

Natural Language Processing with Deep Learning

CS224N/Ling284



Tatsunori Hashimoto

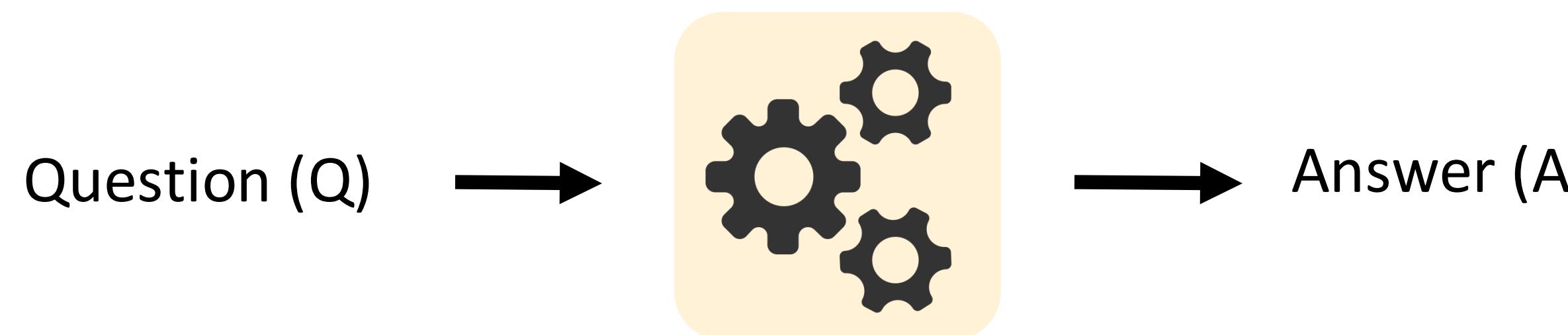
Lecture 13: Question Answering

Lecture plan

- What's QA? Why do we care?
- “Parametric” knowledge and why this is interesting
- Reading comprehension and RAG systems

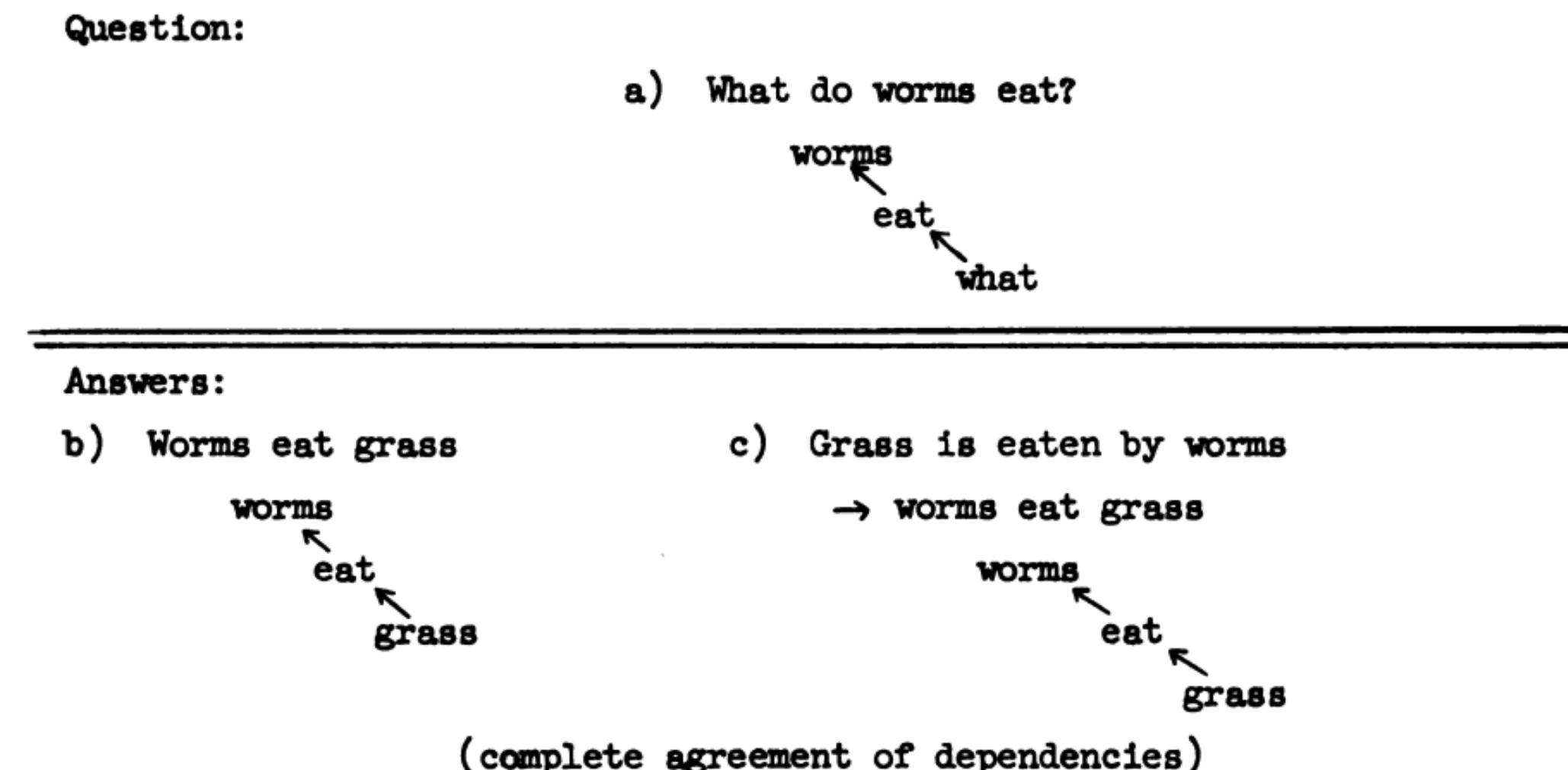
Notes: Guest lecture (Sasha Rush) Thursday!

What is question answering?

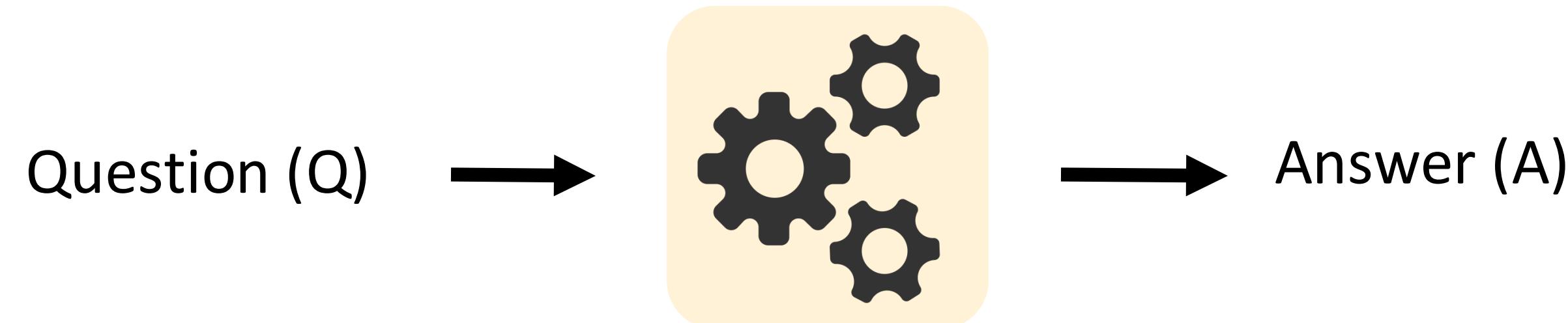


The goal of question answering is to build systems that **automatically** answer questions posed by humans in a **natural language**

The earliest QA systems dated back to 1960s!
(Simmons et al., 1964)



Question answering: a taxonomy



- 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

A screenshot of a Google search results page. The search bar at the top contains the query "Where is the deepest lake in the world?". Below the search bar are navigation links for "All", "Maps", "Images", "News", "Videos", and "More", with "All" being the selected tab. To the right of these are "Settings" and "Tools" buttons. A message indicates "About 21,100,000 results (0.71 seconds)". Below this, there are four thumbnail images: a landscape view of a lake, a 3D cross-section diagram of Lake Baikal, another landscape view of a lake, and a satellite map of a lake.

Siberia

Lake **Baikal**, in Siberia, holds the distinction of being both the deepest lake in the world and the largest freshwater lake, holding more than 20% of the unfrozen fresh water on the surface of Earth.

Lots of practical applications

The screenshot shows a Google search results page. The search query "How can I protect myself from COVID-19?" is entered in the search bar. Below the search bar, there are navigation links for All, Images, News, Shopping, Videos, More, Settings, and Tools. A snippet of text from a CDC article is displayed, followed by a bulleted list of 13 actions to prevent COVID-19 spread. A "Learn more on cdc.gov" button is at the bottom, and a footer note about informational purposes is at the very bottom.

How can I protect myself from COVID-19?

All Images News Shopping Videos More Settings Tools

The best way to prevent illness is to avoid being exposed to this virus. Learn how COVID-19 spreads and practice these actions to help prevent the spread of this illness.

To help prevent the spread of COVID-19:

- Cover your mouth and nose with a mask when around people who don't live with you. Masks work best when everyone wears one.
- Stay at least 6 feet (about 2 arm lengths) from others.
- Avoid crowds. The more people you are in contact with, the more likely you are to be exposed to COVID-19.
- Avoid unventilated indoor spaces. If indoors, bring in fresh air by opening windows and doors.
- Clean your hands often, either with soap and water for 20 seconds or a hand sanitizer that contains at least 60% alcohol.
- Get vaccinated against COVID-19 when it's your turn.
- Avoid close contact with people who are sick.
- Cover your cough or sneeze with a tissue, then throw the tissue in the trash.
- Clean and disinfect frequently touched objects and surfaces daily.

Learn more on [cdc.gov](#)

For informational purposes only. Consult your local medical authority for advice.

2011: IBM Watson beat Jeopardy champions



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

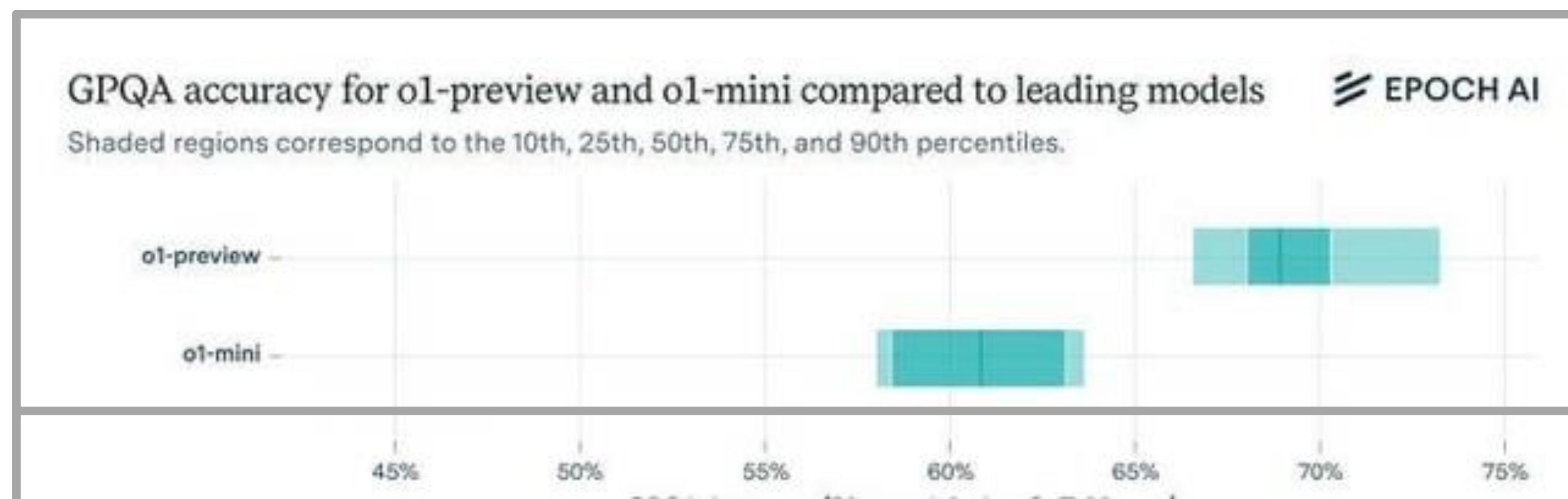
2023-2025: ‘Expert level’ QA tasks

GPQA (PhD level science questions)

Organic Chemistry

Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed (as a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)?

