

Retrieval and RAG



CS 288 Spring 2026
UC Berkeley
cal-cs288.github.io/sp26

Berkeley **BAIR**
EECS

Assignment 3 is out!

- Create an end-to-end Retrieval-Augmented Generation (RAG) model!
 - No starter code
 - Includes building your own retrieval corpus and a QA dataset for evaluation
- **Early milestone due date:** 03/17 (Tue) 05:59PM **(No late days can be applied)**
- **Due date:** 03/19 (Thu) 05:59PM
- **03/17 (Tue) 7:00PM:** A reference corpus and QA data will be released, which can be used for the final RAG model
 - Since only two days are available (up to five with late days) between two deadlines, most of the model should have been built already using your own corpus and QA data before the early milestone. The reference corpus and QA data released on Mar 17 (evening) can be used for refinement.
- **Grading:** Early milestone results (20%), Final model results (40%), Written report (40%)

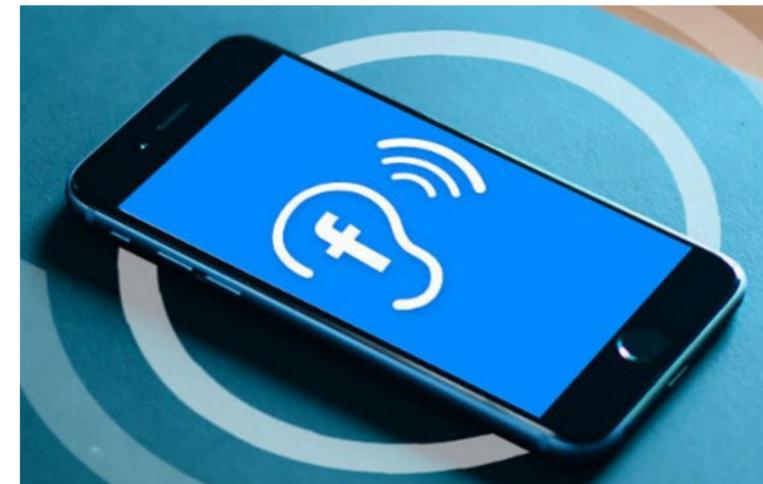
Today's topic: Retrieval and Retrieval-augmented generation (RAG)

- (Information) Retrieval: Given a query, identify a tiny subset of relevant documents or passages from a large corpus (typically 1 to 100 docs from a pool of >1M docs)
- Retrieval-based LMs / Retrieval-augmented LMs / Retrieval-augmented generation (RAG): LMs that use a retrieval model to identify relevant documents or passages and takes them as input to generate the output.
- Most representative task: (Open-domain) question answering!
- Lecture plan
 - Tutorial on question answering (QA) as a task
 - Retrieval 101
 - Retrieval-augmented generation 101 (likely next Tuesday)

Question answering

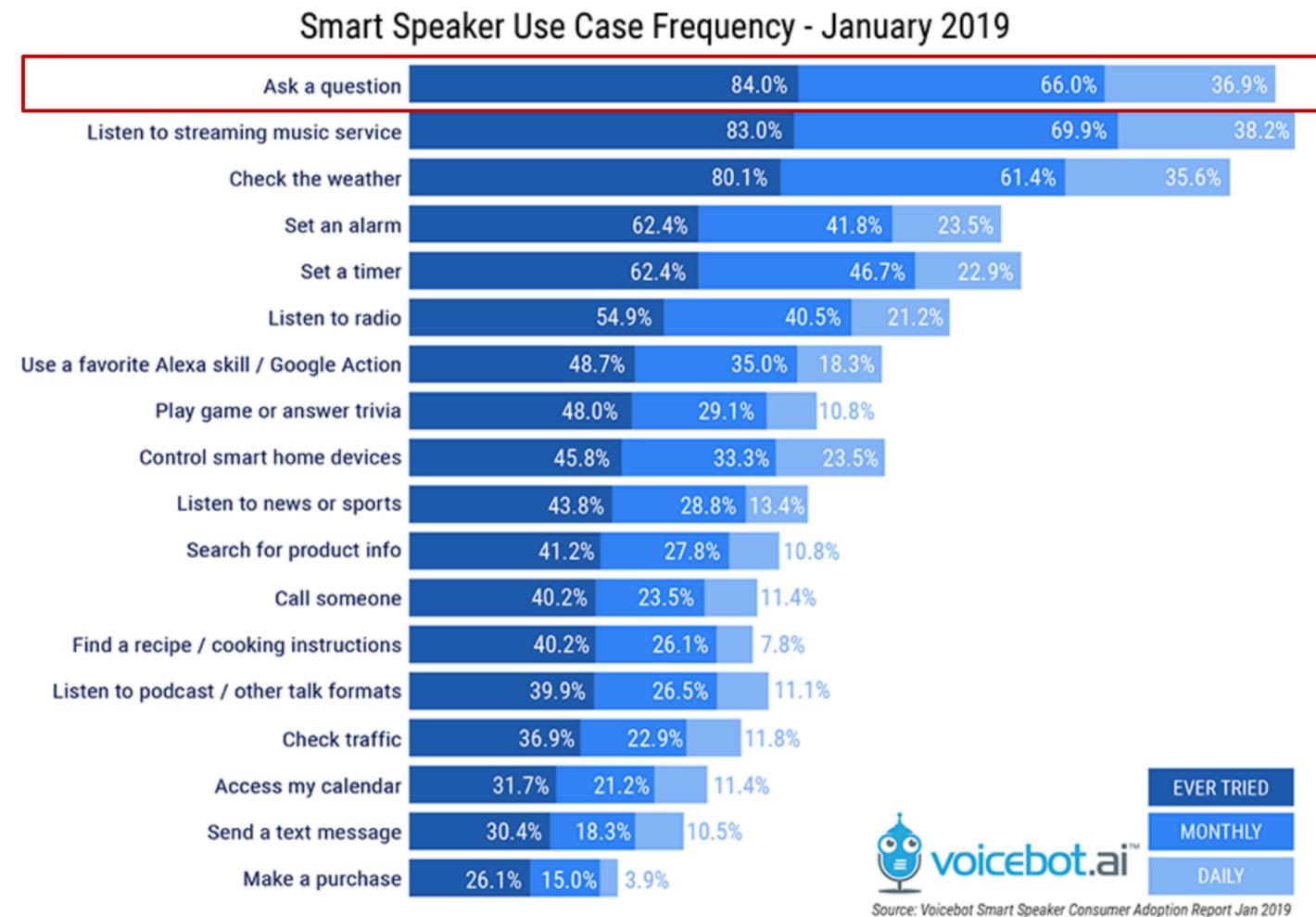
Question answering

- Key element in a wide range of applications (search engines, personal assistants, ...)



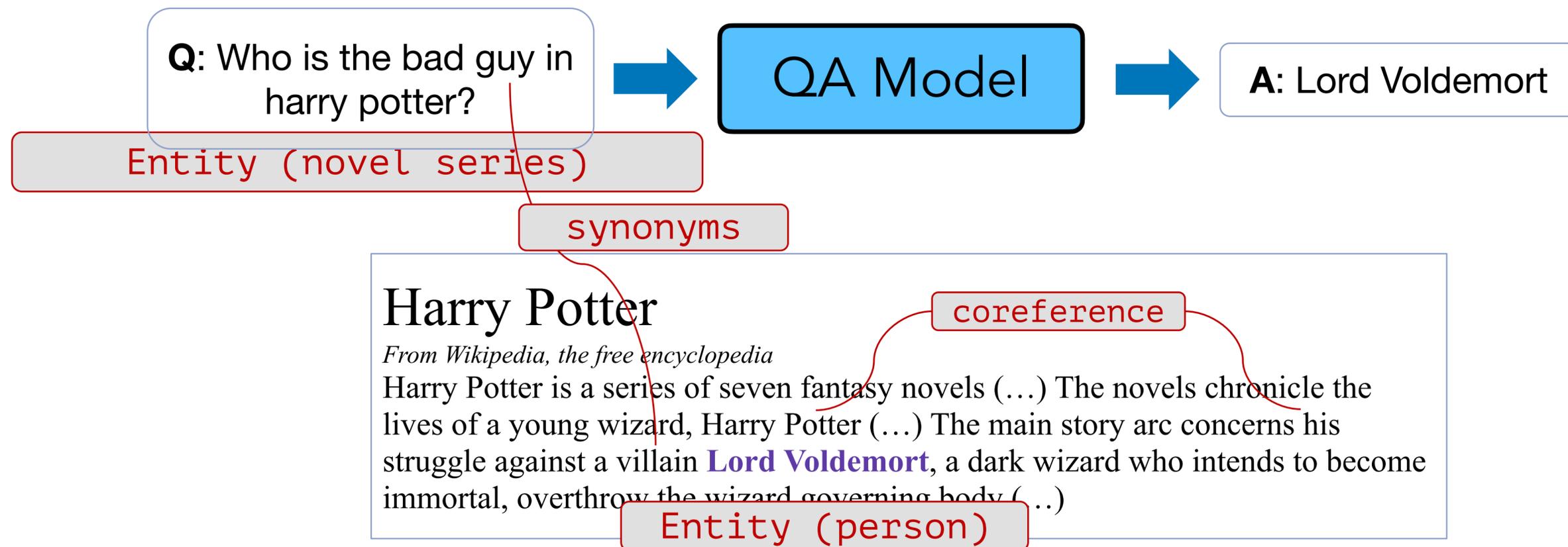
Question answering

- Key element in a wide range of applications (search engines, personal assistants, ...)



