Q:** Who is a writer? **A:** Rebecca Model: Aaron
Neg.	***MFT:*** **Q** has negation, **C** does not	100.0	**C:** Aaron is an editor. Mark is an actor. **Q:** Who is not an actor? **A:** Aaron Model: Mark
Coref.	***MFT:*** Simple coreference, he/she.	100.0	**C:** Melissa and Antonio are friends. He is a journalist, and she is an adviser. **Q:** Who is a journalist? **A:** Antonio Model: Melissa
Coref.	***MFT:*** Simple coreference, his/her.	100.0	**C:** Victoria and Alex are friends. Her mom is an agent **Q:** Whose mom is an agent? **A:** Victoria Model: Alex
Coref.	***MFT:*** former/latter	100.0	**C:** Kimberly and Jennifer are friends. The former is a teacher **Q:** Who is a teacher? **A:** Kimberly Model: Jennifer
SRL	***MFT:*** subject/object distinction	60.8	**C:** Richard bothers Elizabeth. **Q:** Who is bothered? **A:** Elizabeth Model: Richard
SRL	***MFT:*** subj/obj distinction with 3 agents	95.7	**C:** Jose hates Lisa. Kevin is hated by Lisa. **Q:** Who hates Kevin? **A:** Lisa Model: Jose

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

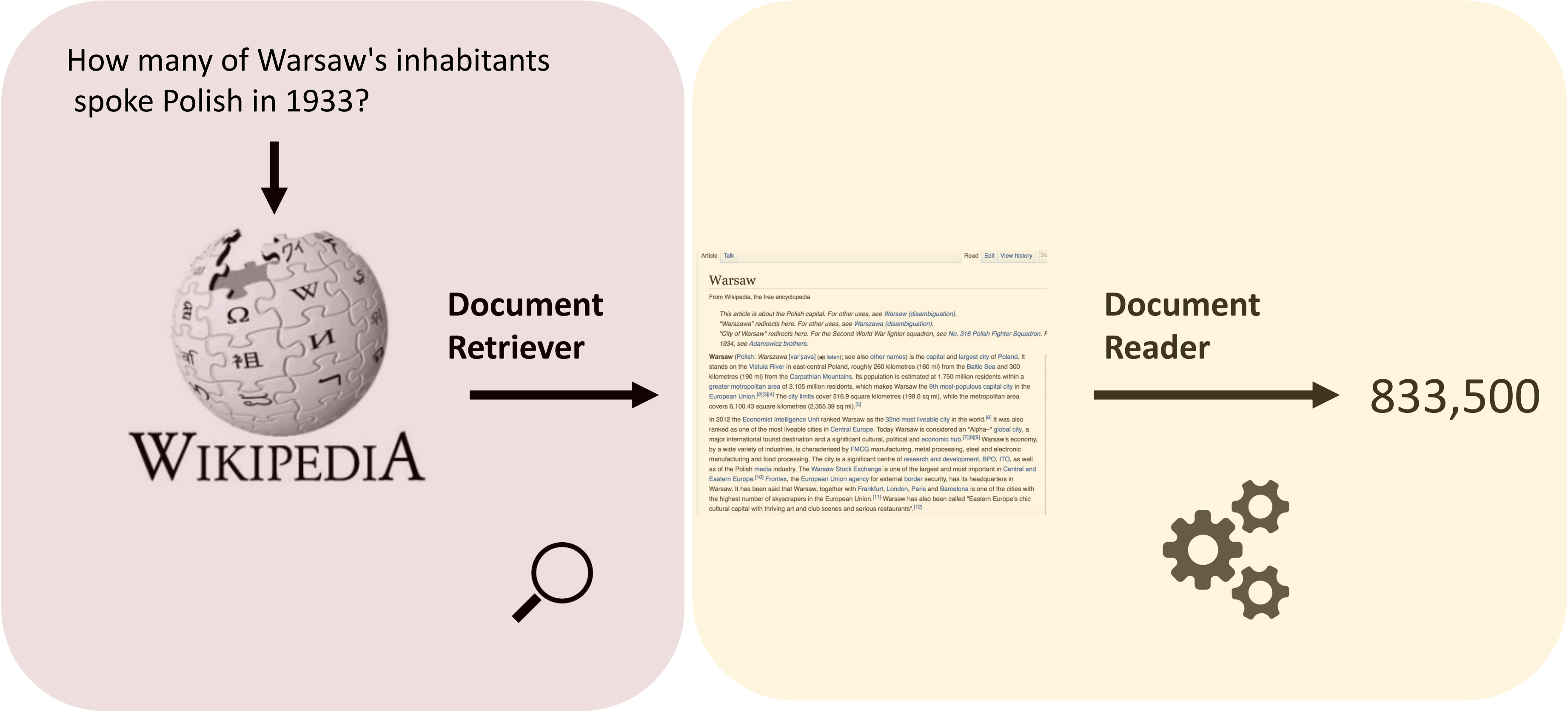
3. Open-domain question answering

- 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.

- Much more challenging and a more practical problem!



Retriever-Reader framework



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 $\mathcal{D} = D_1, D_2, \ldots, D_N$ and Q
- Output: an answer string A

- Retriever: $f(\mathcal{D}, Q) \longrightarrow P_1, \ldots, P_K$ K is pre-defined (e.g., 100)
- Reader: $g(Q, \{P_1, \ldots, P_K\}) \longrightarrow A$ A reading comprehension problem!

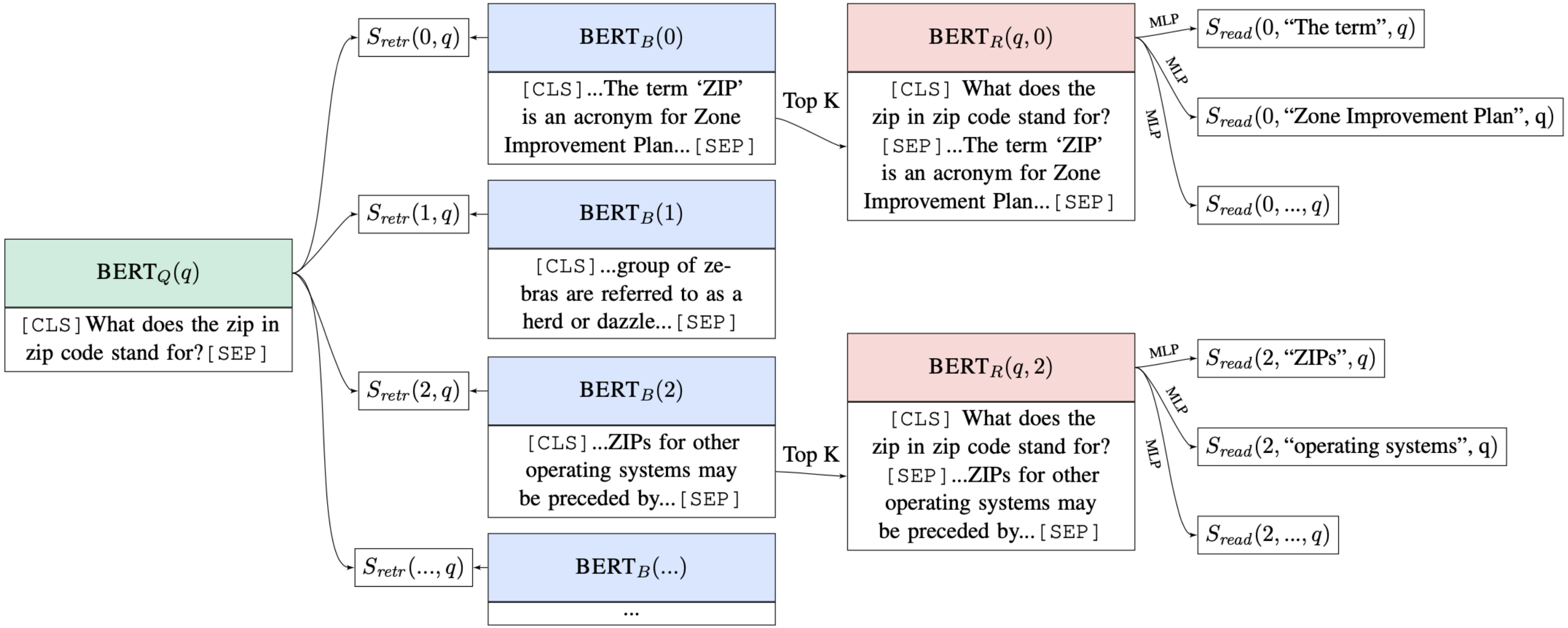
In DrQA,

- Retriever = A standard TF-IDF information-retrieval sparse model (a fixed module)
- Reader = a neural reading comprehension model that we just learned
  - Trained on SQuAD and other distantly-supervised QA datasets