- A) 2
- B) 16
- C) 8
- D) 4



Expert level QA tasks

“Humanity’s last exam”

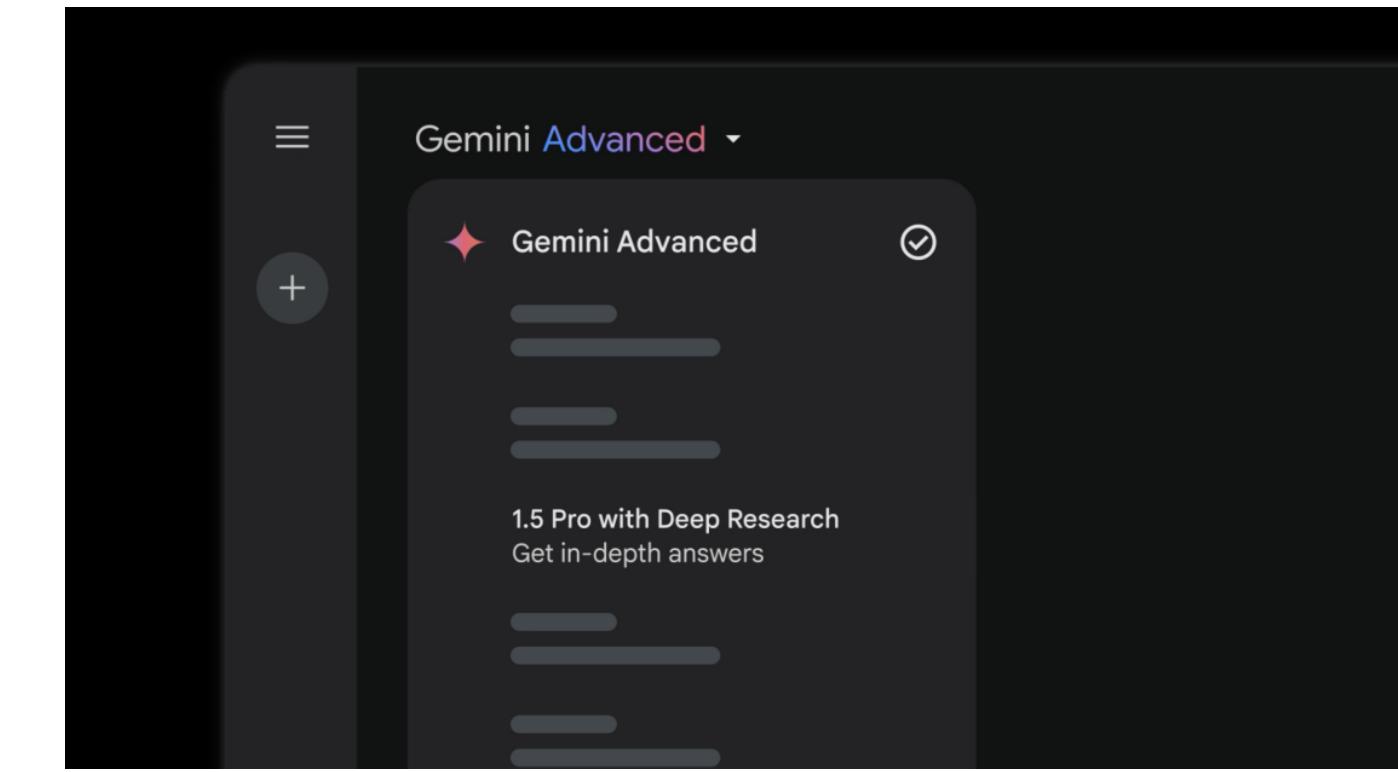
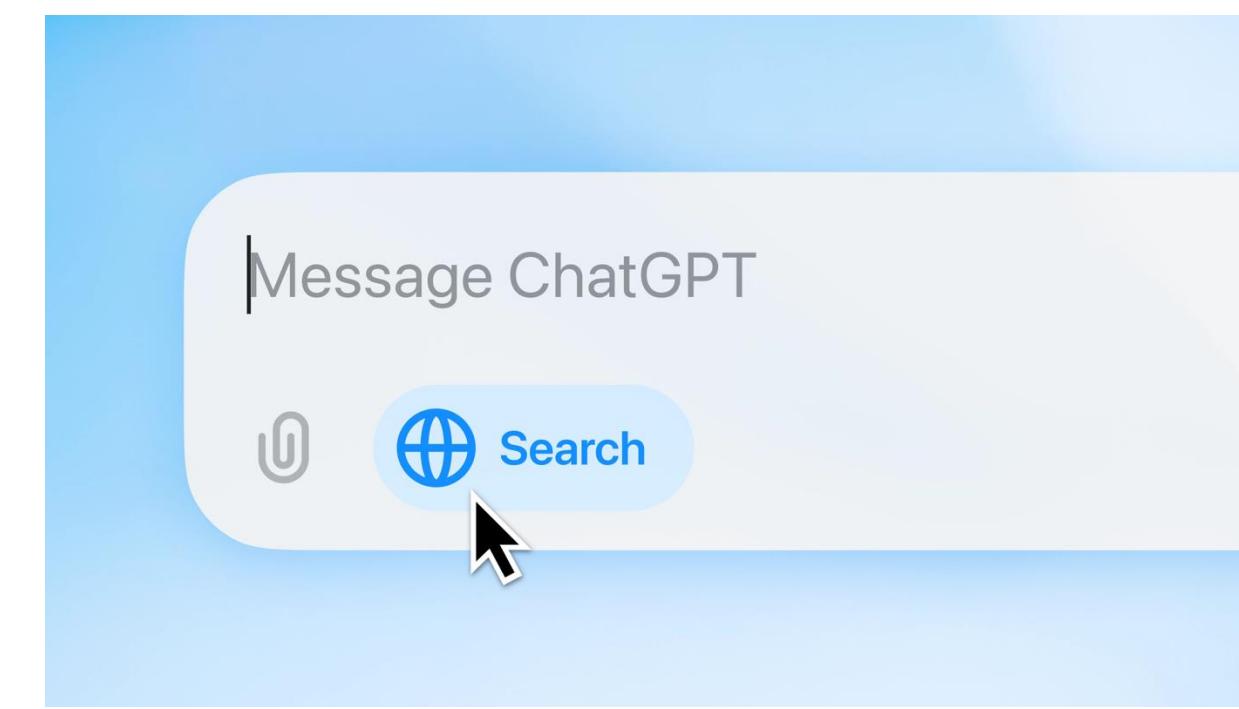
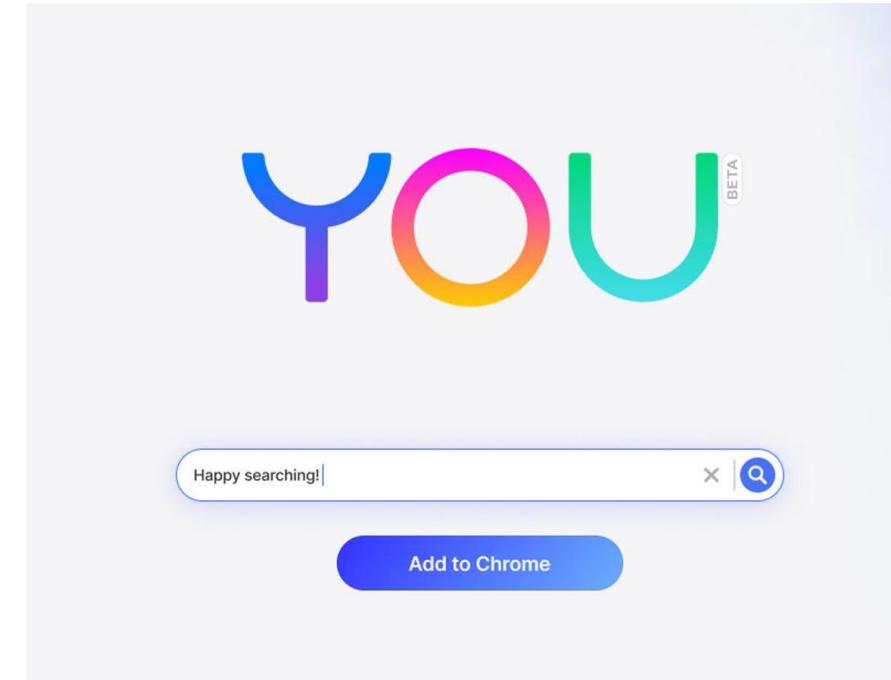
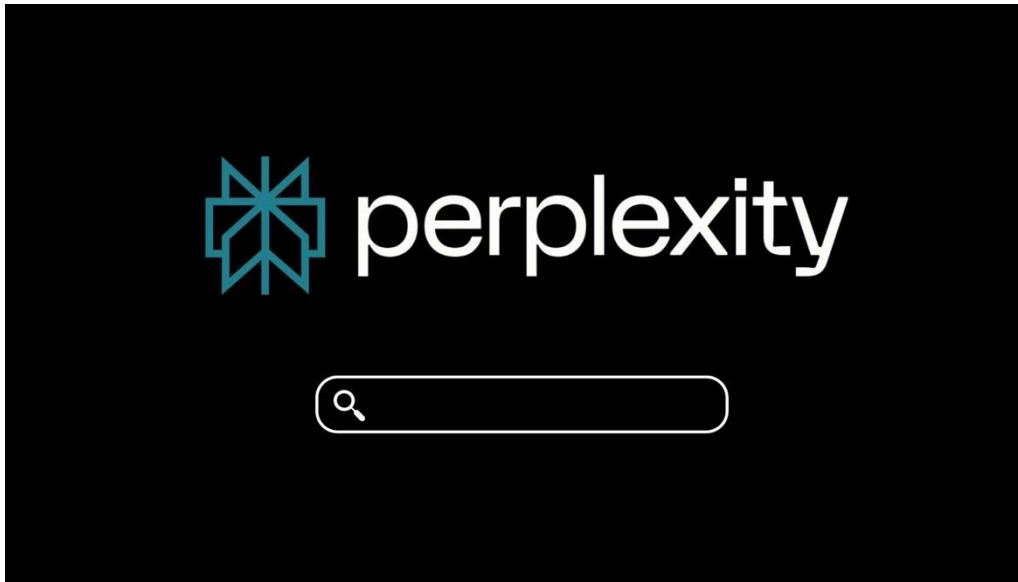
Ecology

Question:

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

Model	Accuracy (%) ↑
GPT-4o	3.1
Grok-2	3.9
Claude 3.5 Sonnet	4.8
Gemini Thinking	7.2
o1	8.8
DeepSeek-R1*	8.6
o3-mini (medium)*	11.1
o3-mini (high)*	14.0

QA remains a huge area of interest



Many products, companies, focusing on QA

Different types of QA

Closed book

- ‘parametric’ QA

Open book

- Reading comprehension
- RAG

Two different settings with different things to think about

Closed book QA

Closed book – answering the question with *no access to external information*

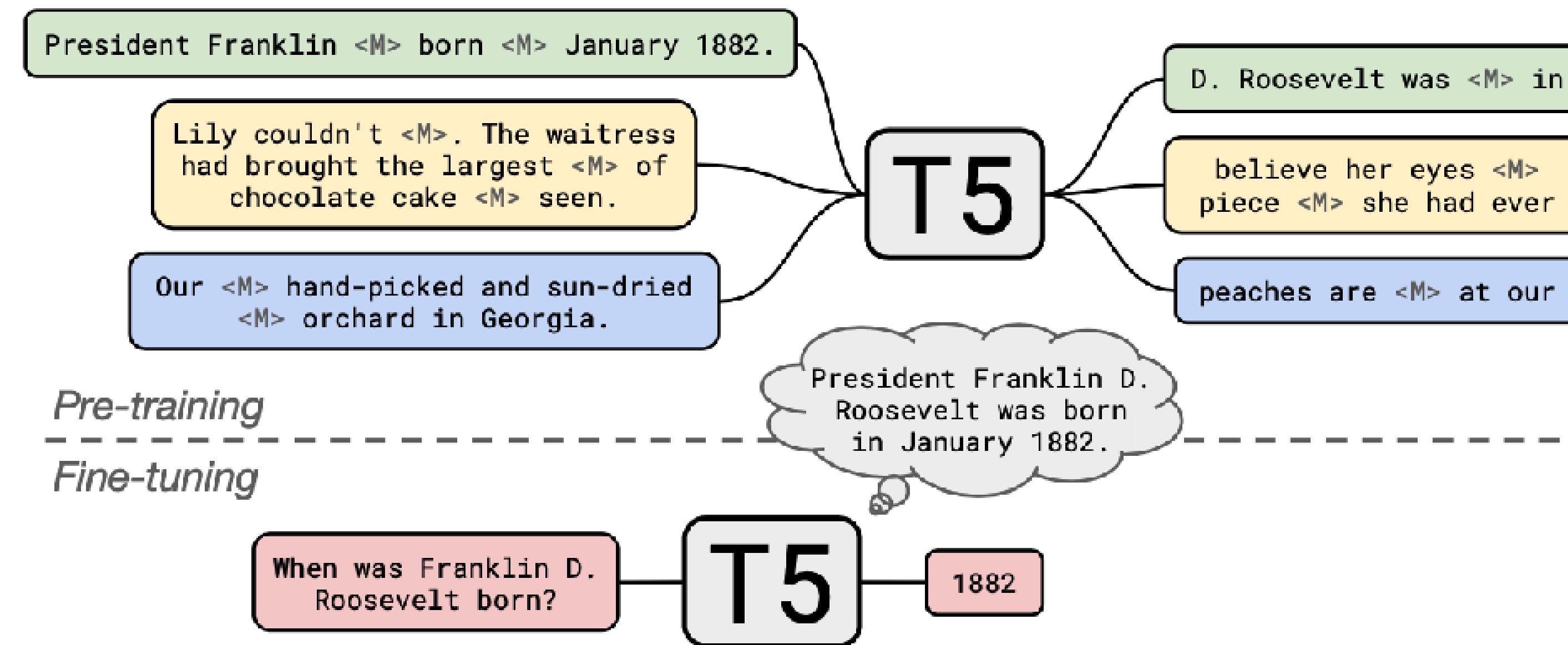
Q&A done just by ‘predicting the next word’

Stanford University is located in _____

When does this work? Why do we care about this setting?

Large language models can do open-domain QA well

Initially – for models like T5, people discovered LMs can directly answer questions



LMs have remarkable knowledge about the world

Closed-book QA then took on an important role as a *capability test* ..