Question answering

- Key element in a wide range of applications (search engines, personal assistants, ...)
- Evaluate the progress for natural language understanding (NLU)

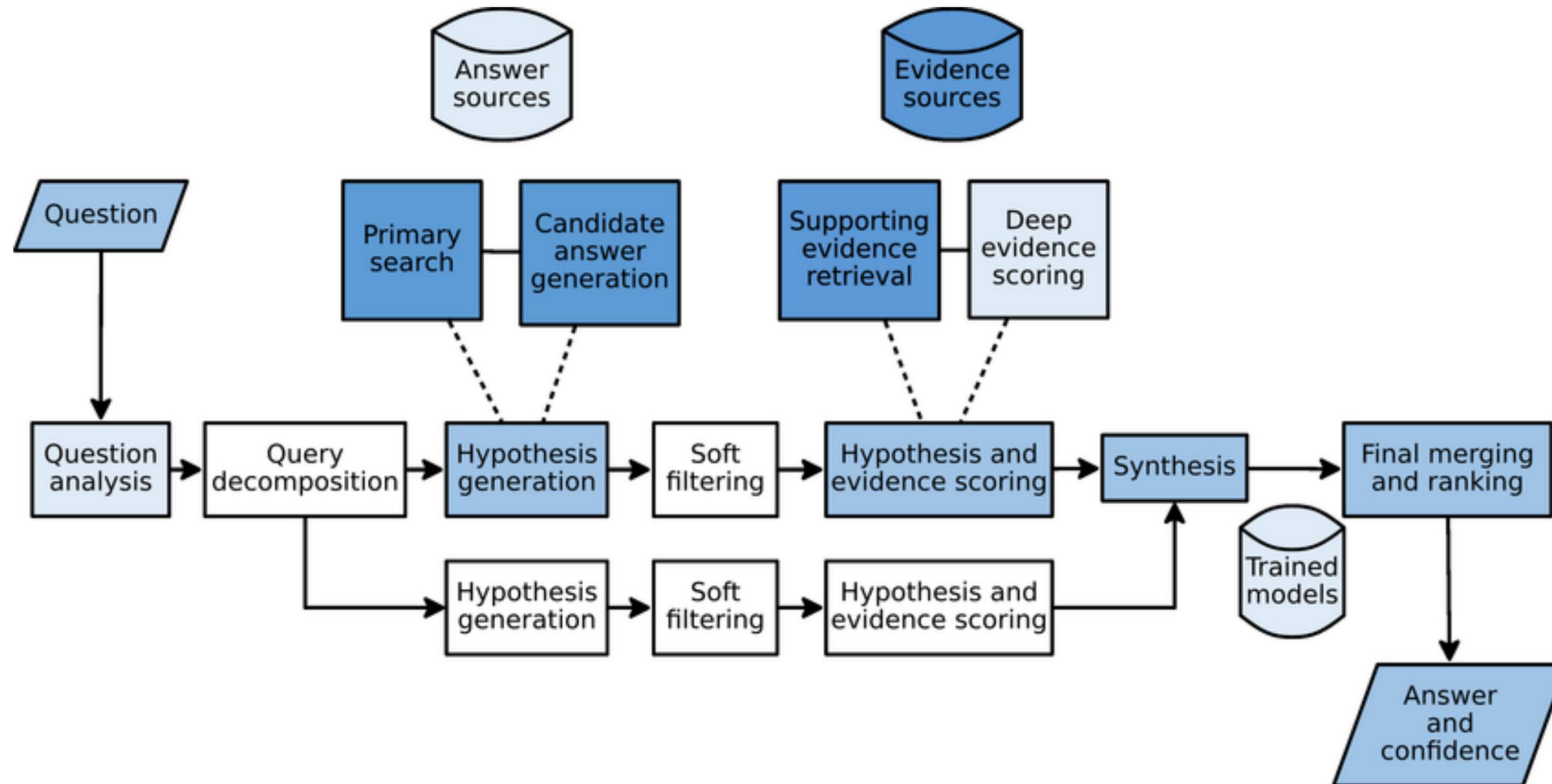


IBM Watson



Two “Jeopardy!” champions, Ken Jennings, left, and Brad Rutter, competed against a computer named Watson, which proved adept at buzzing in quickly. Carol Kaelson/Jeopardy Productions Inc., via Associated

IBM Watson



Earlier QA task: Passage provided

(Also called reading comprehension)

Q: Who is the villain
in Harry Potter?

Harry Potter

From Wikipedia, the free encyclopedia

Harry Potter is a series of seven fantasy novels (...)

Harry Potter (...) The main story arc concerns
his struggle against a villain Lord Voldemort,

Model

A: Voldemort

CNN/Daily Mail Reading Comprehension (2015)

Passage

(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

Question

characters in " @placeholder " movies have gradually become more diverse

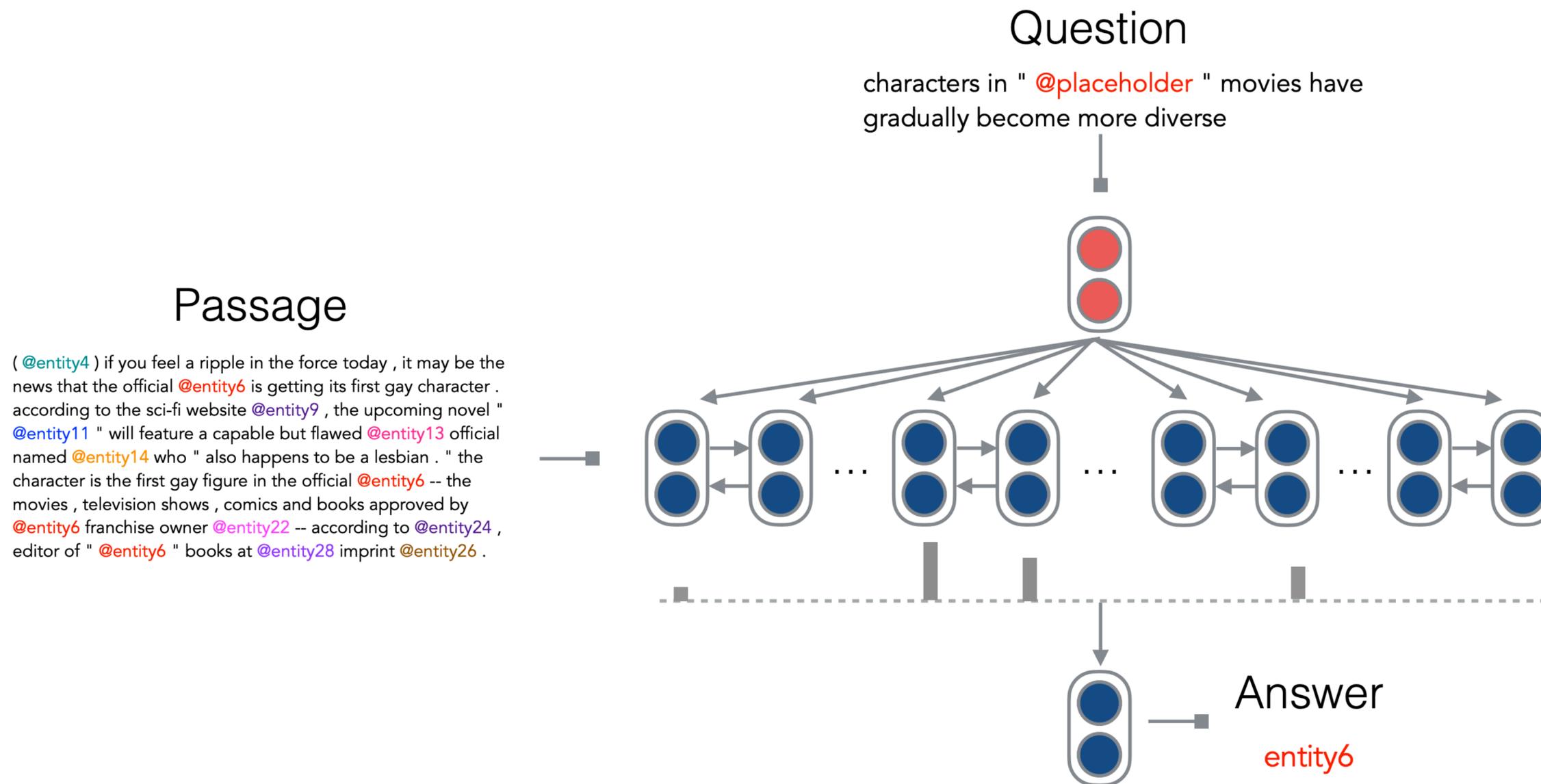
Answer

@entity6

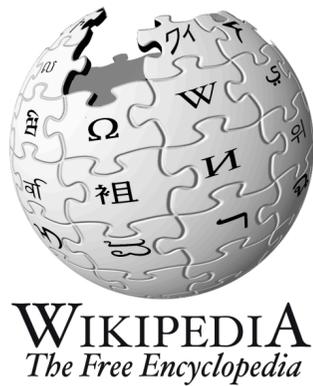
Hermann et al. 2015. "Teaching Machines to Read and Comprehend"