Distantly-supervised examples: (Q, A) $\longrightarrow$ *(P, Q, A)*

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

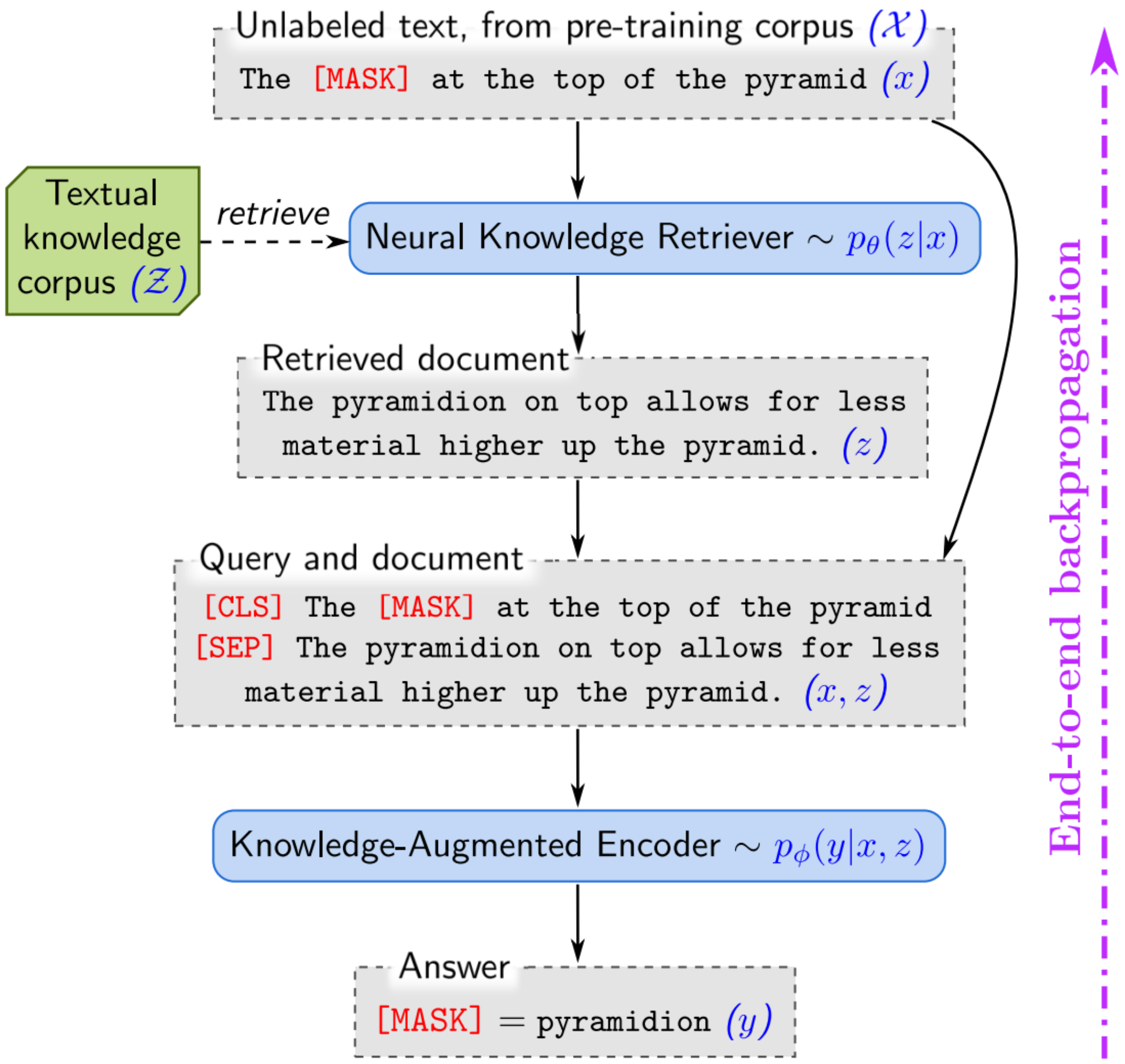
We can train the retriever! Joint training of retriever and reader



- Each text passage can be encoded as a vector using BERT and the retriever score can be measured as the dot product between the question representation and passage representation.
- However, it is not easy to model as there are a huge number of passages (e.g., 21M in English Wikipedia)

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

We can train the retriever! REALM: retrieval-augmented LM

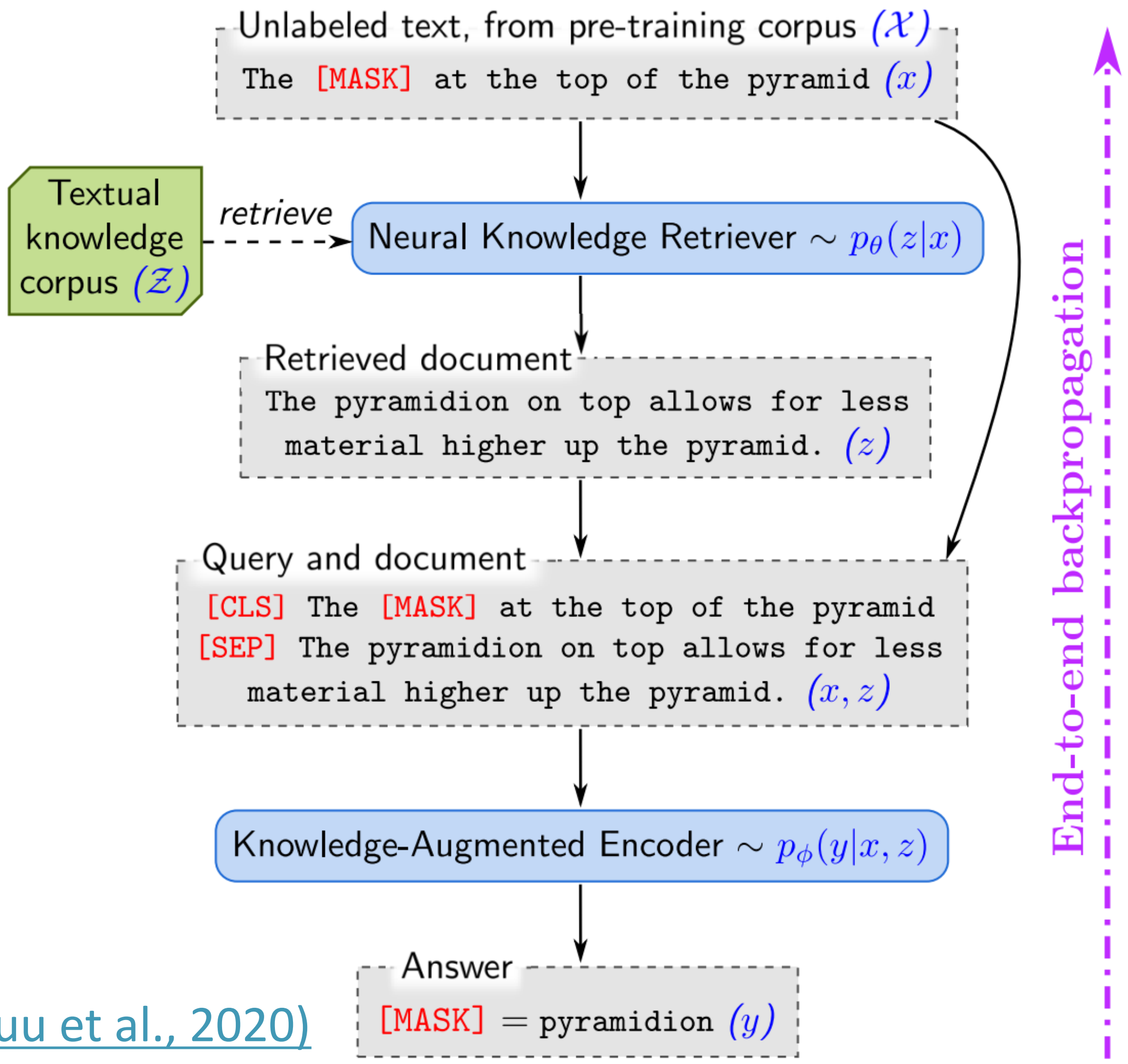


REALM augments language model pre-training with **a neural knowledge retriever** that retrieves knowledge from a **textual knowledge corpus** (e.g., all of Wikipedia).