MMLU

Astronomy

What is true for a type-Ia supernova?

- A. This type occurs in binary systems.
- B. This type occurs in young galaxies.
- C. This type produces gamma-ray bursts.
- D. This type produces high amounts of X-rays.

Answer: A

High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

- A. directional selection.
- B. stabilizing selection.
- C. sexual selection.
- D. disruptive selection

Answer: A

Closed book QA as a capability test

Many popular benchmarks (ARC-C, MMLU, GPQA, etc) are partially knowledge tests

Scenario	Task	What	Who
NarrativeQA narrative_qa	short-answer question answering	passages are books and movie scripts, questions are unknown	annotators from summaries
NaturalQuestions (closed-book) natural_qa_closedbook	short-answer question answering	passages from Wikipedia, questions from search queries	web users
NaturalQuestions (open-book) natural_qa_openbook_longans	short-answer question answering	passages from Wikipedia, questions from search queries	web users
OpenbookQA openbookqa	multiple-choice question answering	elementary science	Amazon Mechanical Turk workers
MMLU (Massive Multitask Language Understanding) mmlu	multiple-choice question answering	math, science, history, etc.	various online sources
GSM8K (Grade School Math) gsm	numeric answer question answering	grade school math word problems	contractors on Upwork and Surge AI
MATH math_chain_of_thought	numeric answer question answering	math competitions (AMC, AIME, etc.)	problem setters
LegalBench legalbench	multiple-choice question answering	public legal and administrative documents, manually constructed questions	lawyers
MedQA med_qa	multiple-choice question answering	US medical licensing exams	problem setters
WMT 2014 wmt_14	machine translation	multilingual sentences	Europarl, news, Common Crawl, etc.

(From HELM)

Is ‘parametric knowledge’ useful?

Why is parametric knowledge a common capability test?

- Easy to collect, easy to automatically grade (GPQA, Humanity’s last exam)
- Knowledge is useful, even if you are going to do things like summarization
- In practice, correlated with *other* model capabilities (Knowledge <-> scale)

How does a LM acquire parametric knowledge?

Stanford University is located in _____

We know ‘filling in the blank’ can lead to knowledge being acquired...

But what kind of data do we need, and how much of it?

Acquiring parametric knowledge

Recent works – careful, controlled synthetic experiments on how LLMs acquire knowledge

Physics of Language Models: Part 3.1, Knowledge Storage and Extraction

Zeyuan Allen-Zhu
zeyuanallen.zhu@meta.com
Meta / FAIR Labs

Yuanzhi Li
Yuanzhi.Li@mbzuai.ac.ae
Mohamed bin Zayed University of AI

September 18, 2023

(version 3)*

Empirical observations – you need many exposures (~1000) or very large models to acquire knowledge

- SECTION 6: HOW TRAINING TIME AFFECTS MODEL CAPACITY.

Achieving a 2bit/param capacity requires each knowledge piece to be visited 1000 times during training, termed ***1000-exposure*** to differentiate from traditional “1000-pass” terminology, as a single data pass can expose a knowledge piece 1000 times.⁵

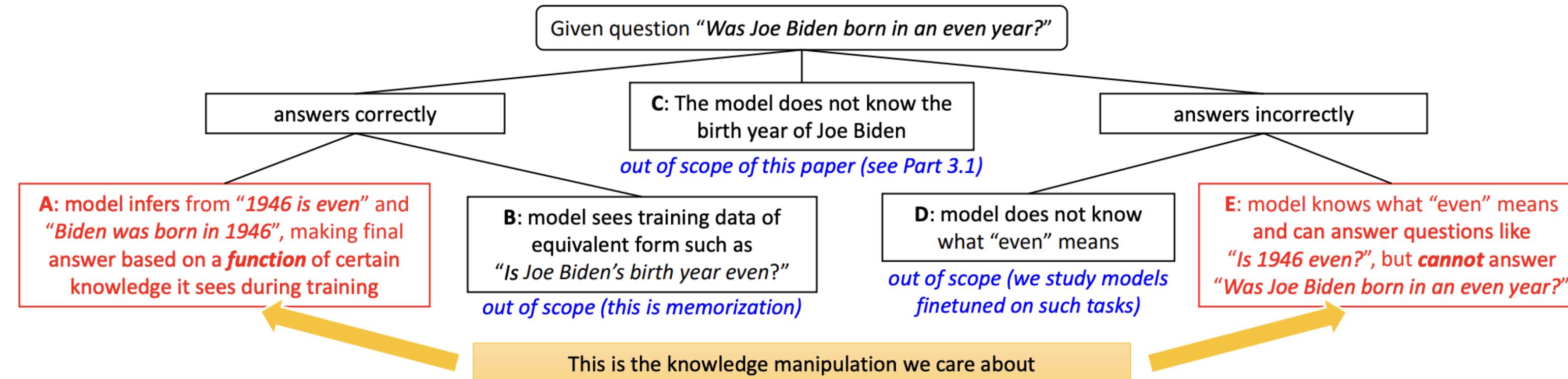
- **RESULT 4:** With 100 exposures, an *undertrained* GPT2’s capacity ratio falls to 1bit/param.

Remark 1.3. Another perspective on Result 4 is that *rare* knowledge, encountered only 100 times during training, is stored at a 1bit/param ratio.

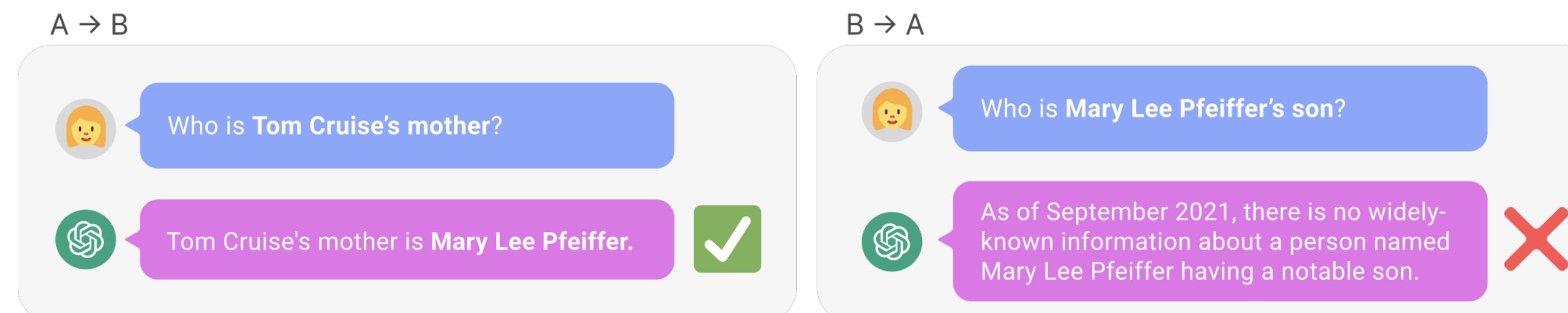
Some challenges with parametric knowledge

Acquiring knowledge *robustly* is difficult.

Entailed knowledge is hard to learn (without CoT)



'Reversal curse' constrains knowledge use



Hallucinations



REUTERS®

Australian mayor readies world's first defamation lawsuit over ChatGPT content

By Byron Kaye

Air Canada ordered to pay customer who was misled by airline's chatbot

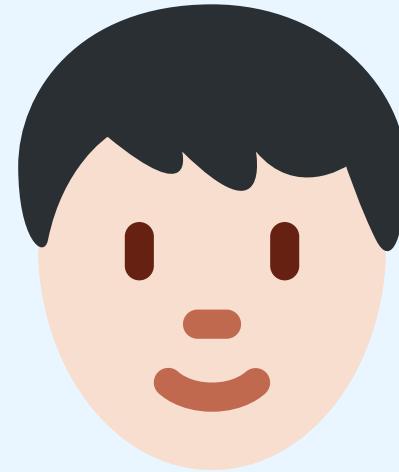
Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare



LLMs are *remarkably* good at many things – even closed-book QA
.. But they aggressively make things up for things they don't know

Abstention – knowing what you know

We don't necessarily know everything in the world



Are there any strictly proper, bounded scoring rules?



I don't know anything about scoring rules.

.. but we generally know how to back off and communicate uncertainty

Methods for abstention

‘Backing off’

- Estimate the *confidence* of a statement (by sampling or self-scoring)
- Abstain whenever the confidence is too low (or remove details)

$$\begin{array}{rcl} x & = & \text{Who was Abe Lincoln?} \\ \hline F_1(x) & = & \text{Abraham Lincoln, } \underbrace{\text{born in Idaho}}_{(1)}, \underbrace{\text{was the 16th}}_{(2)} \dots \\ & & \text{President of the United States. } \underbrace{\text{He is best known}}_{(2)} \dots \\ & & \text{for leading the country through the Civil War.} \underbrace{\dots}_{(3)} \\ \hline F_2(x) & = & \underbrace{\text{Abraham Lincoln was the 16th President of the}}_{(2)} \dots \\ & & \underbrace{\text{United States. } \underbrace{\text{He is best known for leading the}}_{(2)} \dots}_{(3)} \\ & & \text{country through the Civil War.} \underbrace{\dots}_{(3)} \\ \hline F_3(x) & = & \underbrace{\text{Abraham Lincoln was the 16th President of the}}_{(2)} \dots \\ & & \underbrace{\text{United States.}}_{(2)} \dots \\ \hline F_4(x) & = & \emptyset \end{array}$$

Example (from Conformal Factuality, Mohri+ 2024)

Adding new knowledge to a model

To go beyond abstaining..

Can we add new, parametric knowledge after pre-training?



Goal: reliably back off when models *don't* know, but try to add knowledge when we can

Updating knowledge (continued pretraining)

One way to update knowledge – continue to do pretraining with domain-specific data

Study	Domain	Model Parameter Count	Total Unique CPT Tokens
Minerva (Lewkowycz et al., 2022)	STEM	8B, 62B, 540B	26B-38.5B
MediTron (Chen et al., 2023)	Medicine	7B, 70B	46.7B
Code Llama (Rozière et al., 2024)	Code	7B, 13B, 34B	520B-620B
Llemma (Azerbayev et al., 2024)	Math	7B, 34B	50B-55B
DeepSeekMath (Shao et al., 2024)	Math	7B	500B
SaulLM-7B (Colombo et al., 2024b)	Law	7B	30B
SaulLM-{54, 141}B (Colombo et al., 2024a)	Law	54B, 141B	520B
HEAL (Yuan et al., 2024a)	Medicine	13B	14.9B

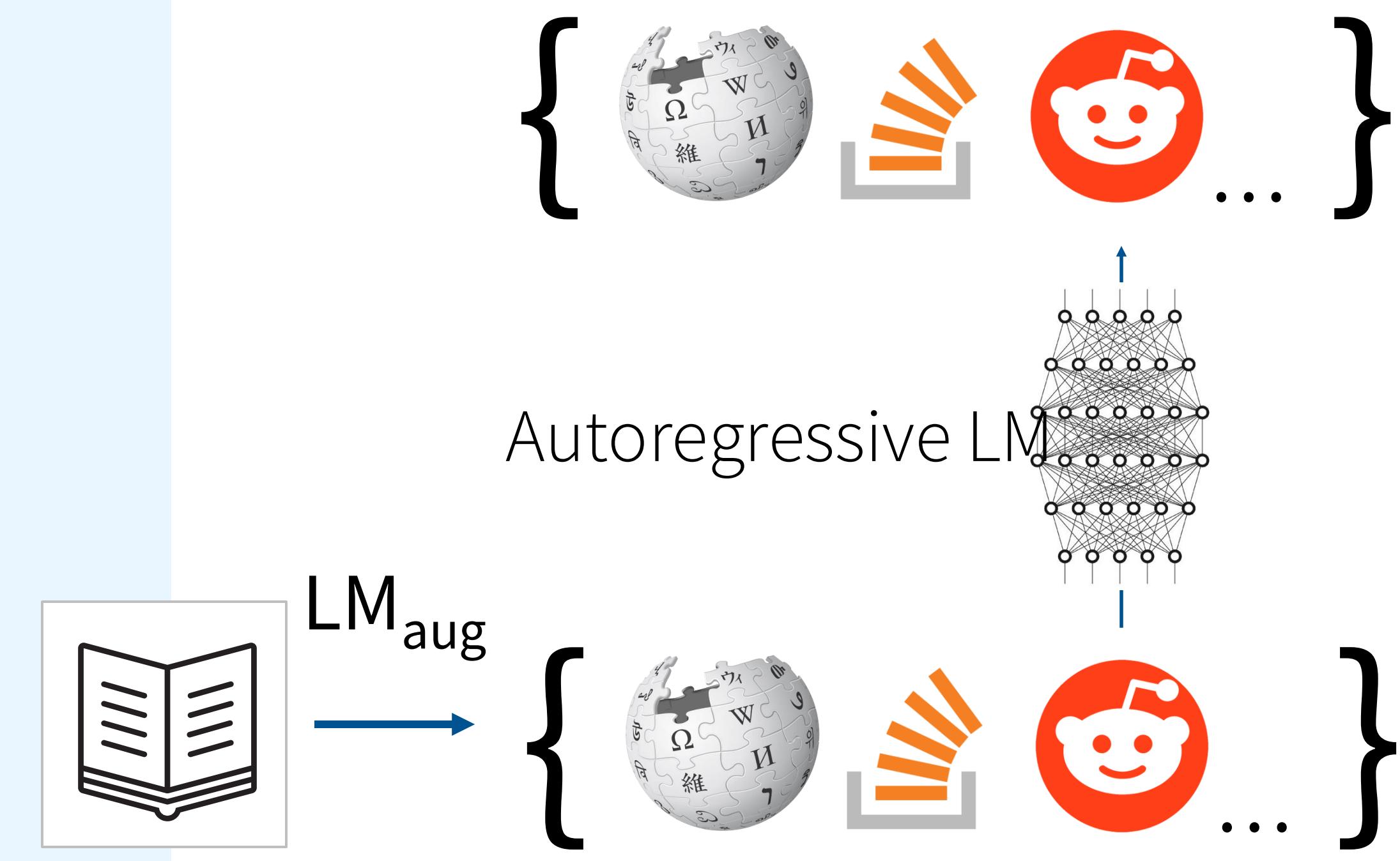
Unfortunately.. require *tons* of data (50+B)