Chen et al. 2016. "A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task"

CNN/Daily Mail Reading Comprehension (2015) Example Model (GRUs, Attention)



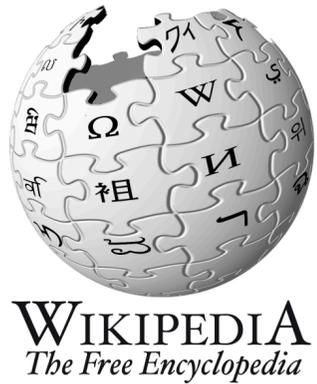
SQuAD (2016)



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".

Question: What causes precipitation to fall?

SQuAD (2016)

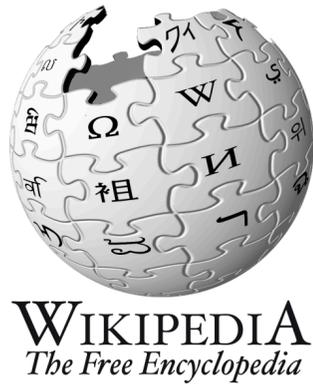


In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice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

Question: What causes precipitation to fall?

Answer: **gravity**

SQuAD (2016)



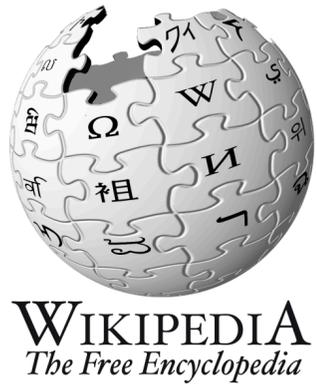
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice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

Question: What causes precipitation to fall?

Answer: **gravity**

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

SQuAD (2016)



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".

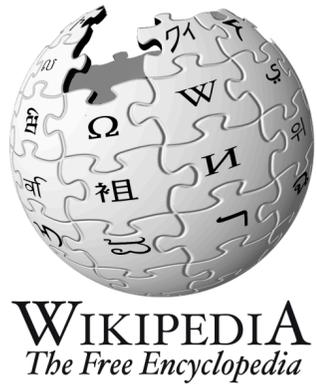
Question: What causes precipitation to fall?

Answer: **gravity**

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

Answer: **graupel**

SQuAD (2016)



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".

Question: What causes precipitation to fall?

Answer: **gravity**

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

Answer: **graupel**

Question: Where do water droplets collide with ice crystals to form precipitation?

Answer: **within a cloud**

Short answers, extractable from the given paragraph

SQuAD (2016)

Paragraph 1 of 43

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O₂.

2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

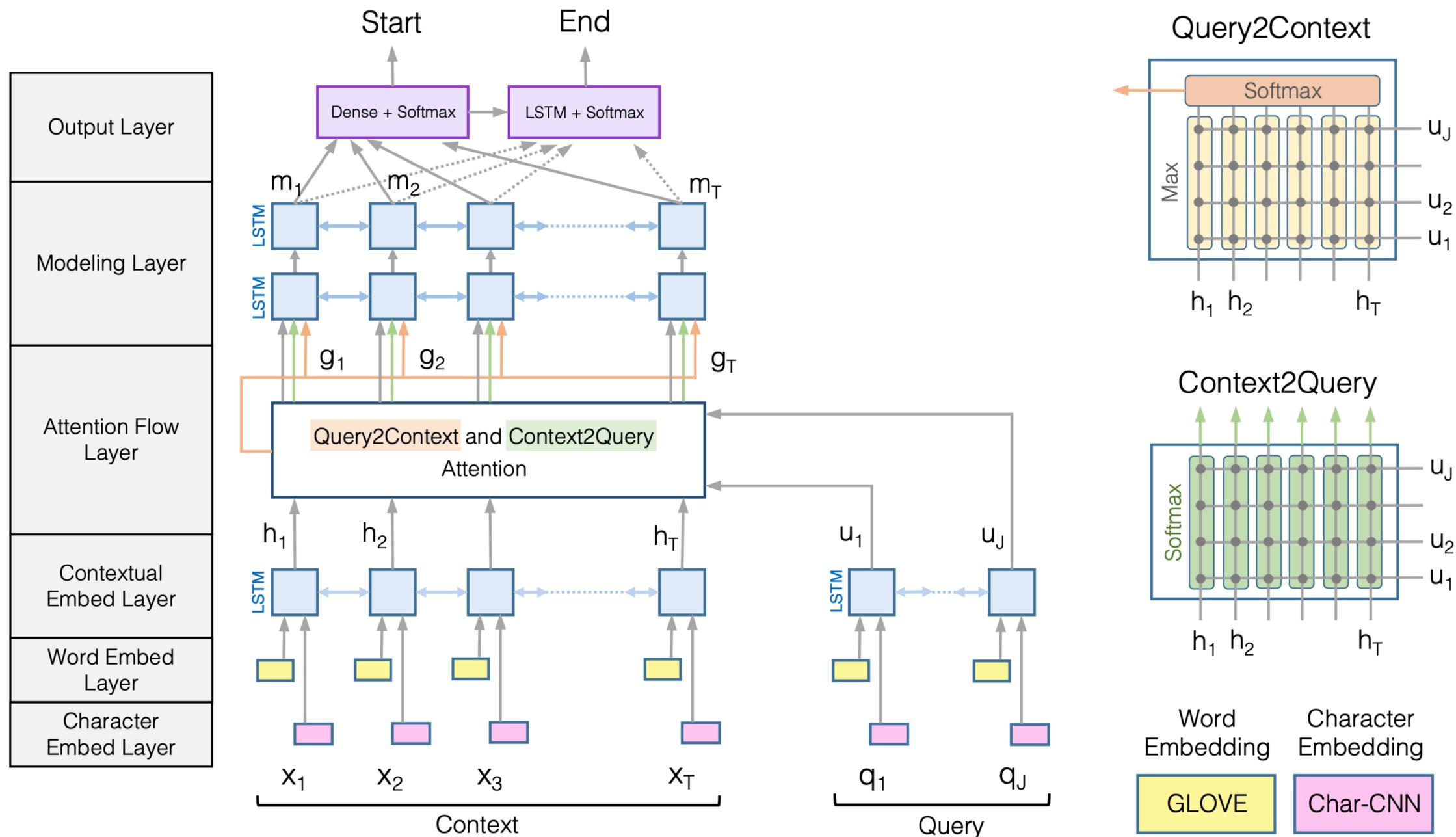
Ask a question here. Try using your own words

Select Answer

Ask a question here. Try using your own words

Select Answer

SQuAD Example Model (LSTMs, Bi-directional Attention)



SQuAD: Evaluation metrics

Exact Match (EM): Percentage of predictions that exactly match any ground-truth answer

F1 Score: Measures token overlap between prediction and ground truth

- Treat answers as bags of tokens and compute F1 for each question
- Final score: average across all questions

Quick quiz (ignore punctuations):

Ground-truth: “Boston, Massachusetts”

Model prediction: “Boston”

- EM = ?
- F1 = ?

SQuAD: Evaluation metrics

Exact Match (EM): Percentage of predictions that exactly match any ground-truth answer

F1 Score: Measures token overlap between prediction and ground truth

- Treat answers as bags of tokens and compute F1 for each question
- Final score: average across all questions

Quick quiz (ignore punctuations):

Ground-truth: “Boston, Massachusetts”

Model prediction: “Boston”

• EM = 0.0

• F1 = $2/3$ (Precision=1, Recall $1/2$)

This span-based answer formulation + metrics become standard in the field

Do they truly require *comprehension*?