Signal from the language modeling objective backpropagates all the way through the retriever, which must consider millions of documents in Z—a significant computational challenge

(Guu et al., 2020)

We can train the retriever! REALM: retrieval-augmented LM



(Guu et al., 2020)

Knowledge augmented encoder

$$p(y \mid x) = \sum_{z \in \mathcal{Z}} p(y \mid z, x)\, p(z \mid x)$$

Neural knowledge retriever

We can train the retriever! REALM: retrieval-augmented LM

$$p(y \mid z, x) = \prod_{j=1}^{J_x} p(y_j \mid z, x)$$

Knowledge augmented encoder

$$p(y \mid x) = \sum_{z \in \mathcal{Z}} p(y \mid z, x)\, p(z \mid x)$$

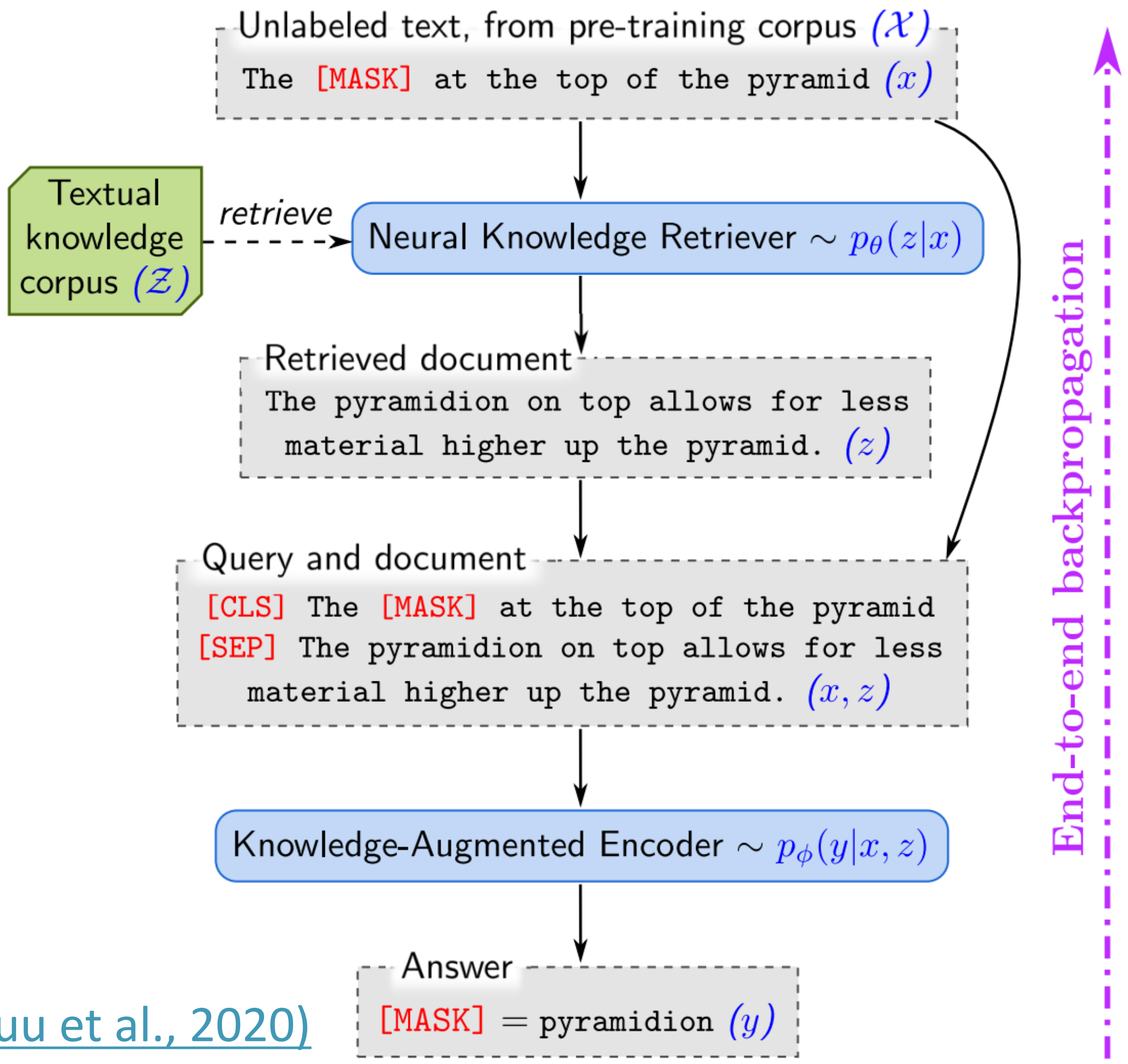
$$p(z \mid x) = \frac{\exp f(x, z)}{\sum_{z'} \exp f(x, z')},$$

$$f(x, z) = \texttt{Embed}_{\texttt{input}}(x)^\top \texttt{Embed}_{\texttt{doc}}(z)$$

Neural knowledge retriever

(Guu et al., 2020)

We can train the retriever! REALM: retrieval-augmented LM



[(Guu et al., 2020)](#)

Knowledge augmented encoder

$$p(y \mid x) = \sum_{z \in \mathcal{Z}} p(y \mid z, x)\, p(z \mid x)$$

- Key computational challenge is that the marginal probability involves a summation over all docs.
- Approximate this by summing over top k docs with highest prob under $p(z|x)$ --- via Maximum inner product search (MIPS)

We can train the retriever! REALM: retrieval-augmented LM

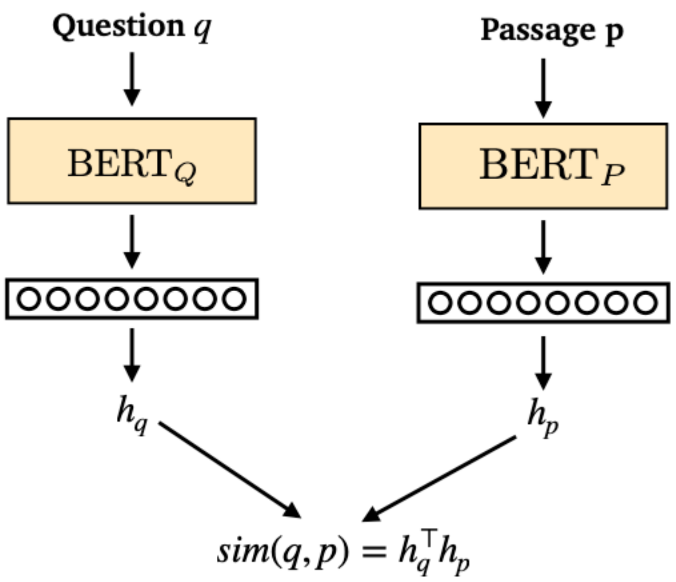
Name	Architectures	Pre-training	NQ (79k/4k)	WQ (3k/2k)	CT (1k /1k)	# params
BERT-Baseline (Lee et al., 2019)	Sparse Retr.+Transformer	BERT	26.5	17.7	21.3	110m
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1	-	223m
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2	-	738m
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4	-	11318m
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7	25.7	34m
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-	110m
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-	110m
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-	110m
ORQA (Lee et al., 2019)	Dense Retr.+Transformer	ICT+BERT	33.3	36.4	30.1	330m
ORQA (more fine-tune epochs)	Dense Retr.+Transformer	ICT+BERT	34.8	35.4	28.7	330m
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2	**46.8**	330m
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**	42.9	330m

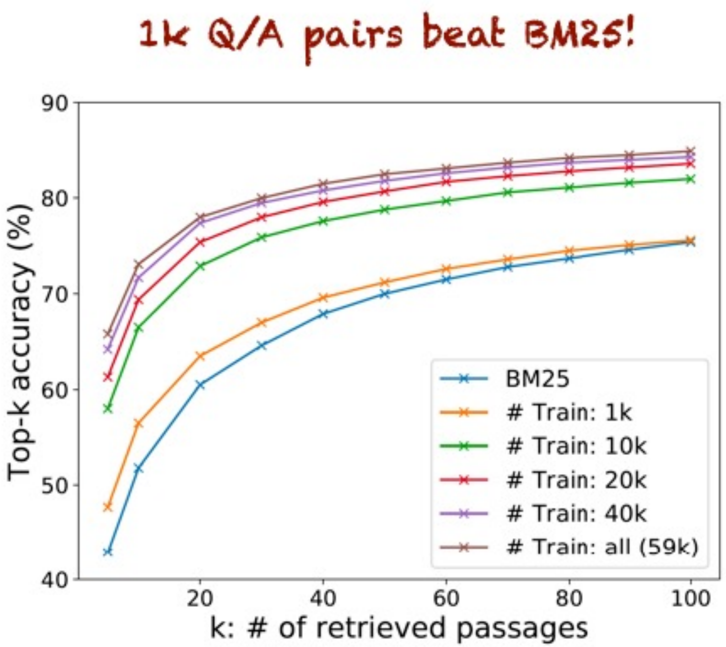
Test results on Open-QA benchmarks: NaturalQuestions, WebQuestions, CuratedTrec dataset

(Guu et al., 2020)

We can train the retriever

- Dense passage retrieval (DPR)
- We can also just train the retriever using question-answer pairs!