Updating knowledge (synthetic CPT)

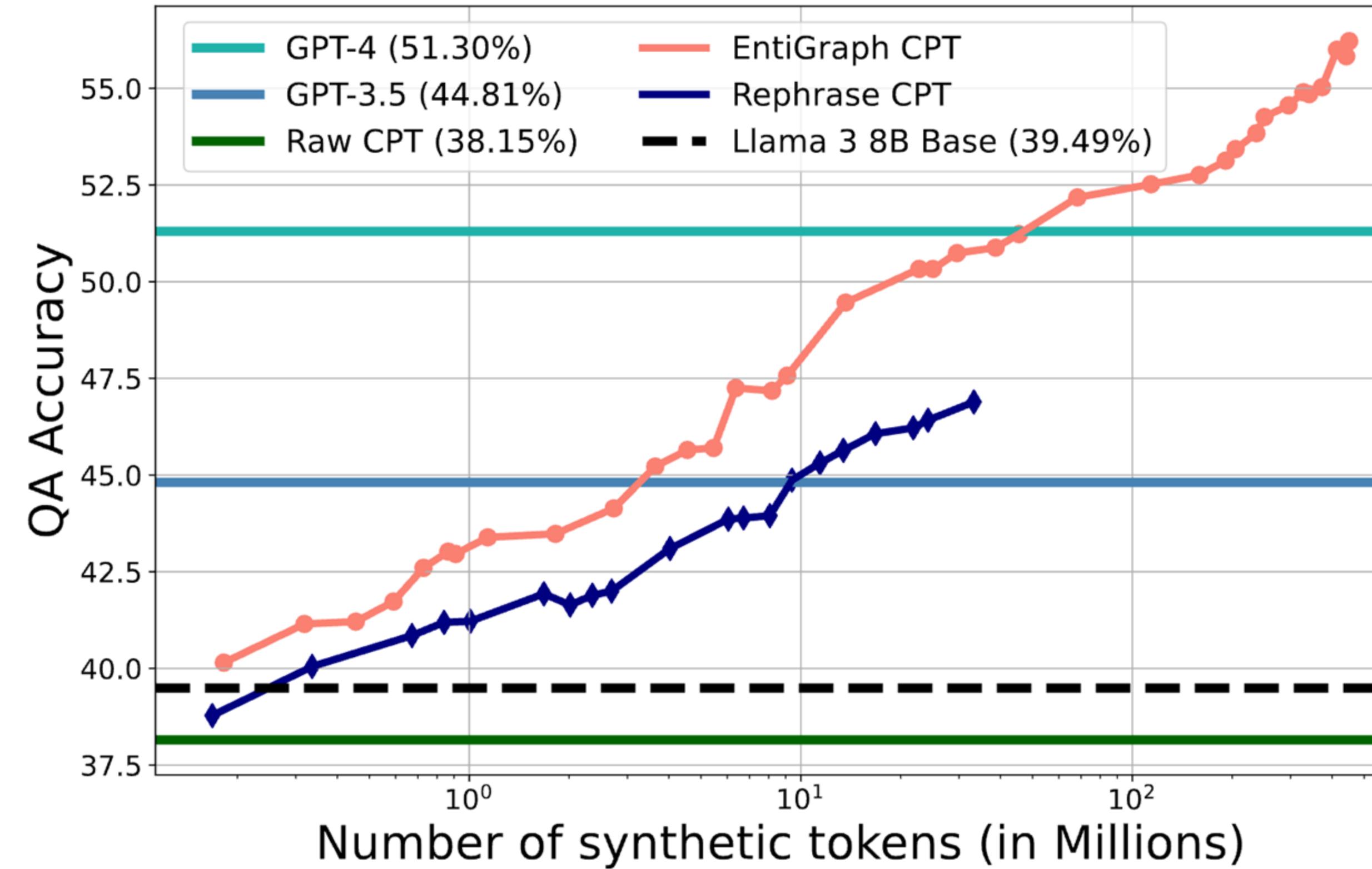
Synthetic continued pretraining: Train on LLM-transformed data

Goal – replicate the diversity of pretraining

- Vary content (topics)
- Vary style (how it's presented)
- Data diversity for generalization



Updating knowledge with synthetic data



- Can significantly boost closed-book QA performance with synthetic data

Black-box LMs are impressive, but parametric memory has its limitations

- (Large) language models 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 LMs are opaque
 - Given a query, it produces an answer, but it's difficult to verify if the answer is correct



Open book QA - Reading comprehension

Reading comprehension = comprehend a passage of text and answer questions about its content (P, Q) → 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

Reading comprehension

Reading comprehension = comprehend a passage of text and answer questions about its content (P, Q) → 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
- 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

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)

Reading comprehension – span based

Archetypical reading comprehension dataset -- SQuAD

A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., 3, $1 \cdot 3$, $1 \cdot 1 \cdot 3$, etc. are all valid factorizations of 3.

What is the only divisor besides 1 that a prime number can have?

Ground Truth Answers: `itself` `itself` `itself` `itself` `itself`

What are numbers greater than 1 that can be divided by 3 or more numbers called?

Ground Truth Answers: `composite number` `composite number` `composite number` `primes`

What theorem defines the main role of primes in number theory?

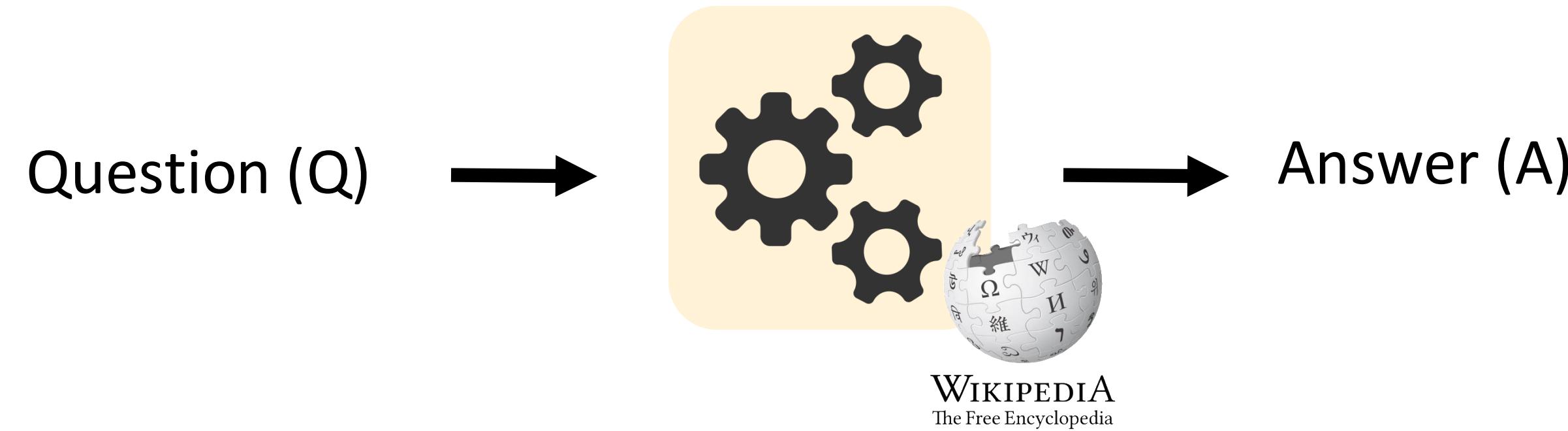
Ground Truth Answers: `The fundamental theorem of arithmetic` `fundamental theorem of arithmetic` `arithmetic` `fundamental theorem of arithmetic` `fundamental theorem of arithmetic` `fundamental theorem of arithmetic`

Major focus of Q&A and research for many years..

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*

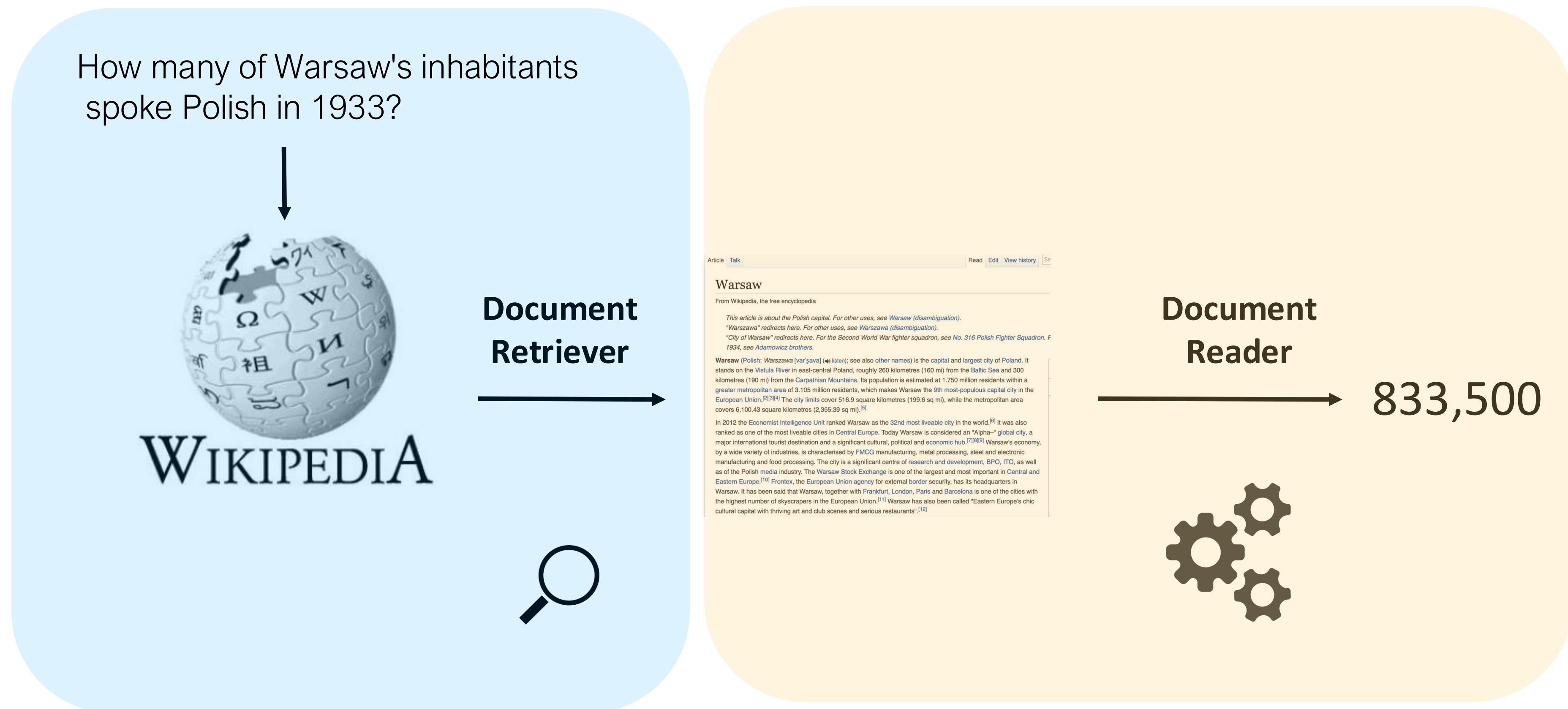
Open-domain question answering



- 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!