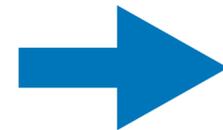
Harry Potter

From Wikipedia, the free encyclopedia

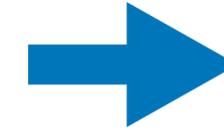
Harry Potter is a series of seven fantasy novels (...)
Harry Potter (...) The main story arc concerns his struggle against a villain Lord Voldemort, a dark wizard who intends to become immortal,



Who is the villain in Harry Potter?



Model



A: Lord Voldemort

You might be able to find answers **without deep natural language understanding**,
due to **annotation artifacts**

Chen et al. ACL 2016, Jia et al. EMNLP 2017, Kaushik & Lipton EMNLP 2018, Min et al. ACL 2019, and more

Do they truly require *comprehension*?

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.

Question: What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Ground-truth: John Elway

Original Prediction: John Elway

Do they truly require *comprehension*?

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?

Ground-truth: John Elway

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Multi-hop QA

Question: What is the former name of the animal whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?

(10 paragraphs given)

Paragraph 5: The Lomako Forest Reserve is found in Democratic Republic of the Congo. It was established in 1991 especially to protect the habitat of the Bonobo apes.

Paragraph 1: The bonobo (“Pan paniscus”), formerly called the pygmy chimpanzee and less often, the dwarf or gracile chimpanzee, is an endangered great ape and one of the two species making up the genus “Pan”.

- Sub-question 1: Which animal is the one whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?
- Sub-question 2: What is the former name of [ANS]?

Multi-hop QA

Question: What is the former name of the animal whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?

(10 paragraphs given)

Paragraph 5: The Lomako Forest Reserve is found in Democratic Republic of the Congo. It was established in 1991 especially to protect the habitat of **the Bonobo apes**.

Paragraph 1: The bonobo (“Pan paniscus”), formerly called the pygmy chimpanzee and less often, the dwarf or gracile chimpanzee, is an endangered great ape and one of the two species making up the genus “Pan”.

- Sub-question 1: Which animal is the one whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?
- Sub-question 2: What is the former name of [ANS]?

Multi-hop QA

Question: What is the former name of the animal whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?

(10 paragraphs given)

Paragraph 5: The Lomako Forest Reserve is found in Democratic Republic of the Congo. It was established in 1991 especially to protect the habitat of **the Bonobo apes**.

Paragraph 1: The bonobo (“Pan paniscus”), formerly called the **pygmy chimpanzee** and less often, the dwarf or gracile chimpanzee, is an endangered great ape and one of the two species making up the genus “Pan”.

- Sub-question 1: Which animal is the one whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?
- Sub-question 2: What is the former name of [ANS]?

Multi-hop QA

Shortcuts still exist!

Question: What is the former name of the animal whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?

(10 paragraphs given)

Paragraph 5: The Lomako Forest Reserve is found in Democratic Republic of the Congo. It was established in 1991 especially to protect the habitat of **the Bonobo apes**.

Paragraph 1: The bonobo (“Pan paniscus”), formerly called the **pygmy chimpanzee** and less often, the dwarf or gracile chimpanzee, is an endangered great ape and one of the two species making up the genus “Pan”.

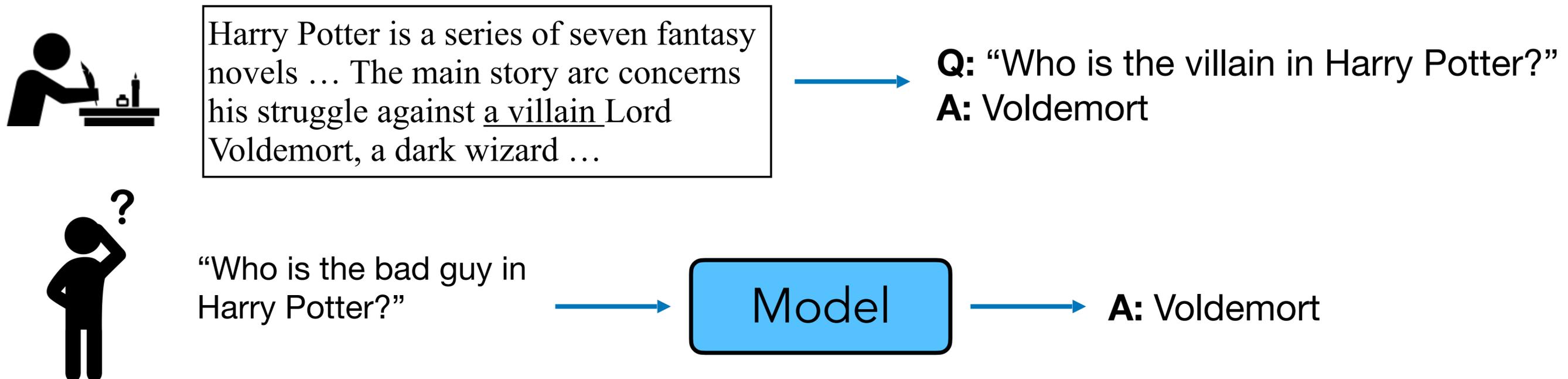
“[...] designed to require reasoning across two paragraphs. [...] However, **since only one of the ten paragraphs is about an animal, one can immediately locate the answer in Paragraph 1** using one hop.”

Yang et al. 2018. "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering"
Min et al. 2019. "Multi-hop Reading Comprehension through Question Decomposition and Rescoring"
Min et al. 2019. "Compositional Questions Do Not Necessitate Multi-hop Reasoning"

Shift Toward Naturally Occurring, Open-Domain QA

Naturally-occurring Qs vs. Annotated Qs

- Previous approaches: Annotators wrote questions with answers in mind → Easily include lexical cues
- New proposal: Let's focus directly on **real questions**, posed by **information-seeking users**.
 - Directly related to end applications
 - Very different distributions from annotated text, little lexical overlap



Natural Questions (NQ; 2019)

Natural Questions: A Benchmark for Question Answering Research

Tom Kwiatkowski^{♣♦♠} Jennimaria Palomaki[♠] Olivia Redfield^{♦♠} Michael Collins^{♣♦♠♥}
Ankur Parikh[♥] Chris Alberti[♥] Danielle Epstein^{♠♦} Illia Polosukhin^{♠♦} Jacob Devlin[♠]
Kenton Lee[♥] Kristina Toutanova[♥] Llion Jones[♠] Matthew Kelcey^{♠♦} Ming-Wei Chang[♥]
Andrew M. Dai^{♣♦} Jakob Uszkoreit[♣] Quoc Le^{♣♦} Slav Petrov[♣]

Google Research
natural-questions@google.com

“Questions consist of real anonymized, aggregated queries issued to the Google search engine.”

Question:

where is the world s largest ice sheet located today?

Short Answer:

Antarctica

*The original NQ provided evidence doc, identified after question collection. Later work removed the evidence document and required models to retrieve it from Wikipedia.

Open-domain QA

Q: Who is the villain in Harry Potter?

Harry Potter
From Wikipedia, the free encyclopedia
Harry Potter is a series of seven fantasy novels by J.K. Rowling. Harry Potter (...)