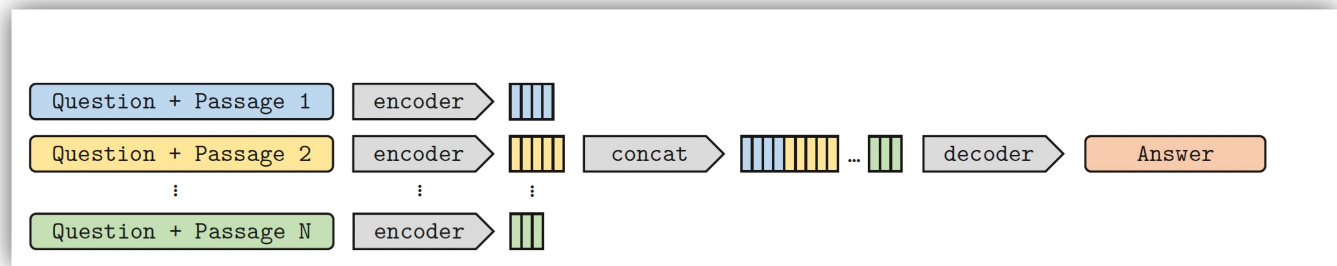


- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models

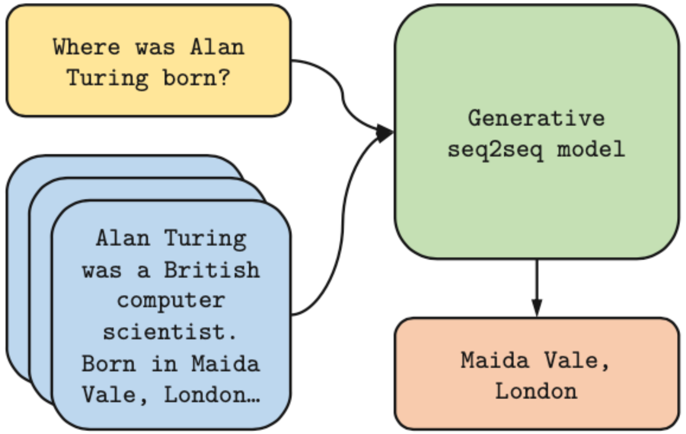
Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Deep retrieval + generative models

- Recent work shows that it is beneficial to generate answers instead of to extract answers.



Fusion-in-decoder (FID)=DPR+T5

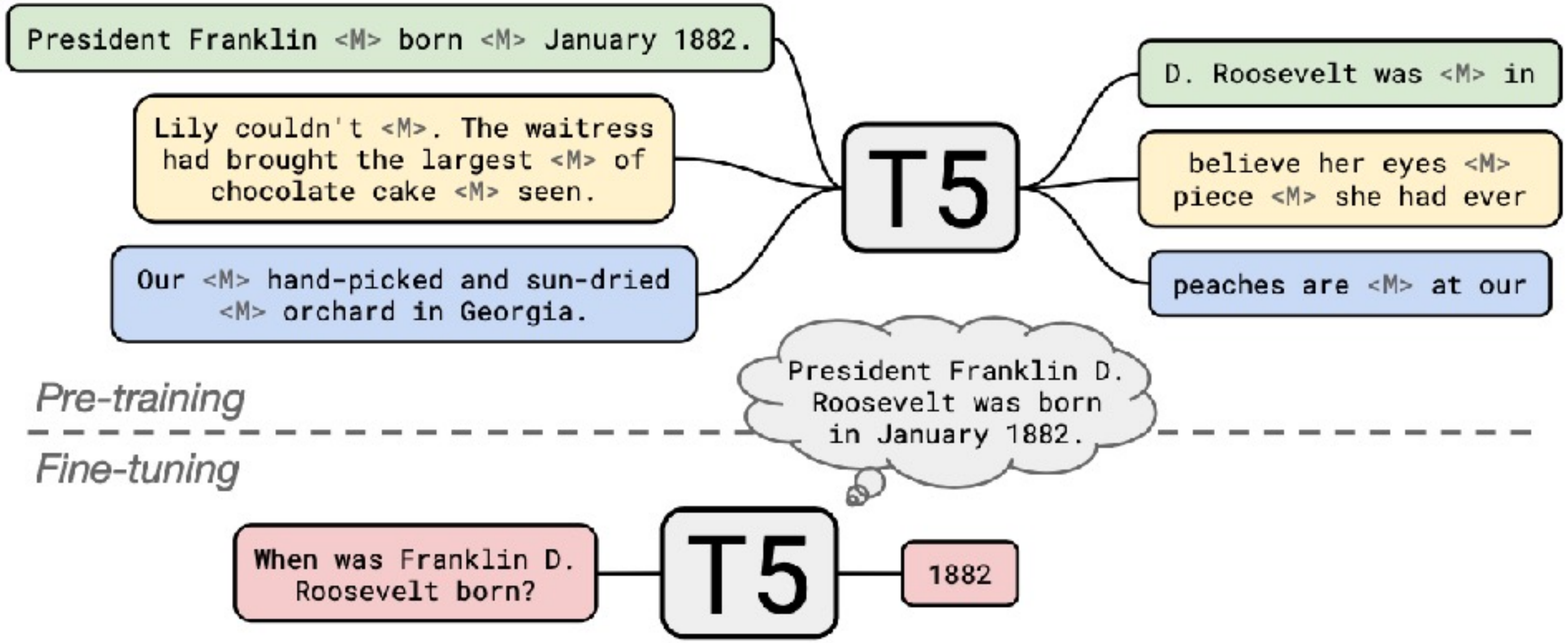


Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2
Fusion-in-Decoder (base)	48.2	65.0	77.1
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**