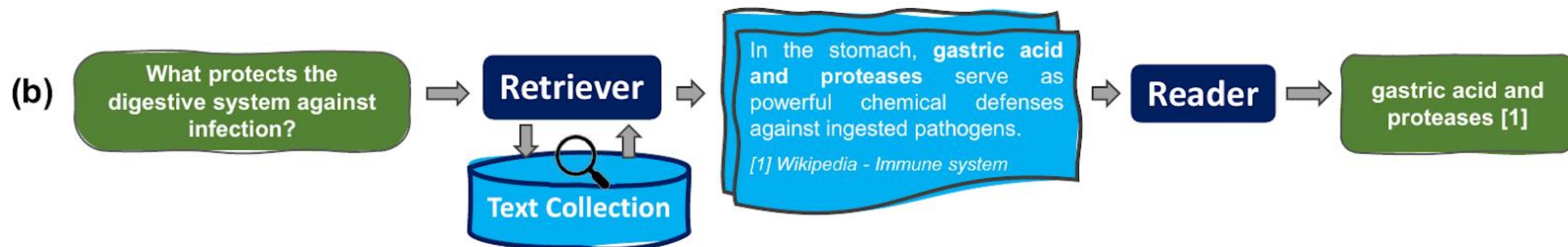
*In contrast to **closed-domain** systems that deal with questions under a specific domain (medicine, technical support).*

Retrieval augmentation



Using retrieval to overcome LMs' shortcomings

- Instead of asking the LM to memorize everything, can we provide the LM with relevant and useful content just-in-time?
- Retrieval / search 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)

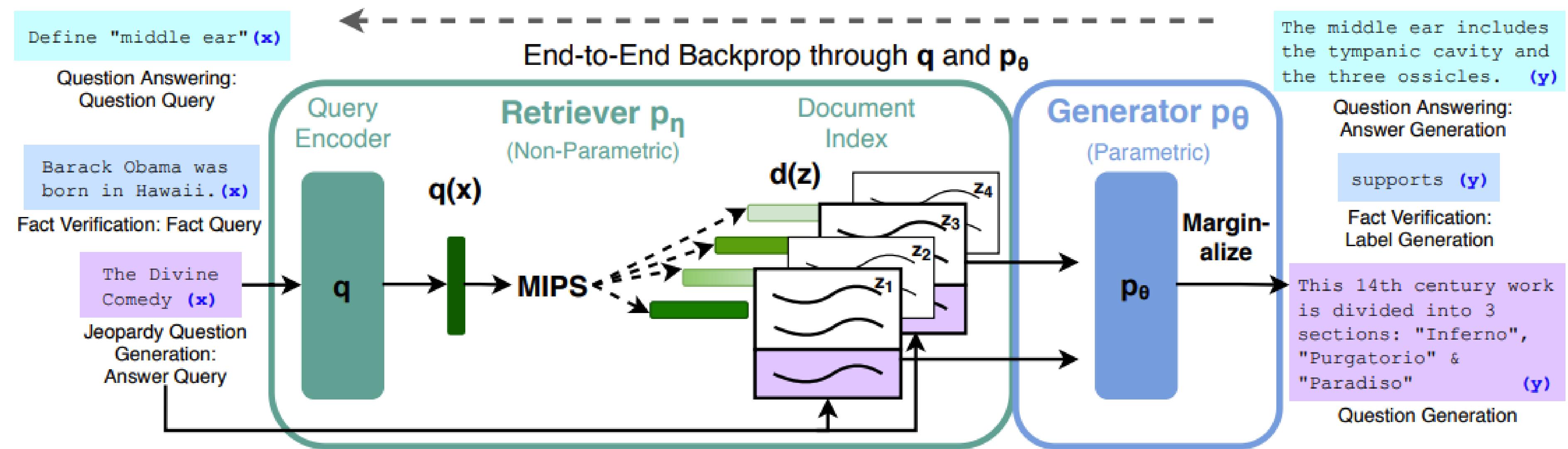


Retriever-reader framework

- Input: a large collection of documents $\mathcal{D} = D_1, D_2, \dots, D_N$ and Q
- Output: an answer string A
- Retriever: $f(\mathcal{D}, Q) \rightarrow P_1, \dots, P_K$ K is pre-defined (e.g., 100)
- Reader: $g(Q, \{P_1, \dots, P_K\}) \rightarrow A$ A reading comprehension problem!
- Retriever = A standard TF-IDF information-retrieval sparse model (a fixed module)
- Reader = a neural reading comprehension model that we just learned
 - Could be Trained on SQuAD and other distantly-supervised QA datasets
 - Or a zero-shot LLM (ChatGPT etc)

RAG and it's open problems

- Retrieval Augmented Generation (RAG) is very powerful!



- See also works like REALM, DPR, ORQA etc..

[Piktus+ 2021]

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

Different types of retrievers

How should we retrieve relevant passages for our retrieval system?

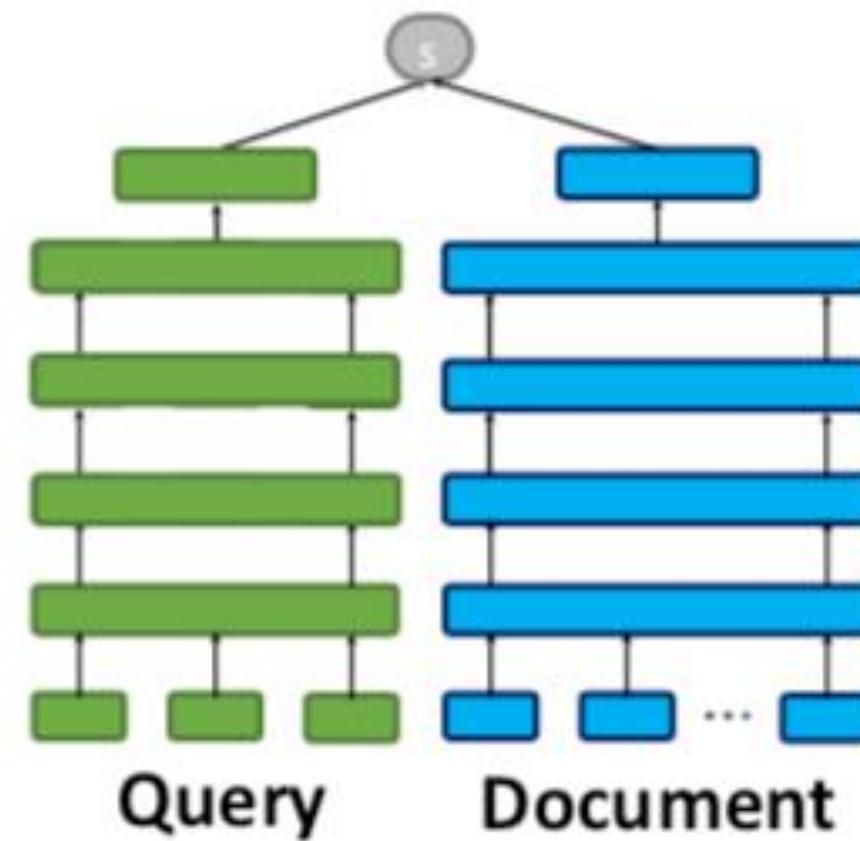
- Word-overlap (BM25)
- Vector retrieval (DPR, Sentence vectors)
- Other, neural systems (ColBERT)

Fast vs slow retrievers

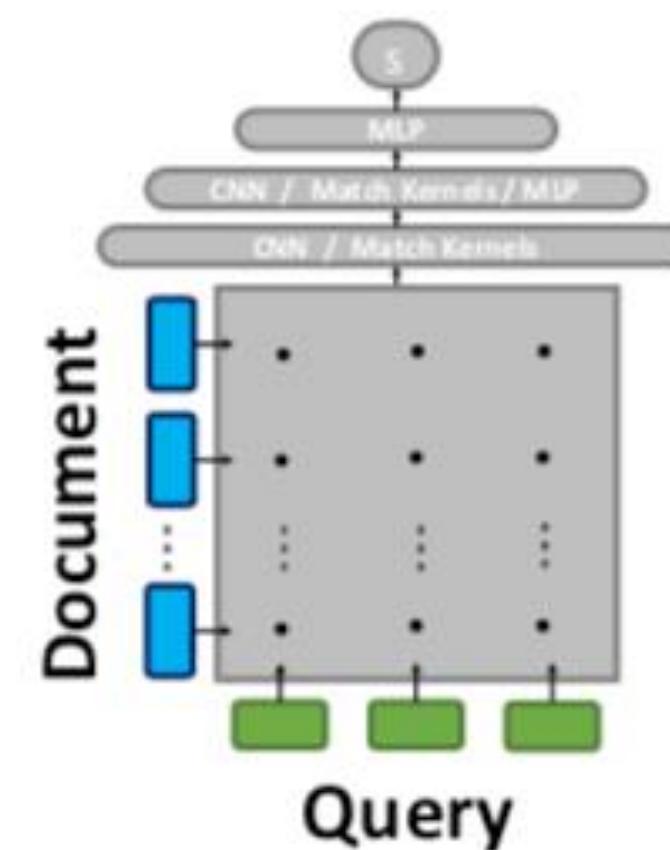
Fastest: computing similarities (or word overlap) on pre-computed vectors

Slowest: using a LM to compute similarities

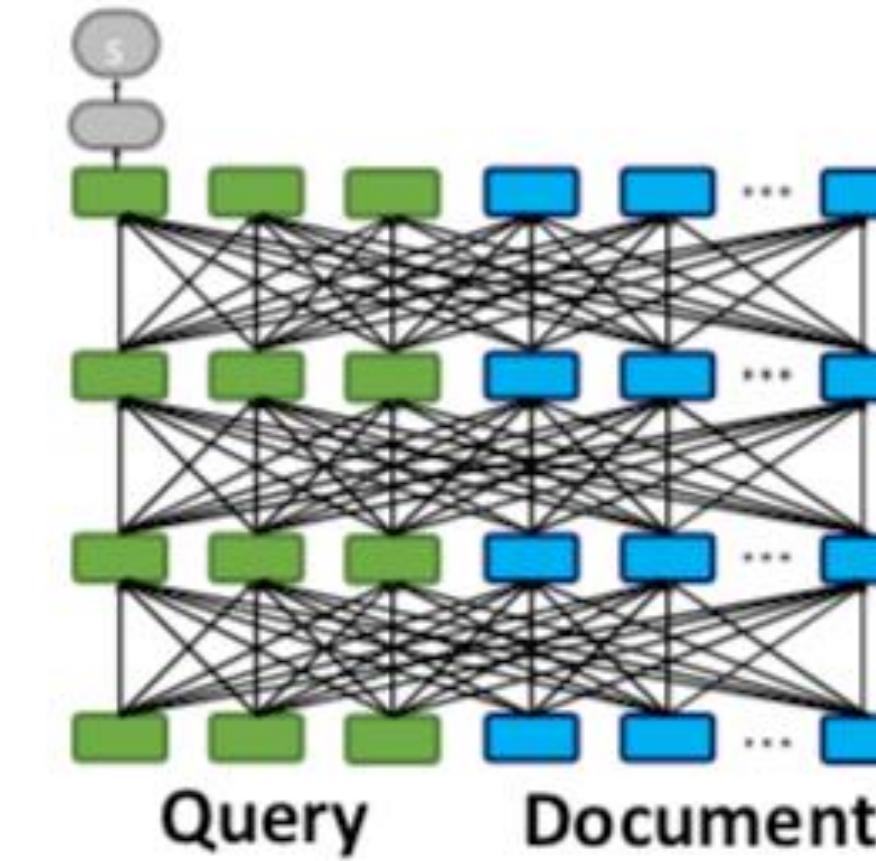
Other hybrid (pre-computing the ‘document index’) e.g. in ColBERT



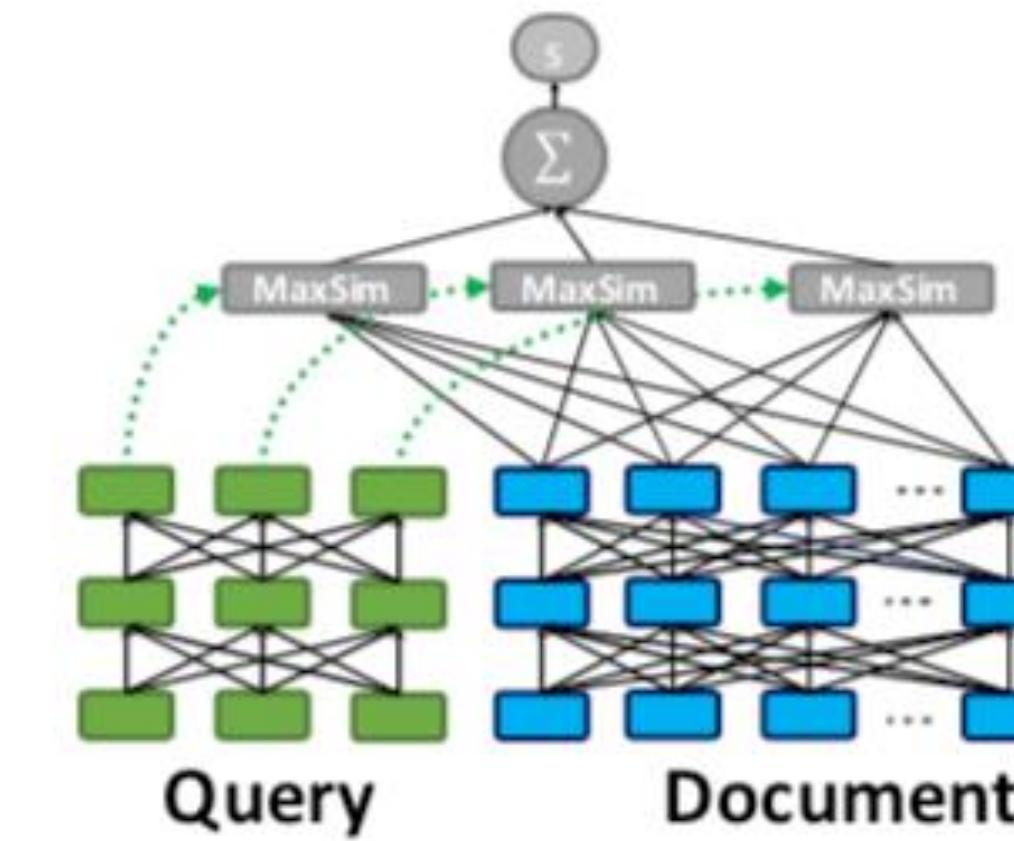
(a) Representation-based Similarity
(e.g., DSSM, SNRM)



(b) Query-Document Interaction
(e.g., DRMM, KNRM, Conv-KNRM)