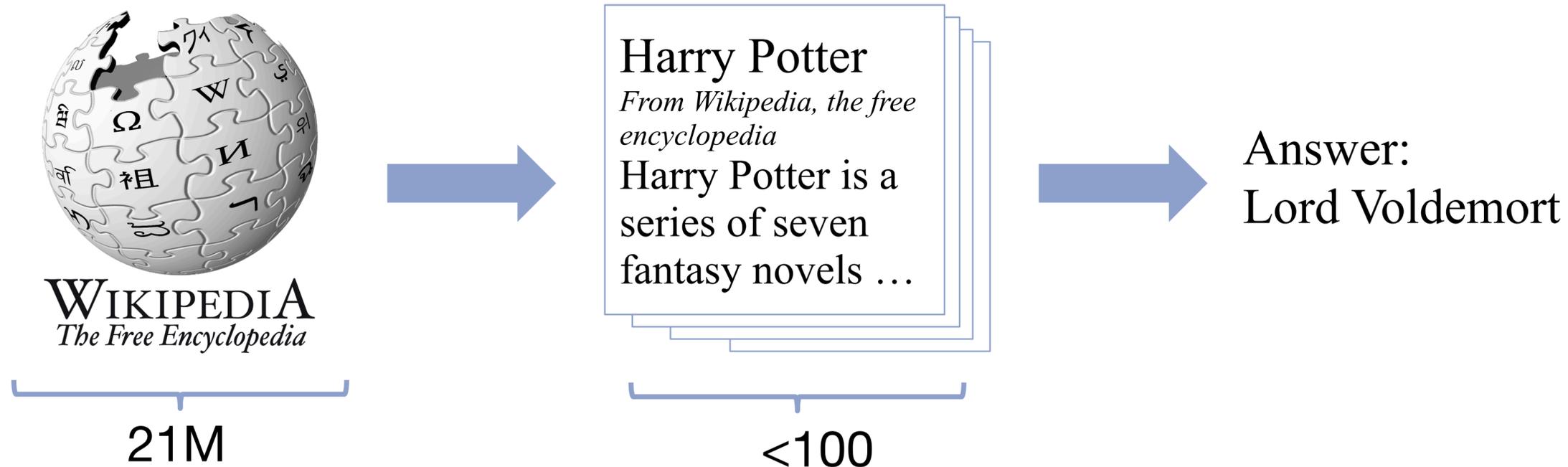
Model

A: Voldemort

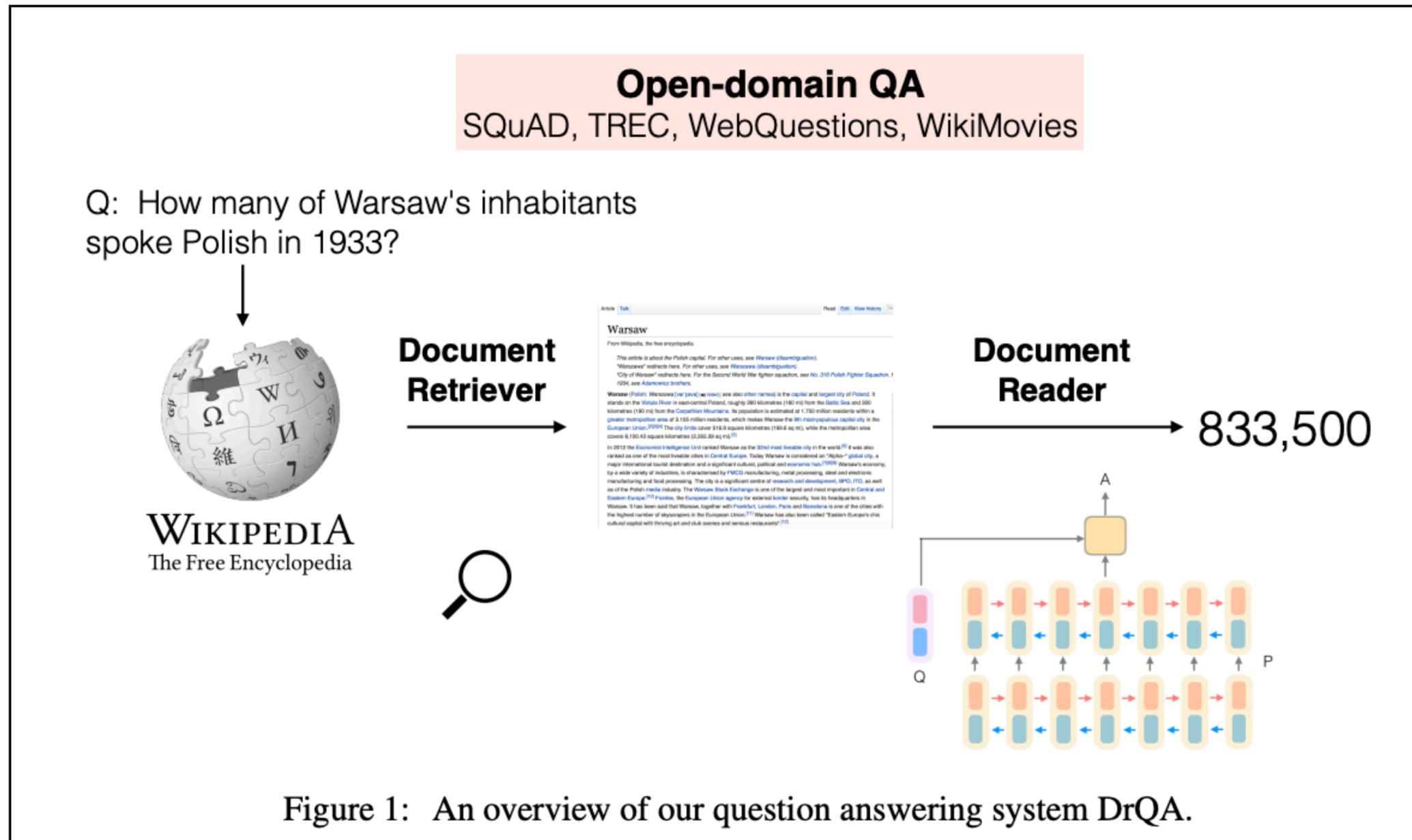
Now that we don't have evidence document, where do we find it?

Simple idea: Let's use retrieval

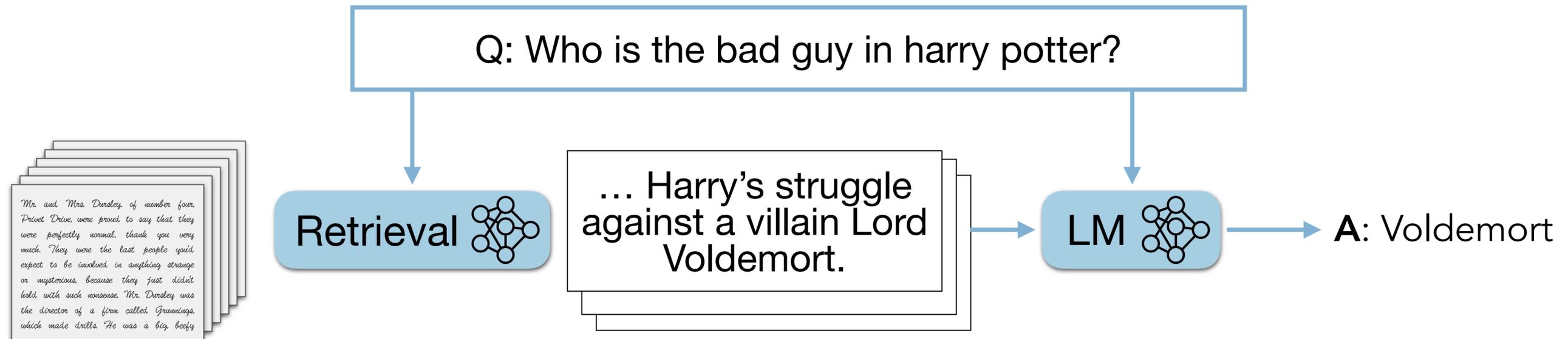
- Ingredient: A text corpus containing world knowledge (e.g., Wikipedia)
- Approach: Retrieval then Answer
 - Retrieval: Given a question, retrieve a small set of documents/passages that contain evidence for the answer
 - Example: retrieve 1–100 passages from Wikipedia (~21M passages)
 - Answering: A neural model takes the retrieved documents/passages and generates the answer



Key paper: DrQA (Chen et al. 2017)



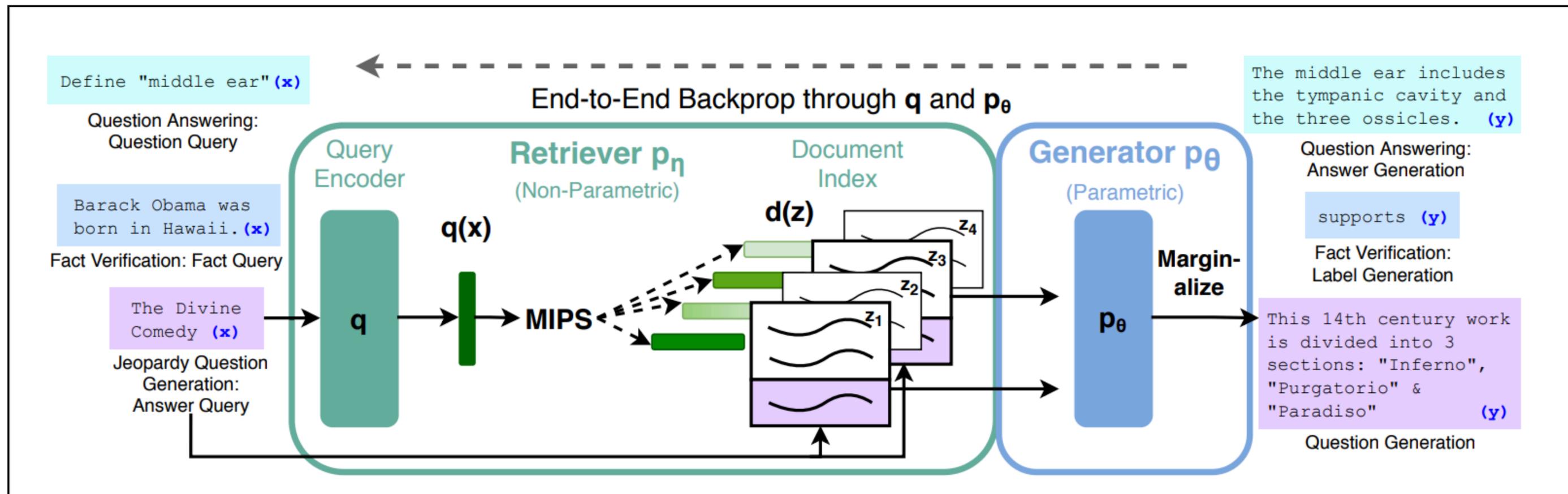
Key paper: DPR (Karpukhin et al. 2020)



Switch lexical-based retrieval to neural retrieval

Switch LSTM-based document reader to BERT-based

Key paper: RAG (Lewis et al. 2020)



- Proposed end-to-end fine-tuning (arguably not really used today)
- Coined the term **Retrieval-Augmented Generation (RAG)**

Inherent properties of naturally-occurring Qs

- Ambiguity: questions may have multiple interpretations
- False presuppositions: questions may assume incorrect facts
- Frequently updated answers: answers may change over time

Q: What season does Meredith and Derek get married in Grey's Anatomy?