Izacard and Grave 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering

LLMs can do open-domain QA well

- … without an explicit retriever stage

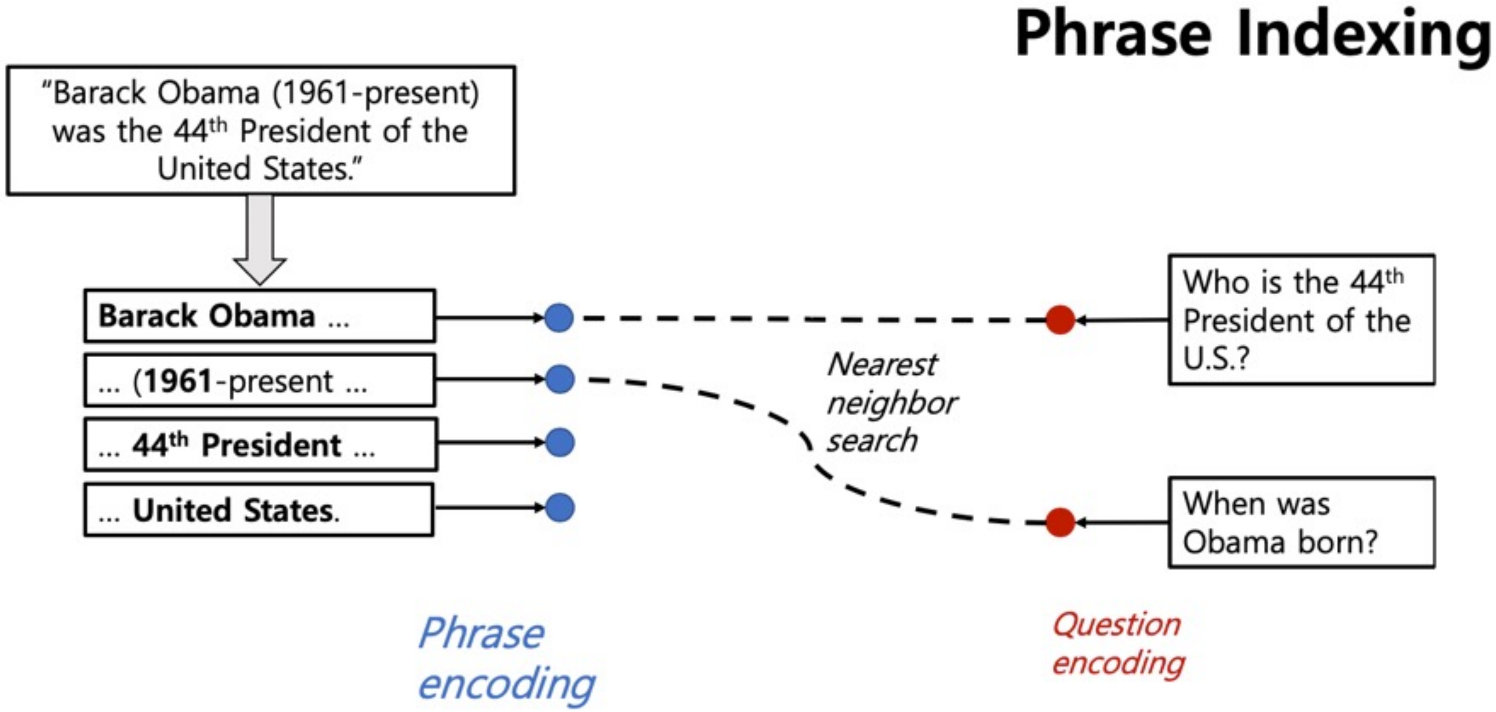


	NQ	WQ	TQA dev	TQA test
Chen et al. (2017)	–	20.7	–	–
Lee et al. (2019)	33.3	36.4	47.1	–
Min et al. (2019a)	28.1	–	50.9	–
Min et al. (2019b)	31.8	31.6	55.4	–
Asai et al. (2019)	32.6	–	–	–
Ling et al. (2020)	–	–	35.7	–
Guu et al. (2020)	40.4	40.7	–	–
Févry et al. (2020)	–	–	43.2	53.4
Karpukhin et al. (2020)	**41.5**	42.4	**57.9**	–
T5-Base	25.9	27.9	23.8	29.1
T5-Large	28.5	30.6	28.7	35.9
T5-3B	30.4	33.6	35.1	43.4
T5-11B	32.6	37.2	42.3	50.1
T5-11B + SSM	34.8	40.8	51.0	60.5
T5.1.1-Base	25.7	28.2	24.2	30.6
T5.1.1-Large	27.3	29.5	28.5	37.2
T5.1.1-XL	29.5	32.4	36.0	45.1
T5.1.1-XXL	32.8	35.6	42.9	52.5
T5.1.1-XXL + SSM	35.2	**42.8**	51.9	**61.6**

Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Maybe the reader model is not necessary too!

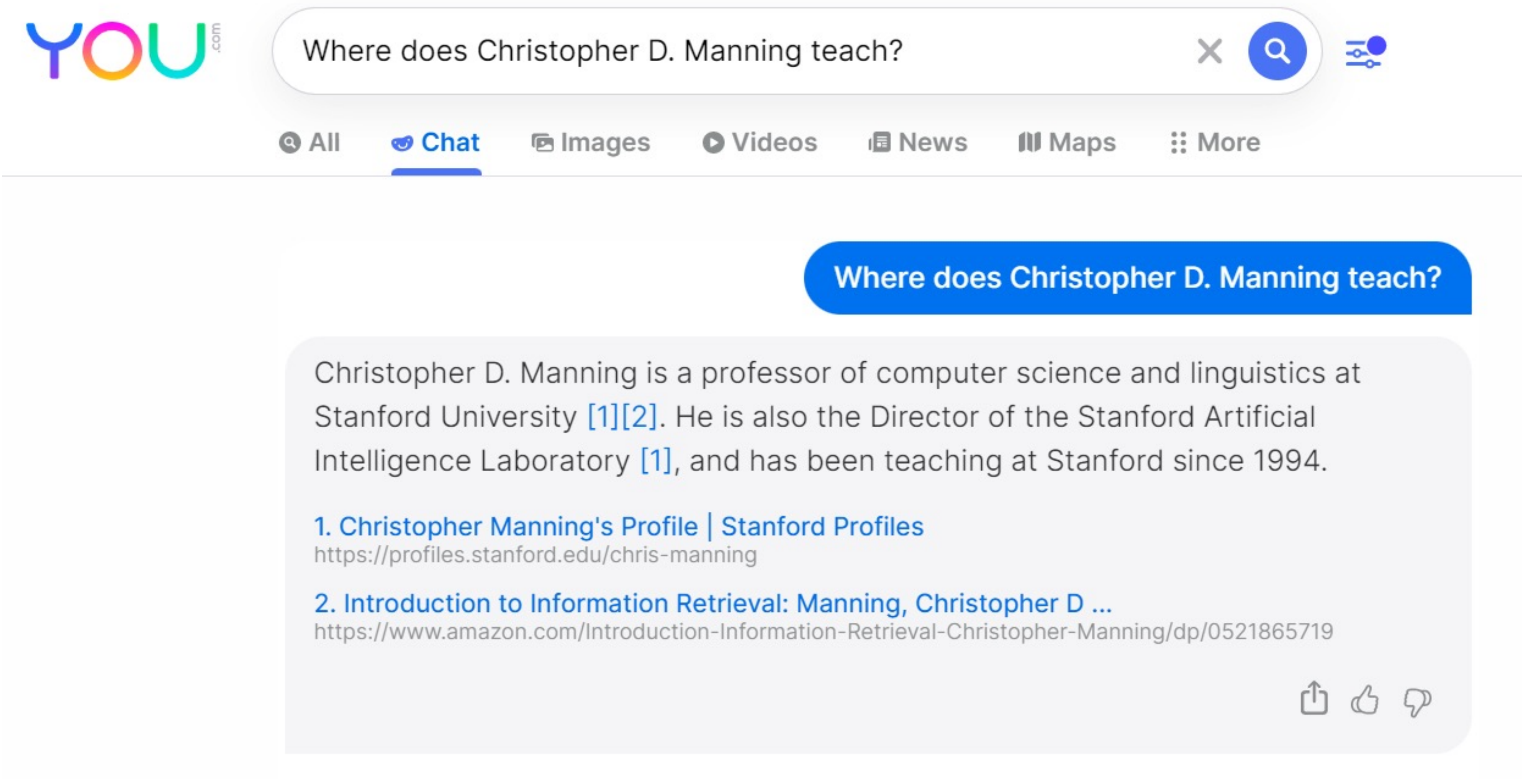
- It is possible to encode all the phrases (60 billion phrases in Wikipedia) using dense vectors and only do nearest neighbor search without a BERT model at inference time!



Seo et al., 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index

Lee et al., 2020. Learning Dense Representations of Phrases at Scale

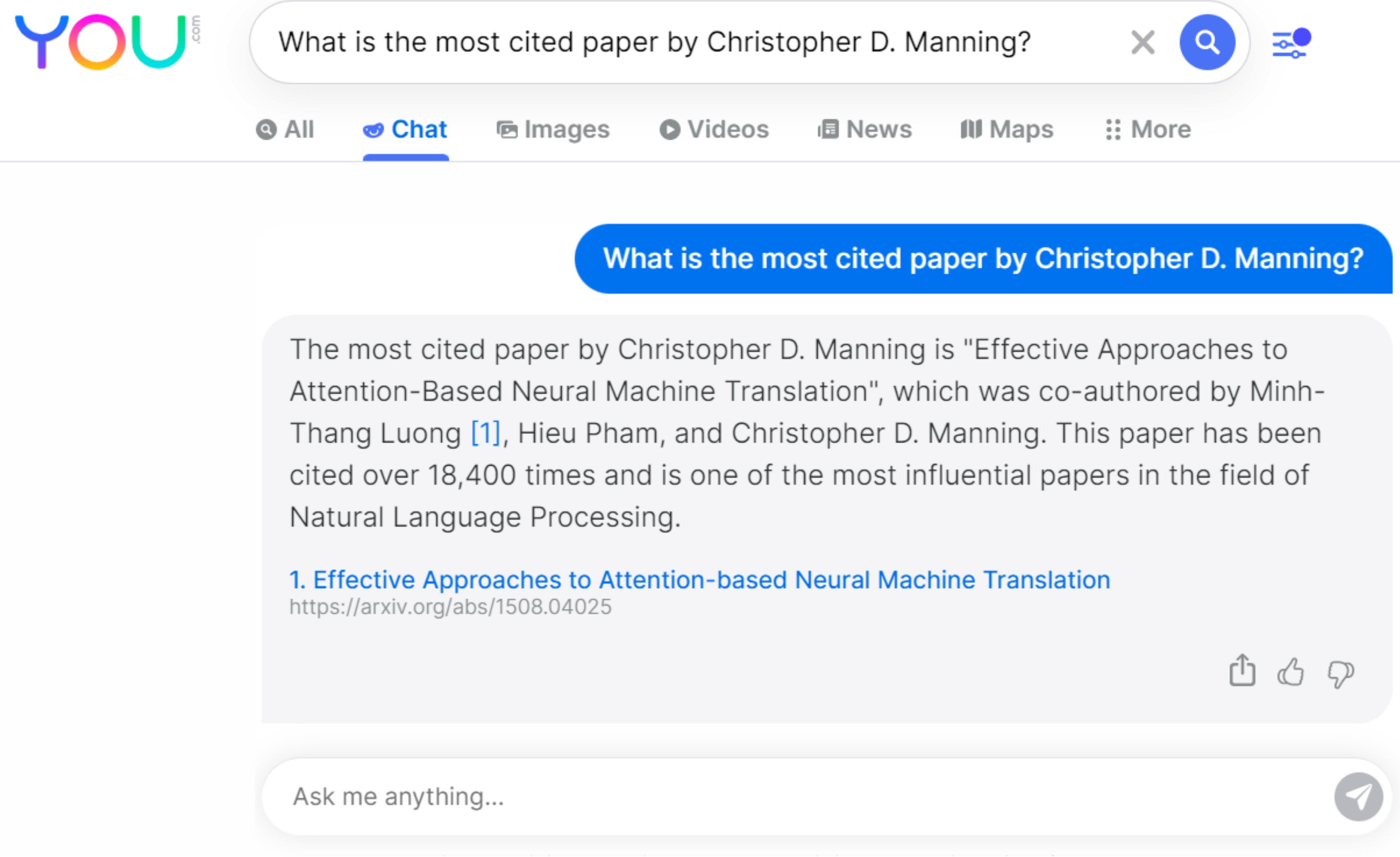
LLM based QA (with web search!)