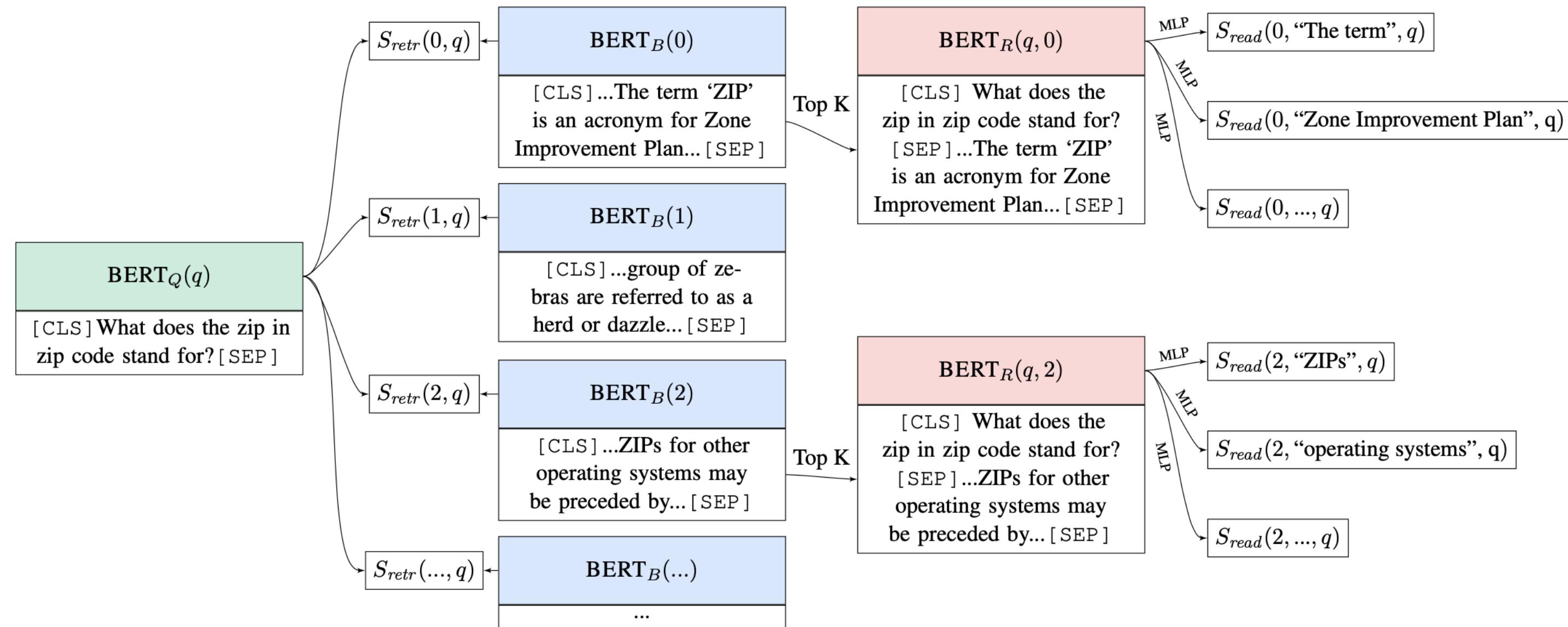
(c) All-to-all Interaction
(e.g., BERT)



(d) Late Interaction
(i.e., the proposed ColBERT)

We can train the retriever too

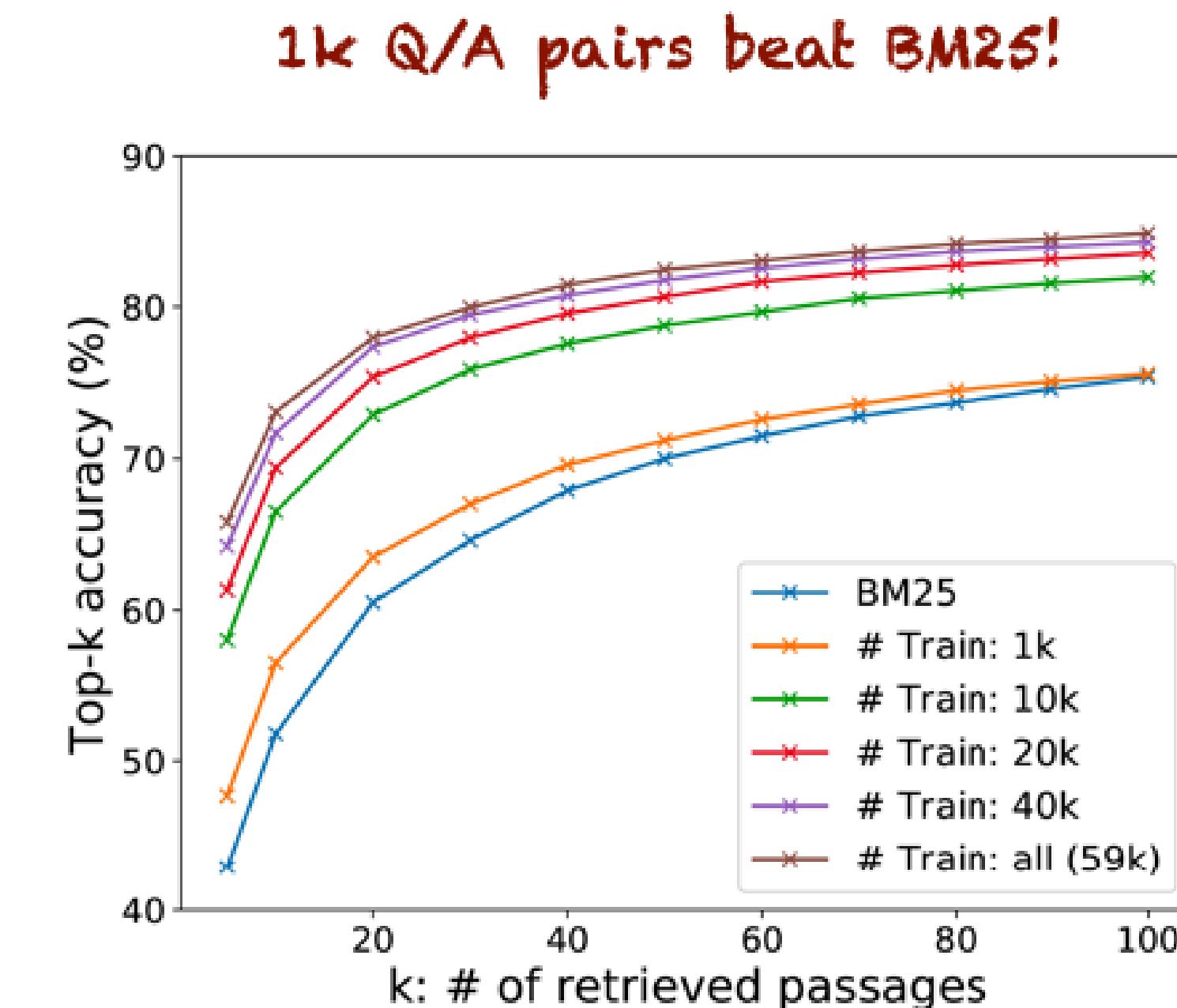
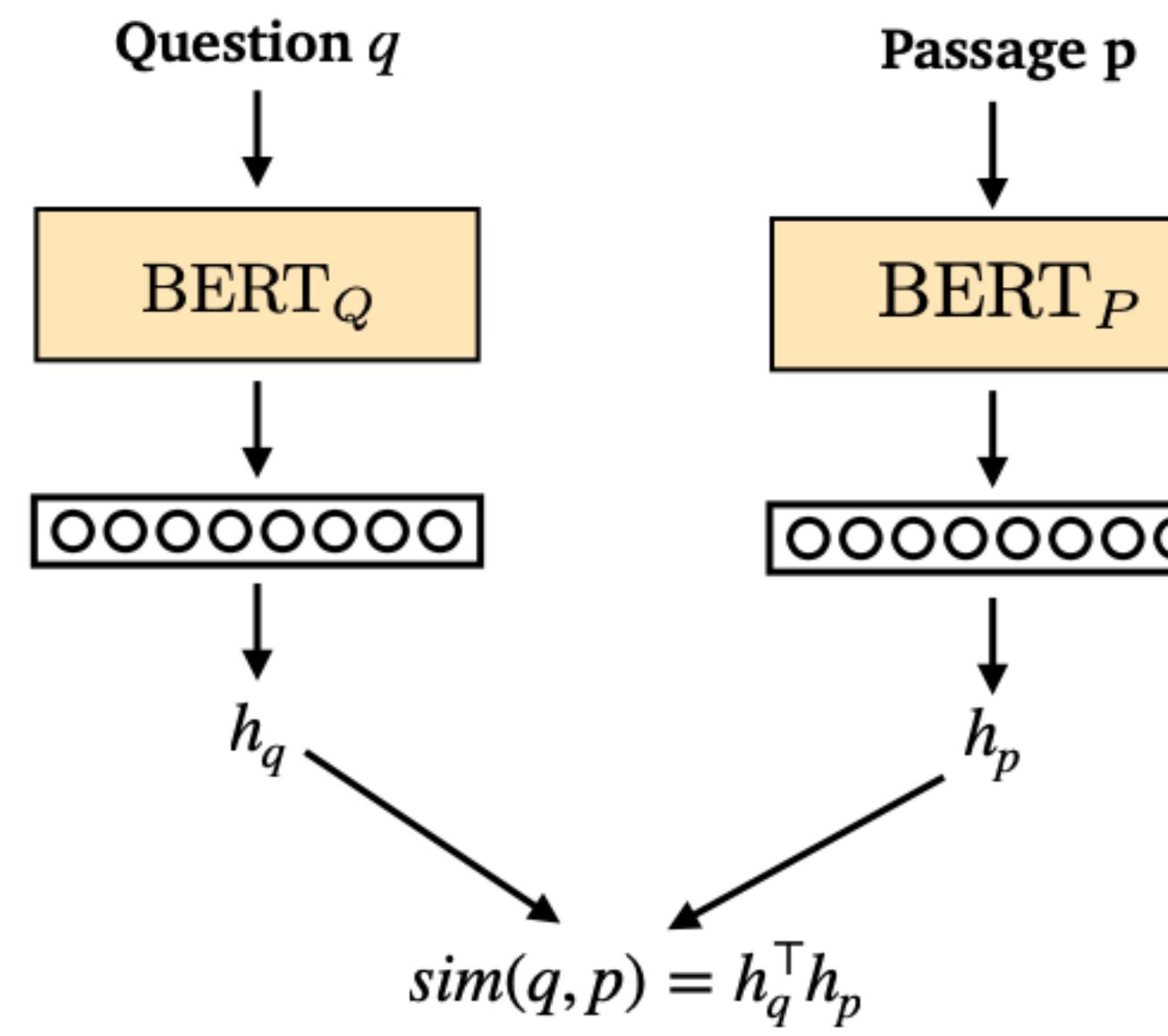
- Joint training of retriever and reader



- 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)

We can train the retriever too

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

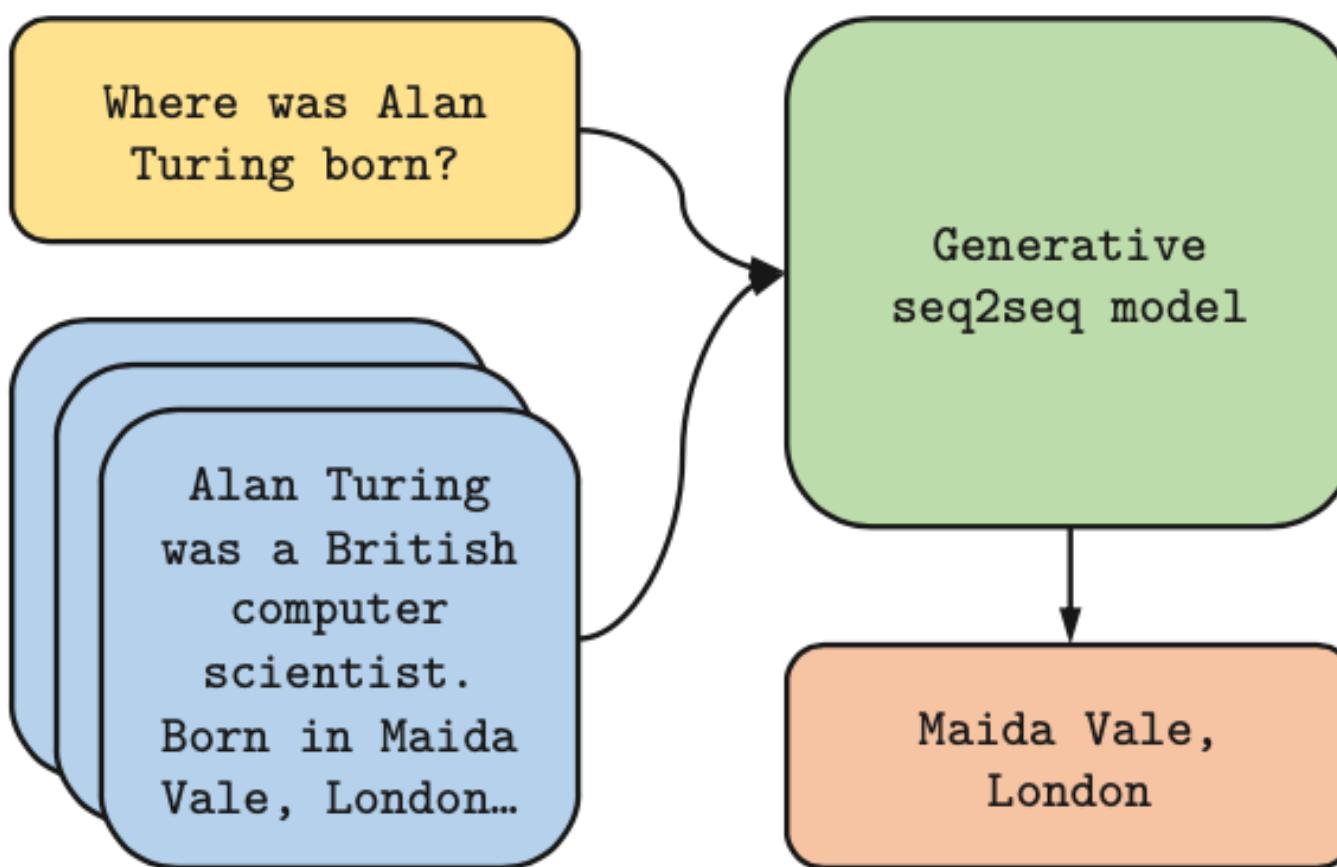


- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models

Dense 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

Problem: How many documents can we use?