Q: What season does Meredith give birth to Zola in Grey's Anatomy?

Now or Never (Grey's Anatomy)

... Season 5 ... Meredith and Derek have decided not to wait any longer to get married and just go to City Hall that evening ... writes their vows down on a post-it note that they both sign ...

Grey's Anatomy (Season 7)

... She and Derek decide to adopt Zola, an orphaned baby, and make their marriage legal.

Another Form of Open-Domain QA: Quizzes and Exams

- Questions are written to test knowledge and comprehension
- Not naturally occurring, but make evaluation easier and more controlled
- Example: TriviaQA (Joshi et al 2017), MMLU (Hendrycks et al. 2020), GPQA (Rein et al. 2023), SimpleQA (Wei et al. 2024), and more

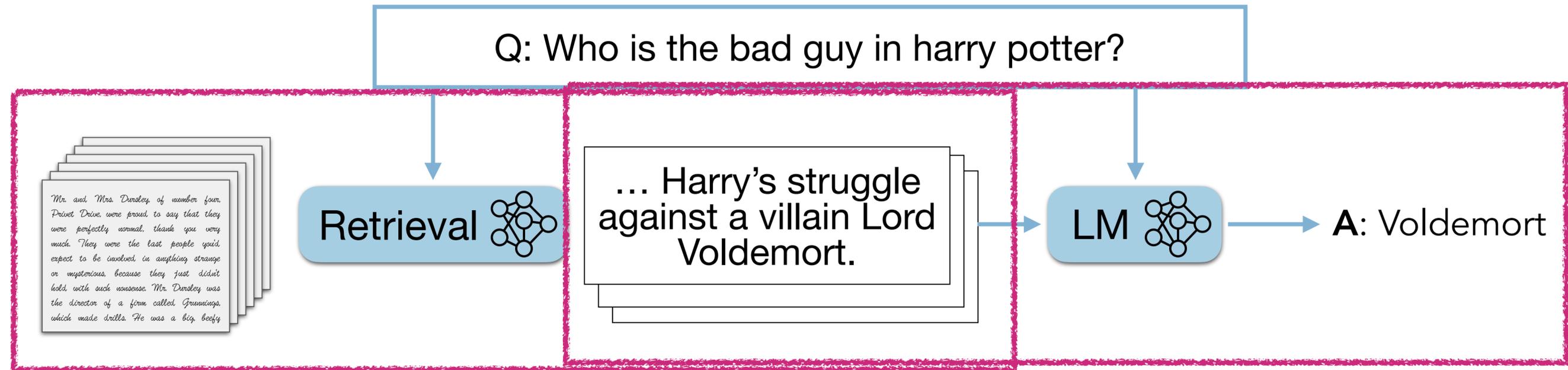
Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

Answer: The Guns of Navarone

Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italianheld Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Retrieval-based LMs (or RAG)

RAG overview

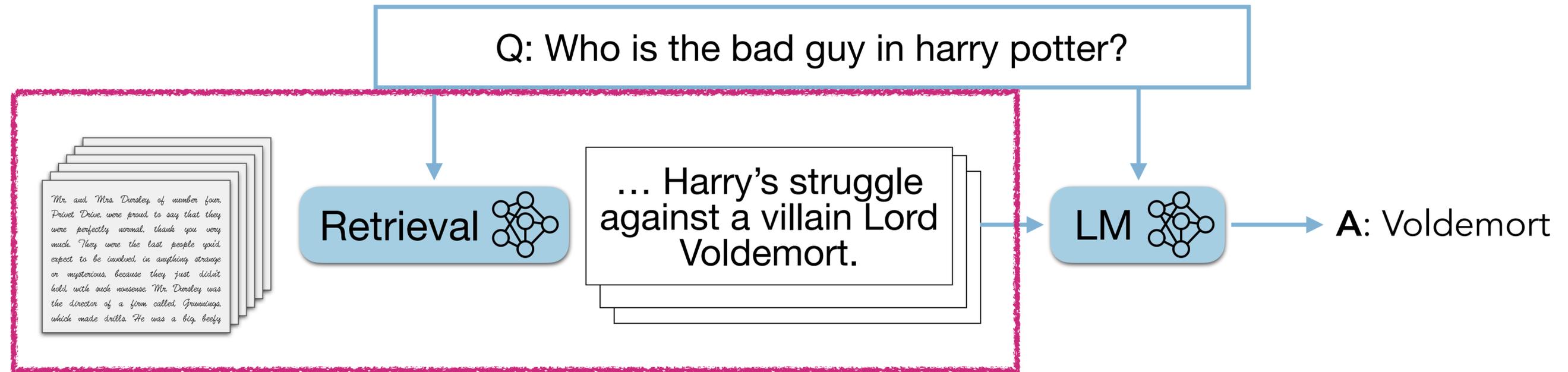


1) Retrieval

2) Augmentation

Simply prepends retrieved docs as additional inputs to an LLM (Very simple!)

RAG overview



1) Retrieval

Retrieval, or Information Retrieval (IR)

- Input: a question, a text corpus (e.g., Wikipedia, which has 21M passages)
- Output: a small number of **(ranked)** text chunks (1 to 100), retrieved from the corpus
 - “Text chunks” can be documents, paragraphs, or fixed-size chunks, typically between 100 to 1000 words.
 - In this lecture, we’ll use documents and paragraphs interchangeably.

Retrieval metrics

Mean Reciprocal Rank (MRR)

- For a single query, find the rank (r) of the gold document. The score is $1/r$.
- MRR = average of these scores across all queries.

Normalized Discounted Cumulative Gain (nDCG)

- Accounts for **graded relevance**, e.g., 3 points for “perfect” documents, and 1 point for “okay” documents
- Cumulative Gain (CG): Sum of the relevance scores
- Discounted (D): penalizes items ranked lower in the list
- Normalized (N): Divide the actual score by the “Ideal” score

Recall@K: The “coverage”

- Checks whether a gold document appears in the top-K retrieved results If yes → counted as a successful retrieval

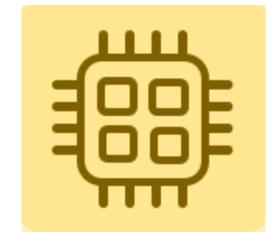
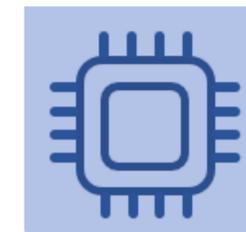
Retrieval methods

- Goal: find a small subset of elements (docs or passages) in a text corpus that are the most similar to the input question
- sim : a similarity score between two pieces of text
 - $\text{sim}(q, d)$: A cosine similarity between $f(q)$ and $f(d)$, where $f(\cdot)$ is a vector representation of text
 - Key challenge in retrieval is how to get (or learn) this f !
- How retrieval typically works:
 - Offline: Given a large collection d_1, \dots, d_N , we pre-compute $f(d_1), \dots, f(d_N)$, and build a **retrieval index** for **fast nearest neighbor search**, so that once q is obtained at run time, finding $\text{argmax}_k \text{sim}(q, d_k)$ is super fast.
 - Fast nearest neighbor search is itself a large research area! (Tradeoffs between speed, accuracy, and memory.)

Software: FAISS, Distributed FAISS, SCaNN, etc...