Problem with LLM based QA

- Seems totally reasonable!

- But (1) it's not his most cited paper, ar (2) it doesn't have that many citations. Yikes! Also the reference to a web page doesn't help.



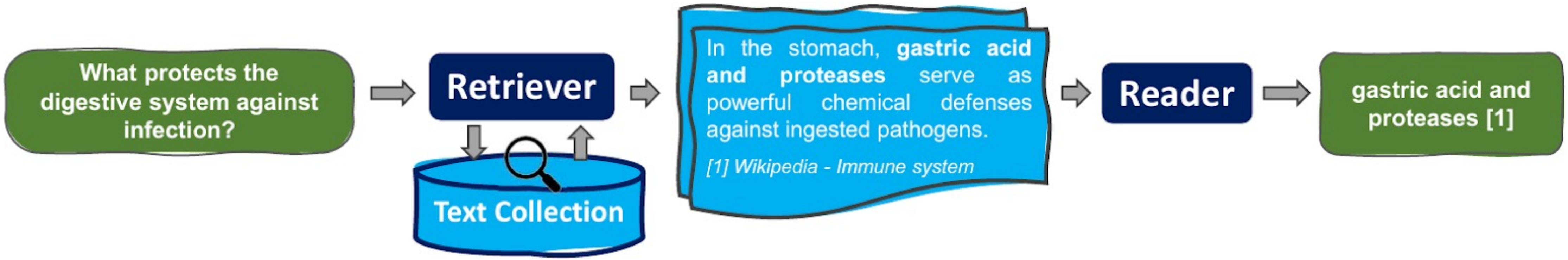
LLMs cannot memorize everything

- LLMs store an impressive amount of information in their parameters
  - But LMs can’t memorize everything
  - The world changes over time
  - You might want it to use private documents that aren’t on the web

- Also, black-box LLMs are opaque
  - Given a query, it produces an answer, but it’s difficult to verify if the answer is correct

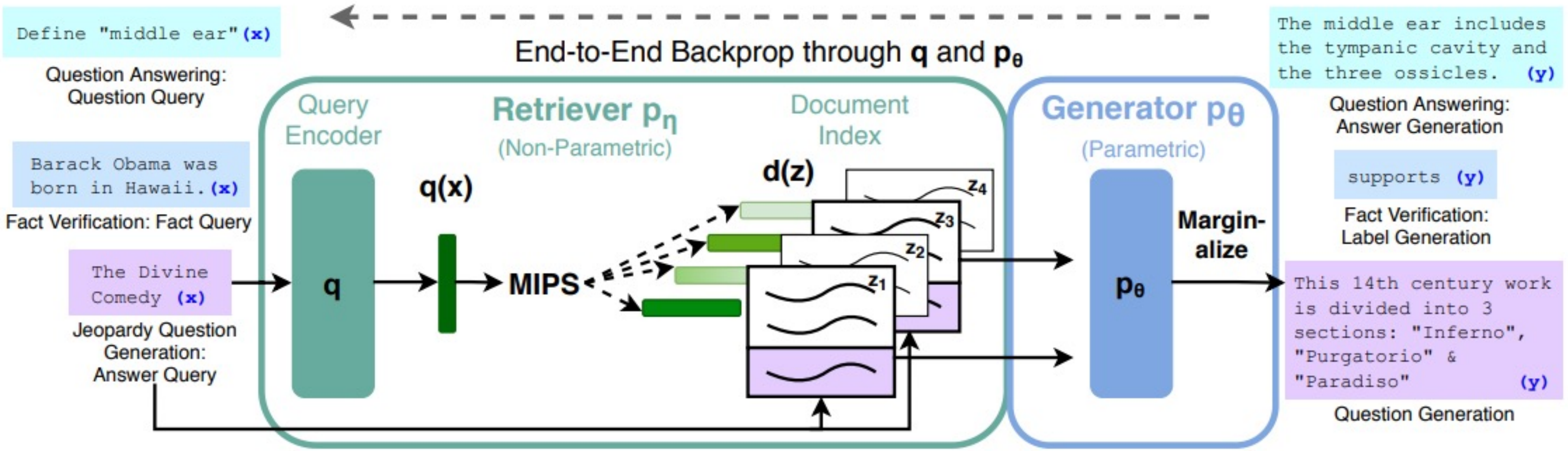
Using retrieval to overcome LLMs' shortcomings

- Instead of asking the LLM to memorize everything, can we provide the LLM with relevant and useful content just-in-time?
- **Retrieval** is a common mechanism for identifying such relevant information.
  - **Dynamic**: it's easy to update / add documents to your retrieval system
  - **Interpretable**: LM can generate pointers to retrieved documents that support human verification of its generations (citations)



4: Retrieval Augmented Generation (RAG)

- Retrieval Augmented Generation (RAG) is very powerful!



[Piktus+ 2021]

4: Retrieval Augmented Generation (RAG)

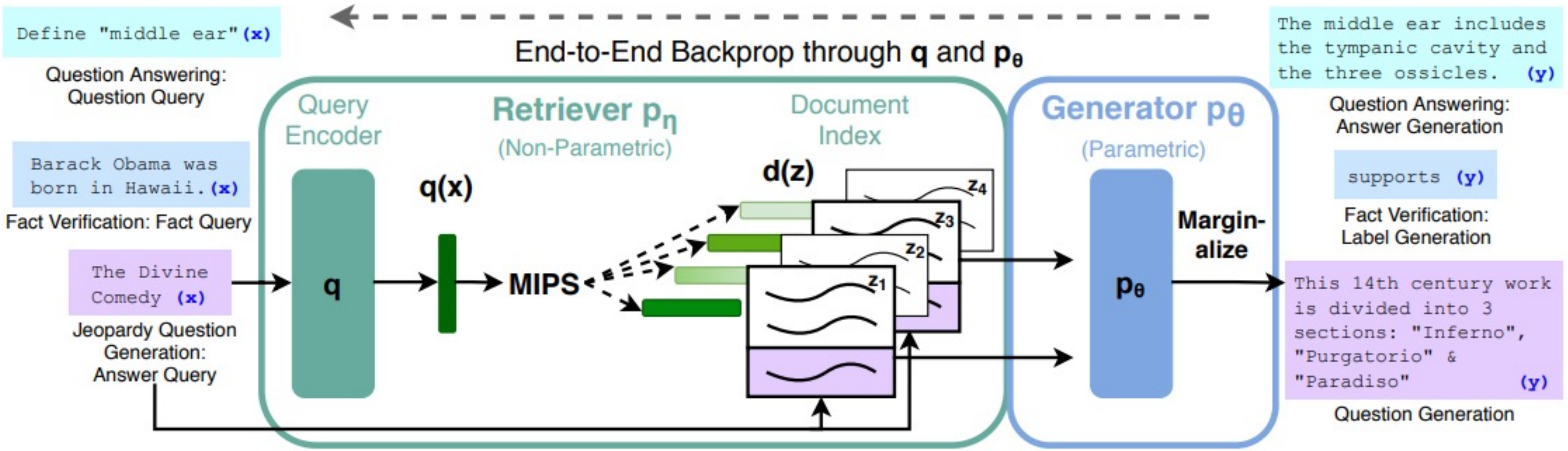
- State of the art for open-domain QA
  - Combine strengths of Open (non-parametric) and Closed-book (parametric)
- More specific, diverse, and factual generation than seq2seq (e.g., BART)

	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

[Piktus+ 2021]

4: Retrieval Augmented Generation (RAG)

How well does RAG work, anyway? Is it more factual?