- The retriever is key – if we have to use only one 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) ...

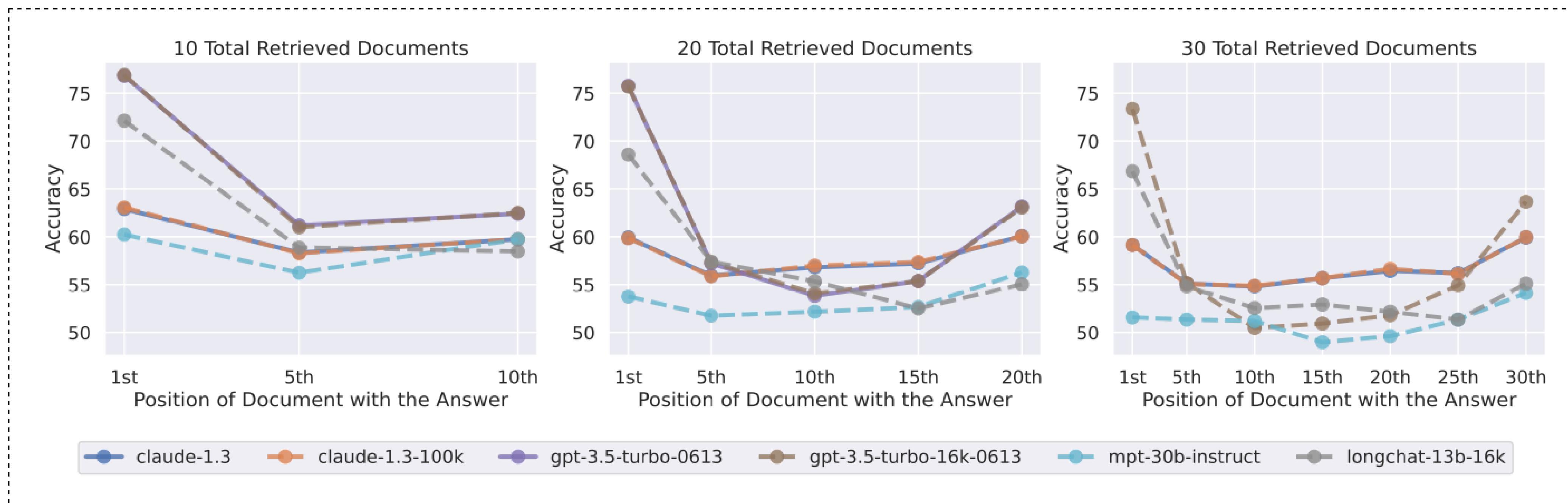
Question: who got the first nobel prize in physics
Answer:

Desired Answer

Wilhelm Conrad Röntgen

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

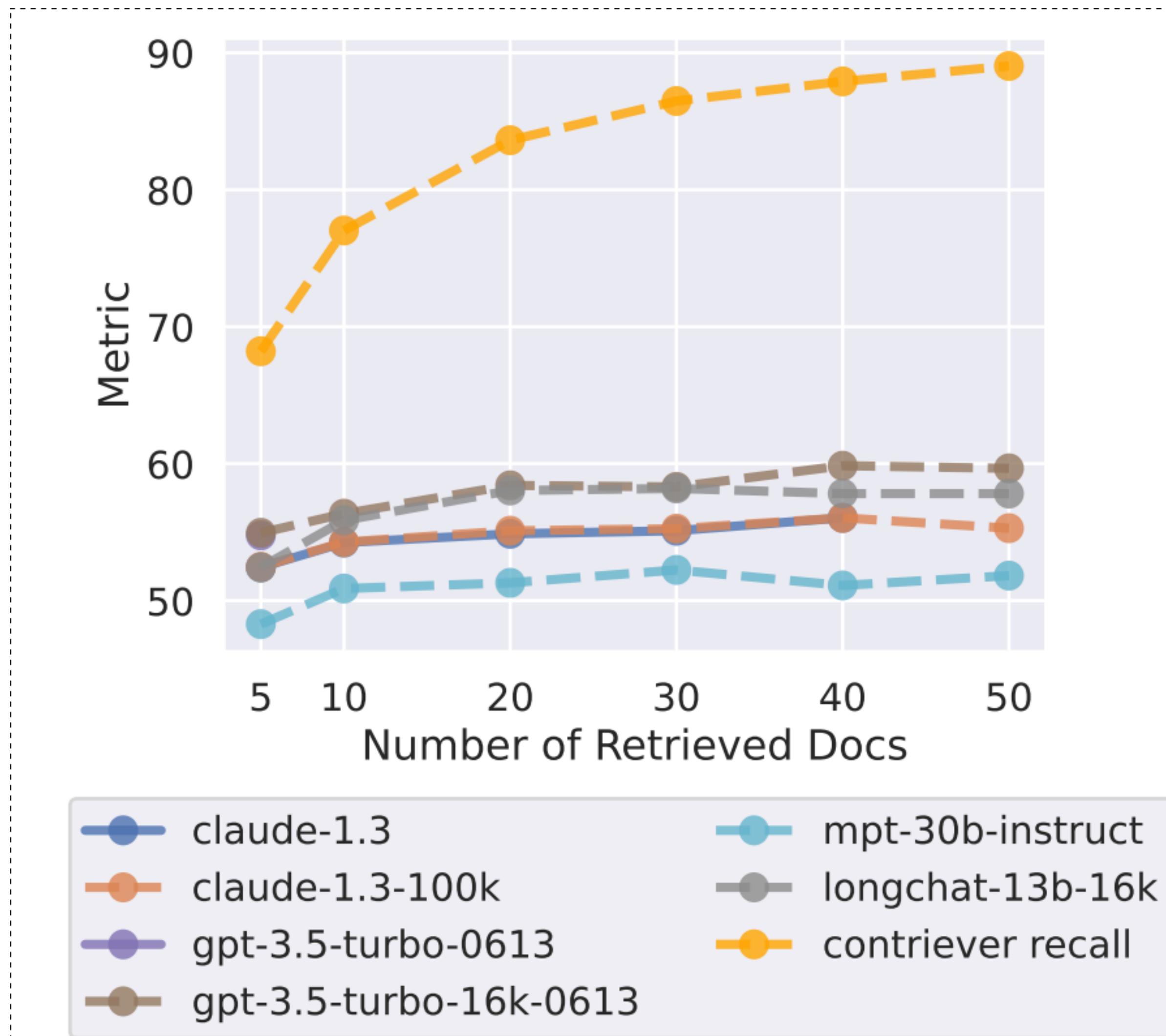
- The long-context problem bites you – LMs do not pay attention to its context well
 - Setup: 1 relevant document, all others irrelevant



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

In practice: LMs can't use many documents

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

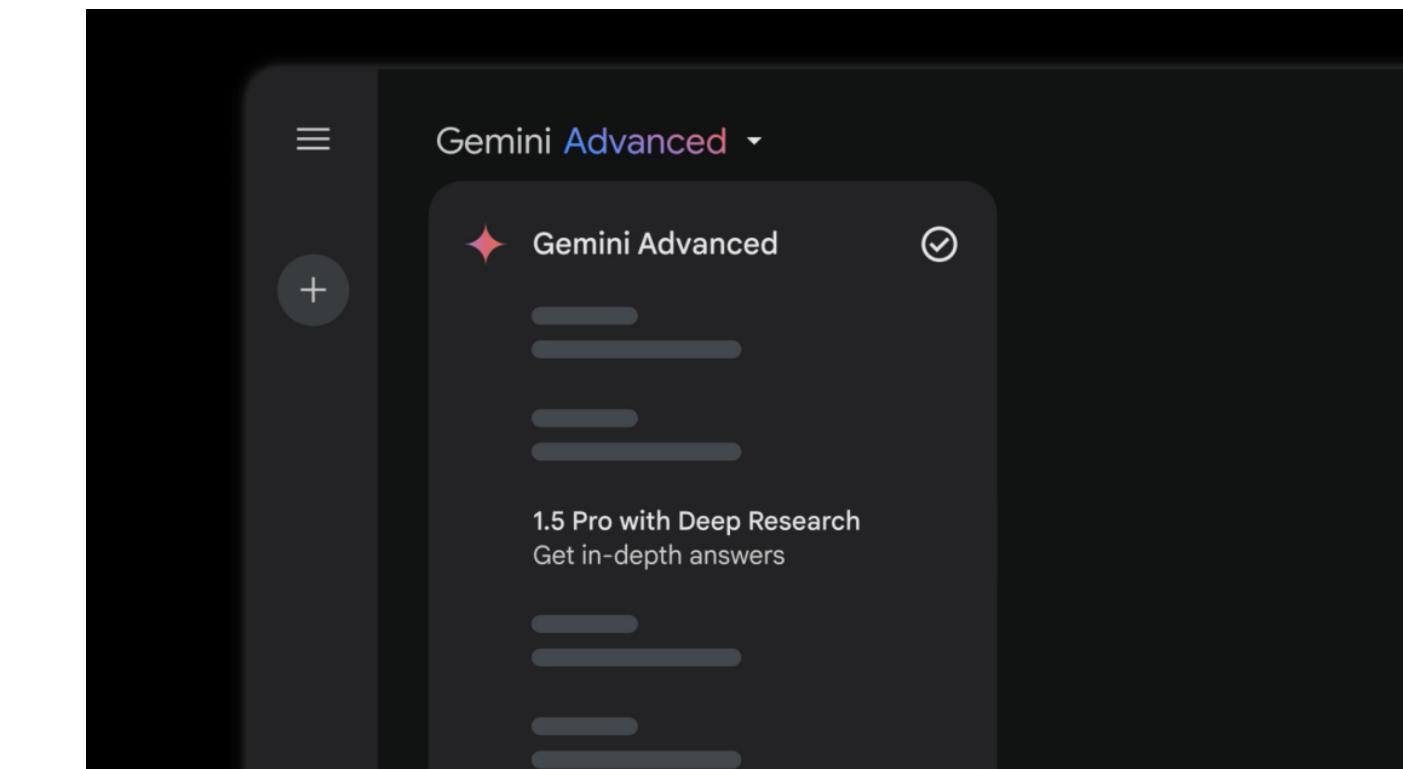
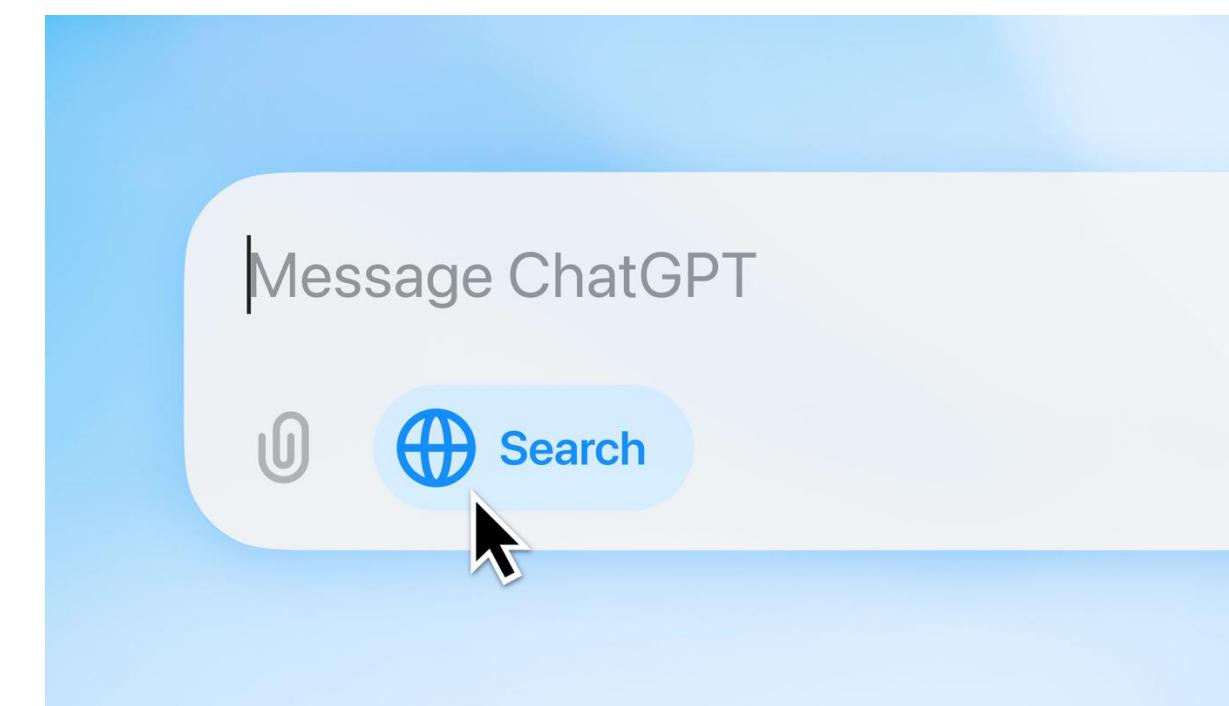
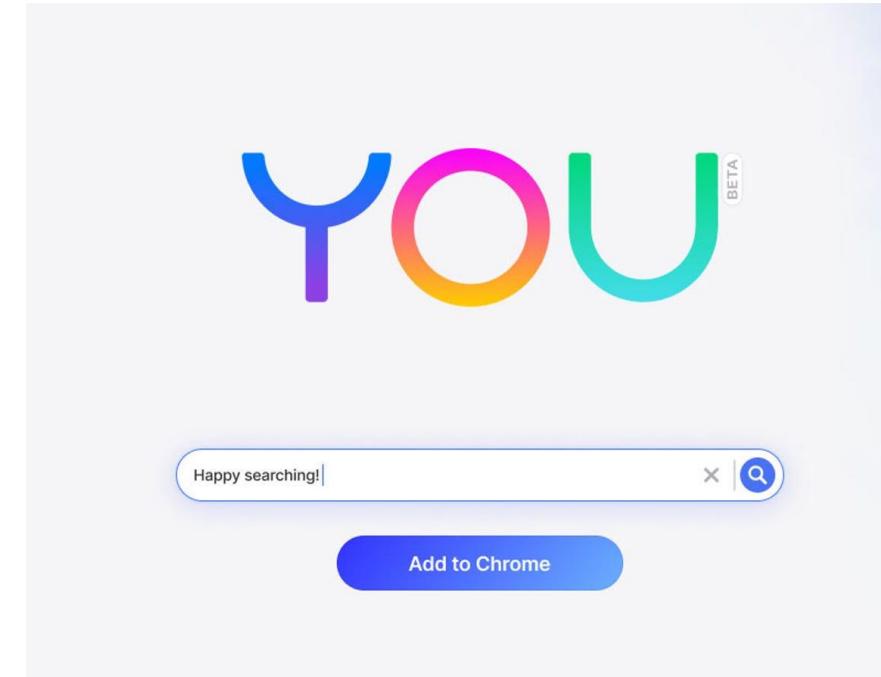
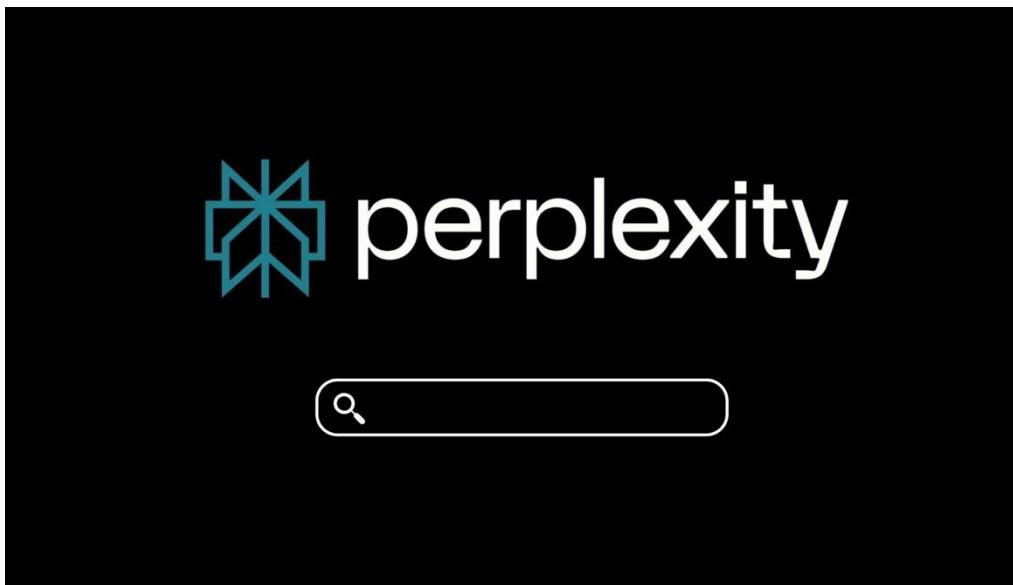