Method	Class name	index_factory	Main parameters	Bytes/vector	Exhaustive	Comments
Exact Search for L2	IndexFlatL2	"Flat"	d	4*d	yes	brute-force
Exact Search for Inner Product	IndexFlatIP	"Flat"	d	4*d	yes	also for cosine (normalize vectors beforehand)
Hierarchical Navigable Small World graph exploration	IndexHNSWFlat	"HNSW,Flat"	d, M	$4*d + x * M * 2 * 4$	no	
Inverted file with exact post-verification	IndexIVFFlat	"IVFx,Flat"	quantizer, d, nlists, metric	4*d + 8	no	Takes another index to assign vectors to inverted lists. The 8 additional bytes are the vector id that needs to be stored.
Locality-Sensitive Hashing (binary flat index)	IndexLSH	-	d, nbits	ceil(nbits/8)	yes	optimized by using random rotation instead of random projections
Scalar quantizer (SQ) in flat mode	IndexScalarQuantizer	"SQ8"	d	d	yes	4 and 6 bits per component are also implemented.
Product quantizer (PQ) in flat mode	IndexPQ	"PQx", "PQM"x"nbits"	d, M, nbits	ceil(M * nbits / 8)	yes	
IVF and scalar quantizer	IndexIVFScalarQuantizer	"IVFx,SQ4" "IVFx,SQ8"	quantizer, d, nlists, qtype	SQfp16: $2 * d + 8$, SQ8: $d + 8$ or SQ4: $d/2 + 8$	no	Same as the IndexScalarQuantizer
IVFADC (coarse quantizer+PQ on residuals)	IndexIVFPQ	"IVFx,PQ"y"x"nbits"	quantizer, d, nlists, M, nbits	ceil(M * nbits/8)+8	no	
IVFADC+R (same as IVFADC with re-ranking based on codes)	IndexIVFPQR	"IVFx,PQy+z"	quantizer, d, nlists, M, nbits, M_refine, nbits_refine	M+M_refine+8	no	

Exact Search



CPU vs. GPU

Approximated Search
(Relatively easy to scale to ~1B vectors)

More info: <https://github.com/facebookresearch/faiss/wiki>

Retrieval methods

- Goal: find a small subset of elements (docs or passages) in a text corpus that are the most similar to the input question
- **sim**: a similarity score between two pieces of text
 - $\text{sim}(q, d)$: A cosine similarity between $f(q)$ and $f(d)$, where $f(\cdot)$ is a vector representation of text
 - Key challenge in retrieval is how to get (or learn) this f !
- How retrieval typically works:
 - Offline: Given a large collection d_1, \dots, d_N , we build a **retrieval index** for **fast nearest neighbor search**. Finding $\text{argmax}_k \text{sim}(q, d_k)$ is super fast.
 - Fast nearest neighbor search is itself a large research area! (Tradeoffs between speed, accuracy, and memory.)

- Lexical-based (sparse) retrieval
- Neural (dense) retrieval

Lexical-matching (sparse) retrieval

- Considers lexical overlap between a question and a document.

- Q: Who is the author of Harry Potter and the Philosopher's Stone?
- Doc: Harry Potter and the Philosopher's Stone is a fantasy novel written by British author J.K. Rowling.

- TF-IDF

1. Each word is assigned a TF-IDF score, which is a multiplication of Term Frequency (TF) and Inverse Document Frequency (IDF).
 - IDF penalizes common words like “the”, and rewards rare, highly specific keywords like “philosopher”.
2. Represent query and document as V -dimensional vectors (the i -th element = TF-IDF score of word i) — **These vectors are sparse**
3. Compute **cosine similarity** between query and document vectors.

- BM25: Improved version of TF-IDF that stops giving extra points once a word appears enough times and adjusts scores for the document length

Lexical-matching (sparse) retrieval

- Considers lexical overlap between a question and a document.

- Q: Who is the author of Harry Potter and the Philosopher's Stone?
- Doc: Harry Potter and the Philosopher's Stone is a fantasy novel written by British author J.K. Rowling.

- Q: Who is the author of harry potter and the sorcerer's stone?
- Doc: Harry Potter and the Philosopher's Stone is a fantasy novel written by British author J.K. Rowling.

- Q: Who is the bad guy in harry potter?
- Doc: The main story arc concerns Harry's struggle against a villain Lord Voldemort.

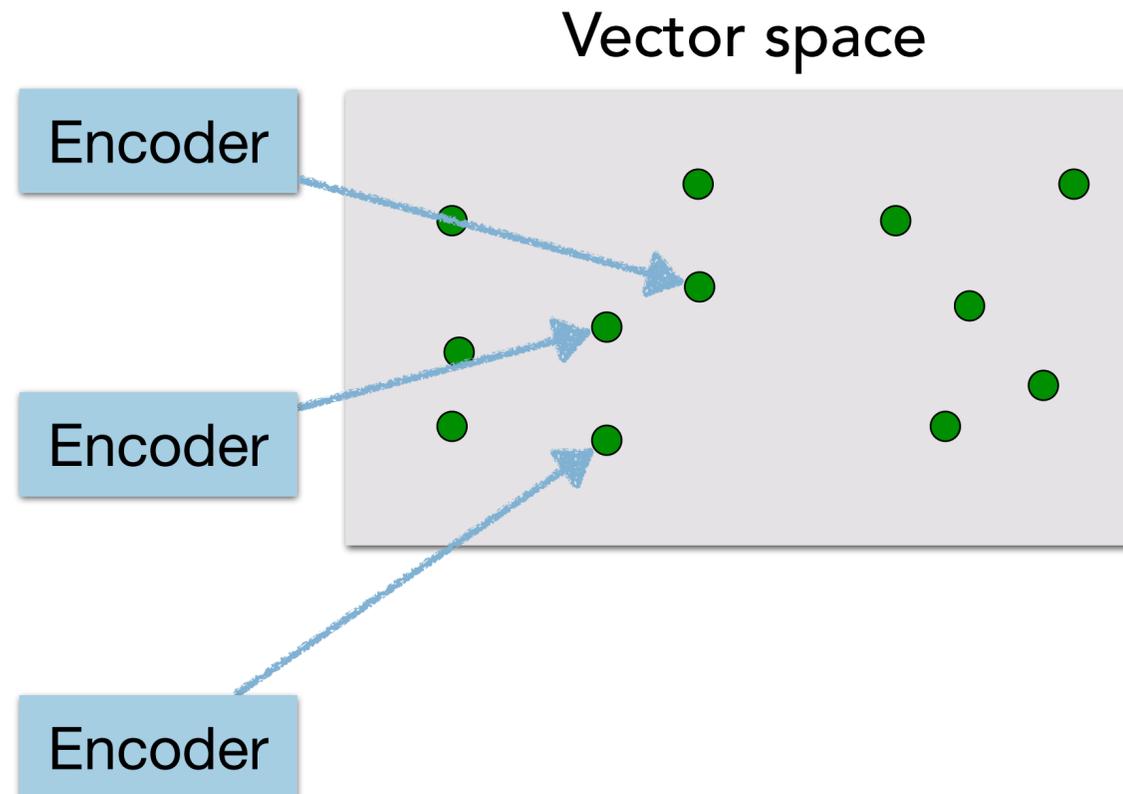
- Lexical-matching retrieval is a very strong baseline!
- However, does not consider semantic meaning.

Neural (dense) retrieval

As Harry shouted,
"Expelliarmus!"
Voldemort cried,
"Avada Kedavra!" A

just as a jet of
red light blasted
from Harry's –
they met in midair

How can a jet of
water be powerful
enough to cut
through steel?