[Piktus+ 2021]

Problem #1: how many documents can use?

- The retriever is key: if we have to use only 1 document, then we have to get that right.
- Why not get lots of documents and pass it to the LM?

Input Context

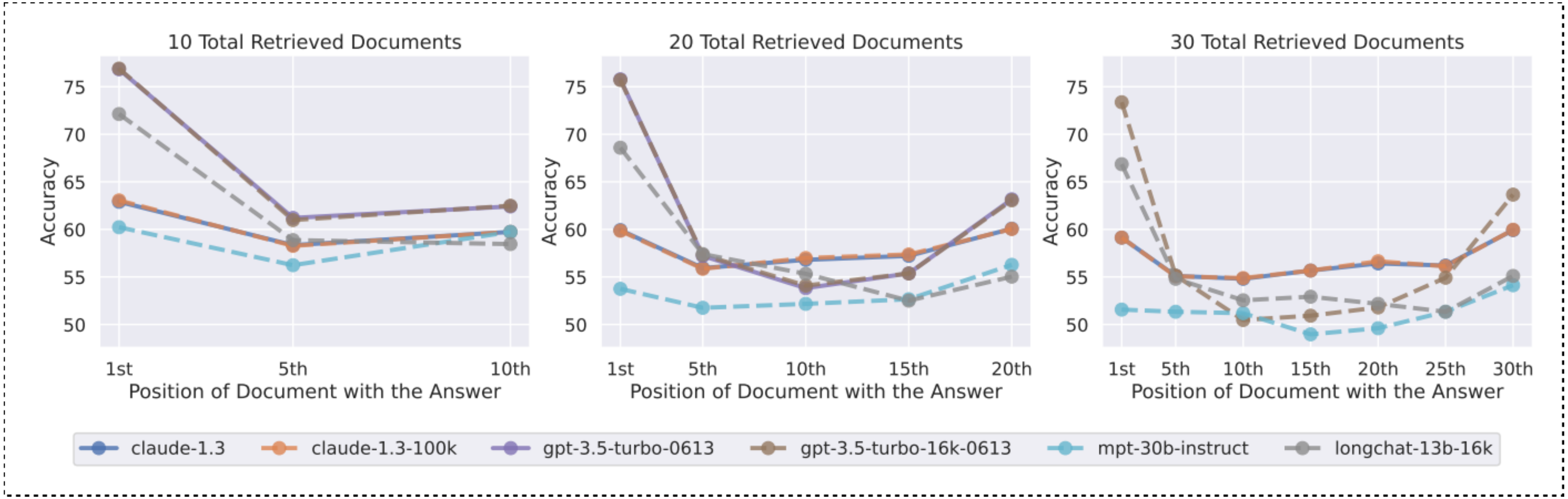
```
Write a high-quality answer for the given question
using only the provided search results (some of
which might be irrelevant).

Document [1](Title: List of Nobel laureates in
Physics) ...
Document [2](Title: Asian Americans in science and
technology) ...
Document [3](Title: Scientist) ...

Question: who got the first nobel prize in physics
Answer:
```

Desired Answer

```
Wilhelm Conrad Röntgen
```

LLM's can't pay attention to the entire context

- The long-context problem bites you – LLMs do not pay attention to its context well!
- Setup: 1 relevant document, all others irrelevant

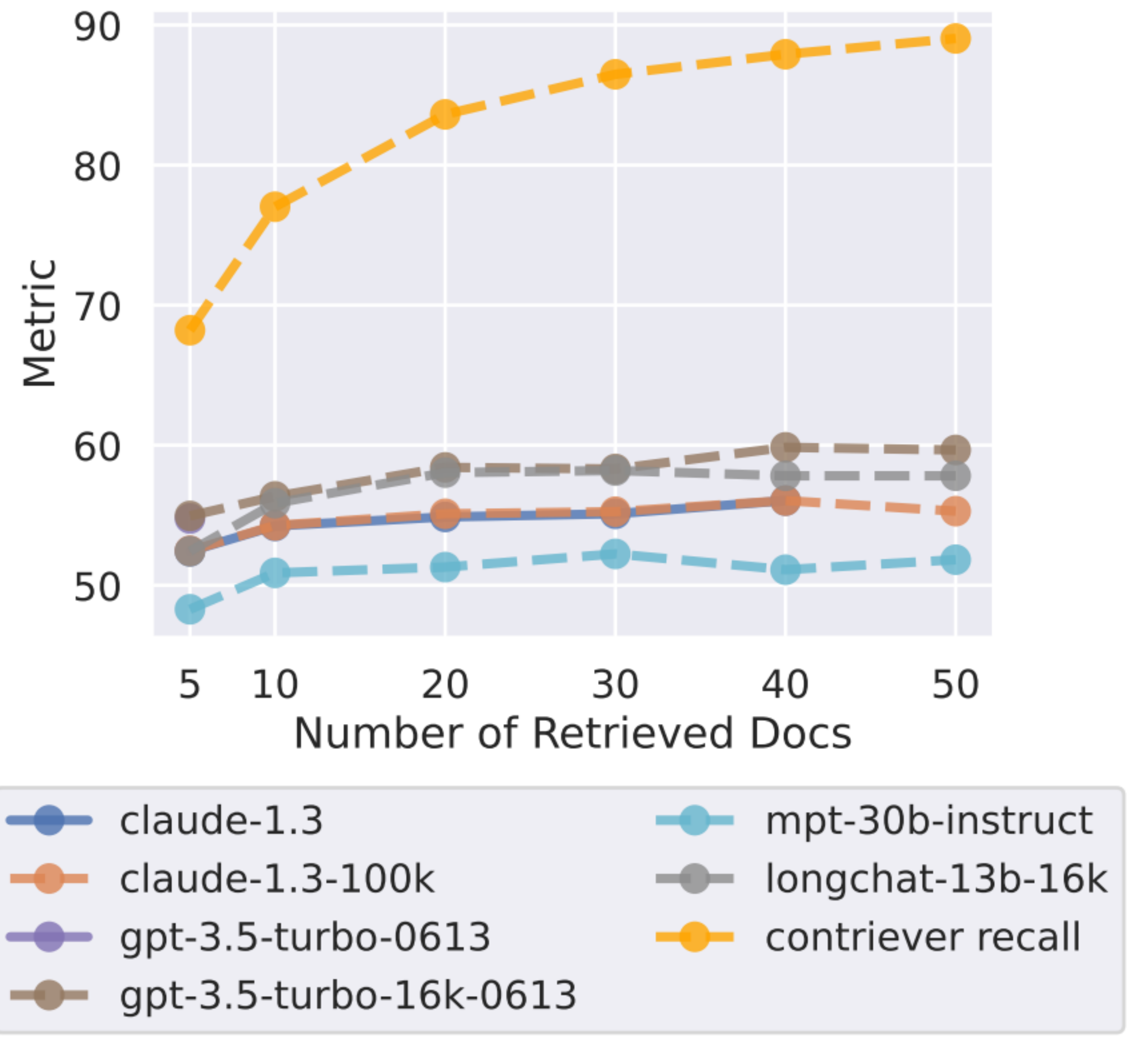


Best Closed-Book performance: GPT-3.5-Turbo, ~56%
Best Oracle (only feed in relevant doc) performance: GPT-3.5-Turbo, ~88.5%

[Liu+ 2023]