Even more open domain: LLM (with web search!)

The screenshot shows a search interface with the following elements:

- Search Bar:** The query "Where does Christopher D. Manning teach?" is entered.
- Search Buttons:** Includes a magnifying glass icon, a refresh/circular arrow icon, and a gear/refresh icon.
- Filter Buttons:** Options include All, Chat (which is selected), Images, Videos, News, Maps, and More.
- Result Summary:** A blue button displays the search query again: "Where does Christopher D. Manning teach?"
- Text Description:** Christopher D. Manning is a professor of computer science and linguistics at Stanford University [1][2]. He is also the Director of the Stanford Artificial Intelligence Laboratory [1], and has been teaching at Stanford since 1994.
- Links:**
 - [1. Christopher Manning's Profile | Stanford Profiles](https://profiles.stanford.edu/chris-manning)
https://profiles.stanford.edu/chris-manning
 - [2. Introduction to Information Retrieval: Manning, Christopher D ...](https://www.amazon.com/Introduction-Information-Retrieval-Christopher-Manning/dp/0521865719)
https://www.amazon.com/Introduction-Information-Retrieval-Christopher-Manning/dp/0521865719
- Interaction Buttons:** Upvote, like, and dislike icons.

Other examples of LLMs with web search

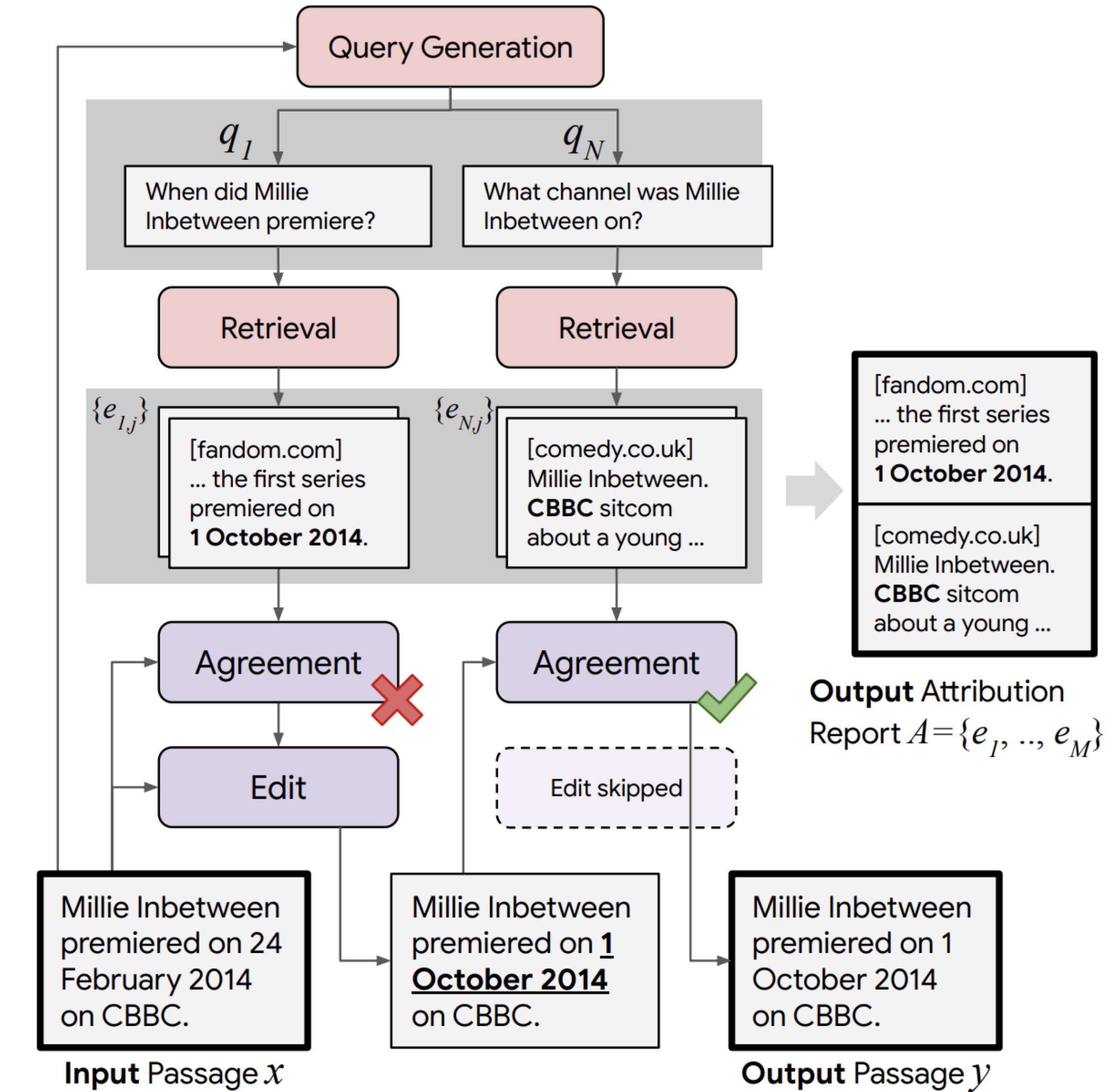


Lots of systems that search and cite the web!

Systems architectures of such systems

Google setup (RARR)

- Uses QA to check for consistency between sources and responses
- Returns citations for each of these items using a LM.



Are LLM-based QA systems the answer?

- A unique benefit of RAG: citing your sources

The screenshot shows a user interface for generating an itinerary. At the top, a search bar contains the query: "Create an itinerary for exploring Toronto over a span of two days." Below the search bar are two cards:

Card 1 (Left):

Toronto is a vibrant and diverse city with many attractions and activities to suit different tastes and preferences. Here are some suggestions for creating an itinerary for exploring Toronto over two days:

- Day 1: Start your day with a visit to the **CN Tower**, the iconic landmark that offers panoramic views of the city and beyond ¹. You can also experience the thrill of walking on the glass floor or the EdgeWalk, the world's highest hands-free walk ¹. Next, head to the nearby **St. Lawrence Market**, one of the world's best food markets, where you can sample a variety of cuisines and local specialties ². After lunch, take a stroll along **Queen West**, a trendy neighborhood with eclectic shops, galleries, cafes and street art ³. In the evening, enjoy a show at one of the many theaters or comedy clubs in the **Entertainment District**, or catch a game at the **Scotiabank Arena** if you're a sports fan.

Card 2 (Right):

- Day 2: Explore the history and culture of Toronto at the **Royal Ontario Museum**, one of the largest and most comprehensive museums in North America, featuring exhibits on art, natural history, world cultures and more ⁴. Then, hop on a ferry to the **Toronto Islands**, a group of islands that offer a relaxing escape from the city, with beaches, parks, trails and amusement rides ³ ⁵. You can also rent a bike or kayak to explore the islands at your own pace. For dinner, head to **Chinatown**, one of the largest and most vibrant in North America, where you can find a variety of Asian cuisines and shops ³.

I hope this helps you plan your trip to Toronto. Have fun! 😊

Learn more:

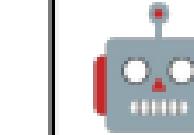
- 1. cntower.ca
- 2. travel.usnews.com
- 3. bing.com
- 4. rom.on.ca
- 5. tripadvisor.com

How accurate are these citations?

- These citations themselves are generated by LLMs



What are the latest discoveries from the James Webb Space Telescope?



The James Webb Space Telescope is designed to peer into the dusty clouds of gas where stars and planetary systems are born. Webb has captured the first direct image of an exoplanet, and the Pillars of Creation in the Eagle Nebula[1][2]. Additionally, the telescope will be used to study the next interstellar interloper[3].

(*Some generated statements may not be fully supported by citations, while others are fully supported.)

Cited Webpages

[1]: nasa.gov (citation does not support its associated statement)
[NASA's Webb Confirms Its First Exoplanet](#)

... Researchers confirmed an exoplanet, a planet that orbits another star, using NASA's James Webb Space Telescope for the first time. ...

[2]: cnn.com (citation partially supports its associated statement)
[Pillars of Creation: James Webb Space Telescope ...](#)

... The Pillars of Creation, in the Eagle Nebula, is a star-forming region captured in a new image (right) by the James Webb Space Telescope that reveals more detail than a 2014 image (left) by Hubble ...

[3]: nasa.gov (citation fully supports its associated statement)
[Studying the Next Interstellar Interloper with Webb](#)

... Scientists have had only limited ability to study these objects once discovered, but all of that is about to change with NASA's James Webb Space Telescope...The team will use Webb's spectroscopic capabilities in both the near-infrared and mid-infrared bands to study two different aspects of the interstellar object.

Existing systems have high fluency

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

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

Bing Chat	4.40
NeevaAI	4.43
perplexity.ai	4.51
YouChat	4.59
Average	4.48

Low correctness for citations

- But precision and recall are both low..

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

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

Recap

QA is one of the top uses for a LLM

- QA as a capability evaluation, and for understanding knowledge
- QA as a system – RAG and internet retrieval
- Limitations: retrievers are not perfect, and models can't handle lots of documents