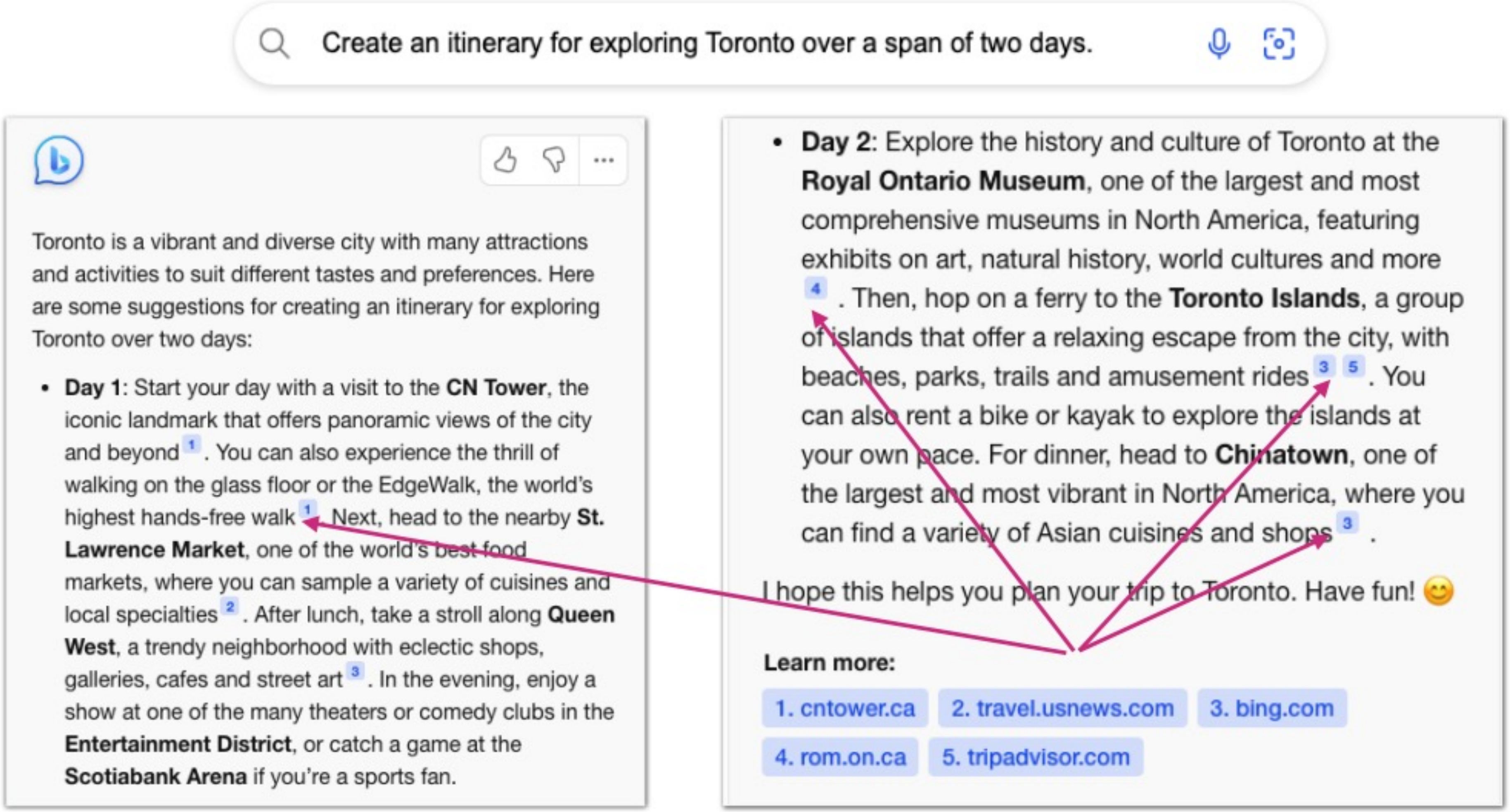
In practice: LLMs can't use many documents

- Retriever performance (yellow) rises slowly to 90% recall
- RAG performance (other lines) saturate very quickly – after 10-20 documents.



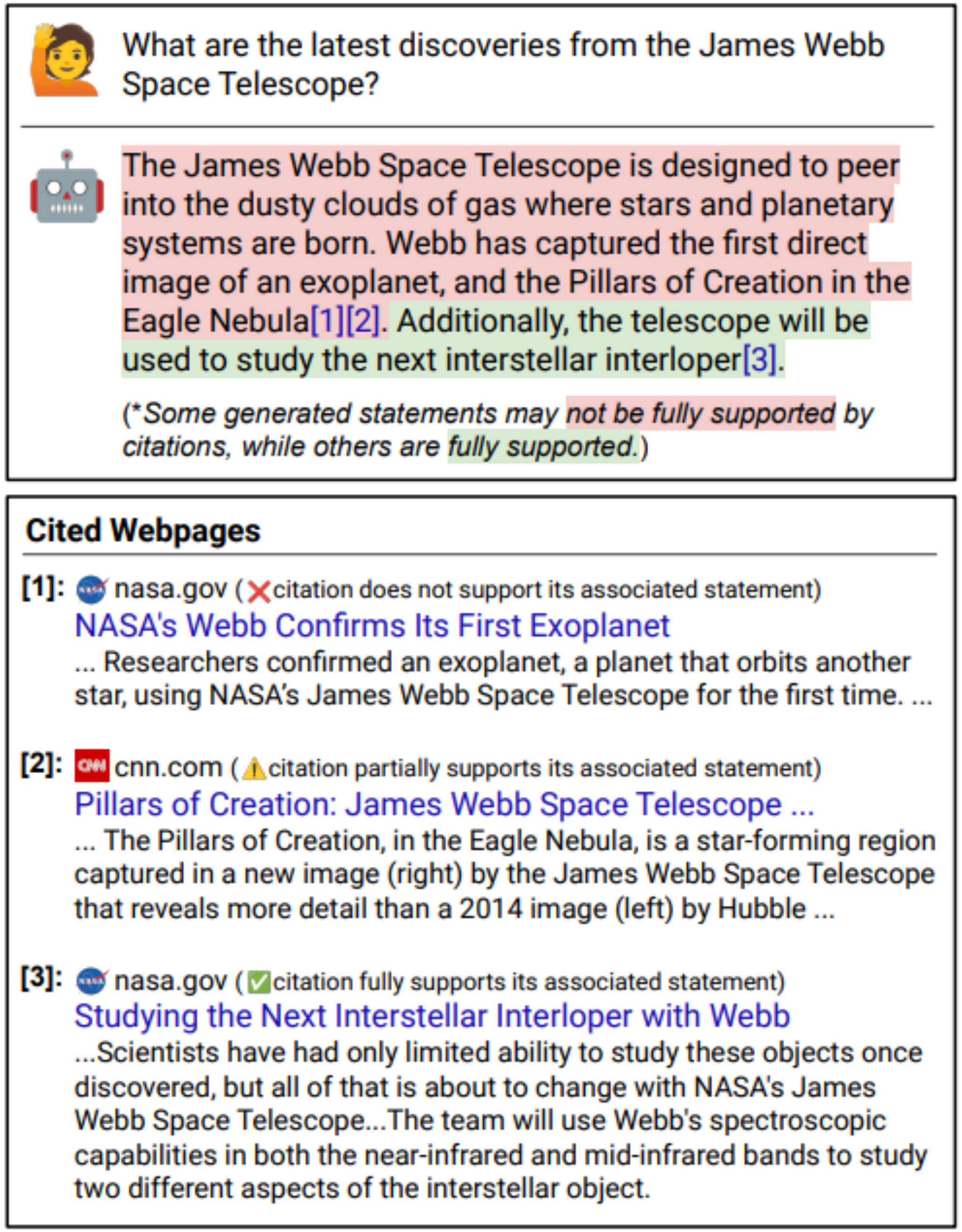
Problem #2: useful and usable citations

- A unique benefit of RAG: citing your sources



How accurate are these citations?

- These citations themselves are generated by LLMs
- So the citations could also be hallucinated .. How often does that happen?



Existing systems have high fluency, but low correctness for citations

- Outputs are easy to read / appear useful to rater (1-5 scale)

Perceived Utility (↑)	Average Over All Queries
Bing Chat	4.34
NeevaAI	4.48
perplexity.ai	4.56
YouChat	4.62
Average	4.50

Fluency (↑)	Average Over All Queries
Bing Chat	4.40
NeevaAI	4.43
perplexity.ai	4.51
YouChat	4.59
Average	4.48

[Liu+ 2023]

Existing systems have high fluency, but low correctness for citations

- But precision and recall are both low …

Citation Precision (%; ↑)	Average Over All Queries
Bing Chat	89.5
NeevaAI	72.0
perplexity.ai	72.7
YouChat	63.6
Average	74.5

Citation Recall (%; ↑)	Average Over All Queries
Bing Chat	58.7
NeevaAI	67.6
perplexity.ai	68.7
YouChat	11.1
Average	51.5

Overview

1. What is question answering?
2. Reading comprehension
   - ✓ How to answer questions over **a single passage of text**
3. Open-domain (textual) question answering
   - ✓ How to answer questions over **a large collection of documents**
4. Retrieval-augmented generation for question answering