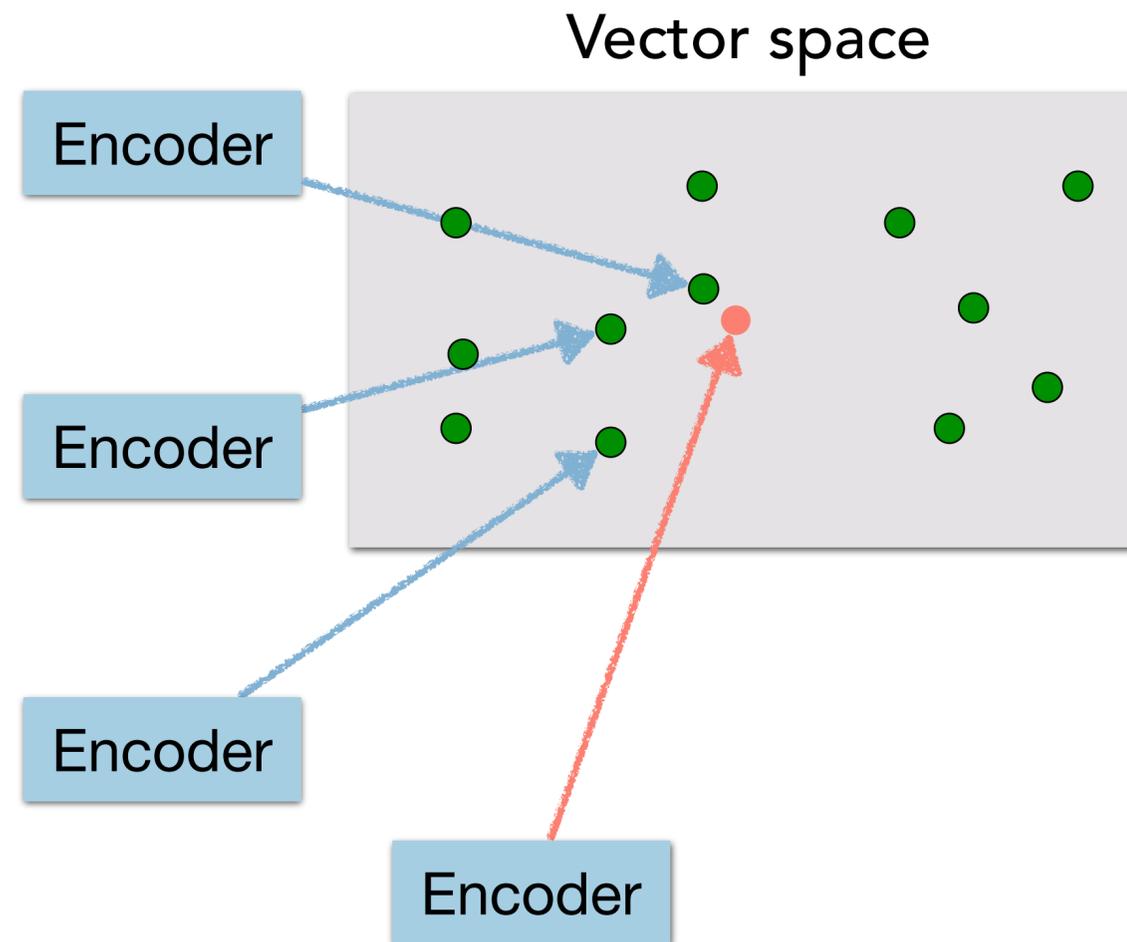
$$\mathbf{z}_i = \text{Enc}(z_i) \in \mathbb{R}^h \quad (1 \leq i \leq N)$$

Neural (dense) retrieval

As Harry shouted,
"Expelliarmus!"
Voldemort cried,
"Avada Kedavra!" A

just as a jet of
red light blasted
from Harry's –
they met in midair

How can a jet of
water be powerful
enough to cut
through steel?



$$\mathbf{z}_i = \text{Enc}(z_i) \in \mathbb{R}^h \quad (1 \leq i \leq N)$$

$$\mathbf{x} = \text{Enc}(x) \in \mathbb{R}^h$$

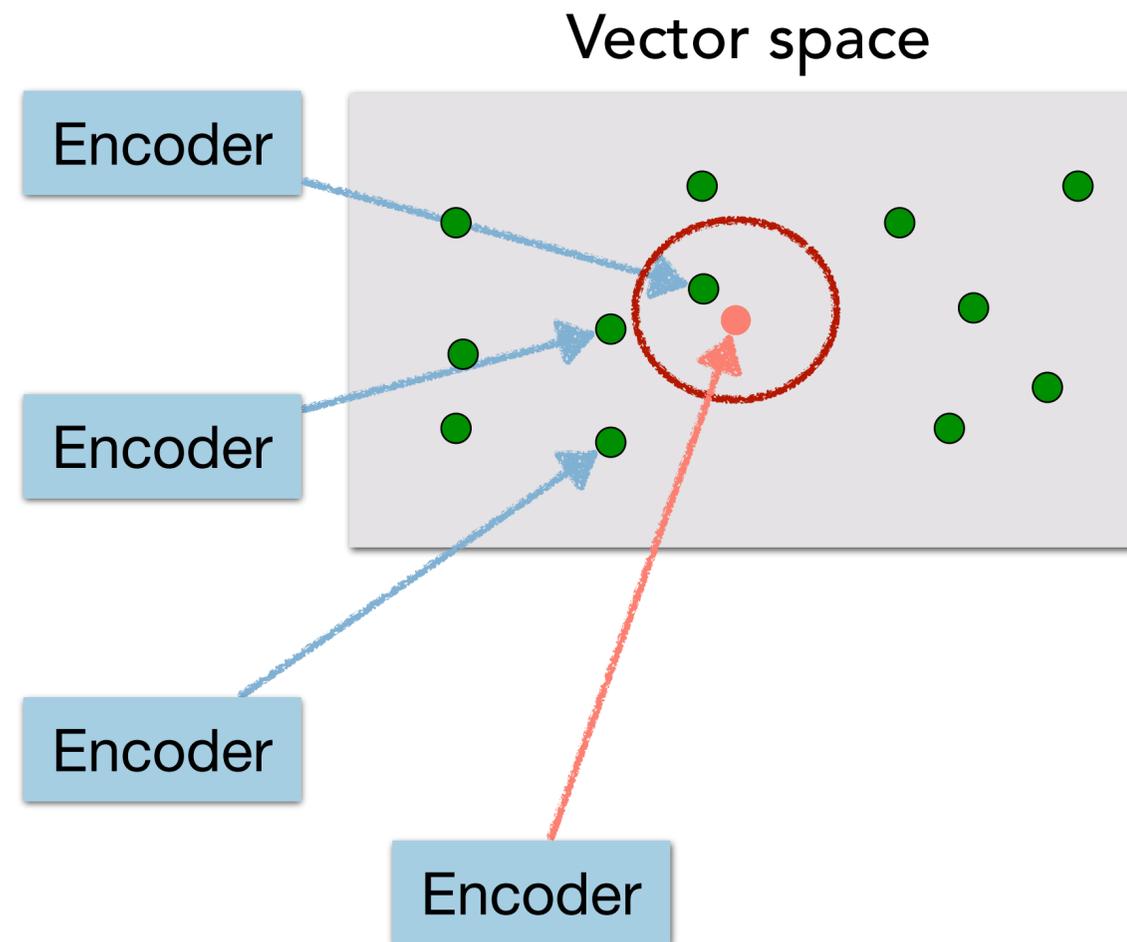
Question: What color is the Killing Curse in Harry Potter?

Neural (dense) retrieval

As Harry shouted,
"Expelliarmus!"
Voldemort cried,
"Avada Kedavra!" A

just as a jet of
red light blasted
from Harry's –
they met in midair

How can a jet of
water be powerful
enough to cut
through steel?



$$\mathbf{z}_i = \text{Enc}(z_i) \in \mathbb{R}^h \quad (1 \leq i \leq N)$$

$$\mathbf{x} = \text{Enc}(x) \in \mathbb{R}^h$$

$$\text{sim}(x, z_i) = \mathbf{x}^\top \mathbf{z}_i$$

Question: What color is the Killing Curse in Harry Potter?

Neural (dense) retrieval history

- Studied for decades
 - “Siamese” network (Bromley et al. 1993, Chopra et al 2005, Yih et al 2011, Huang et al 2013)
- No good recipe for training the encoder
 - Required massive compute & labeled dataset
 - Not much better than alternatives (e.g. lexical-matching)
- Neural retrieval suddenly started to work in 2020
 - Pre-trained encoders such as BERT
 - Development of improved learning objectives (DPR and its variants)

Dense Passage Retrieval for Open-Domain Question Answering

Vladimir Karpukhin*, Barlas Oğuz*, Sewon Min[†], Patrick Lewis,

Ledell Wu, Sergey Edunov, Danqi Chen[‡], Wen-tau Yih

Facebook AI [†]University of Washington [‡]Princeton University

March 10 lecture starts from here

CS 288 Advanced Natural Language Processing

Course website: cal-cs288.github.io/sp26

Ed: edstem.org/us/join/XvztdK

- Class starts at 15:40!
- Reminder to refresh the lecture slides (PDFs) on the lecture day
- Check out updates on A3
- Lecture plans:
 - Retrieval and RAG (40 min)
 - Advanced architectures (40min)

Recap: Retrieval methods

- Goal: find a small subset of elements (docs or passages) in a text corpus that are the most similar to the input question
- **sim**: a similarity score between two pieces of text
 - $\text{sim}(q, d)$: A cosine similarity between $f(q)$ and $f(d)$, where $f(\cdot)$ is a vector representation of text
 - Key challenge in retrieval is how to get (or learn) this f !
- How retrieval typically works:
 - Offline: Given a large collection d_1, \dots, d_N , we build a **retrieval index** for **fast nearest neighbor search**. Finding $\text{argmax}_k \text{sim}(q, d_k)$ is super fast.
 - Fast nearest neighbor search is itself a large research area! (Tradeoffs between speed, accuracy, and memory.)

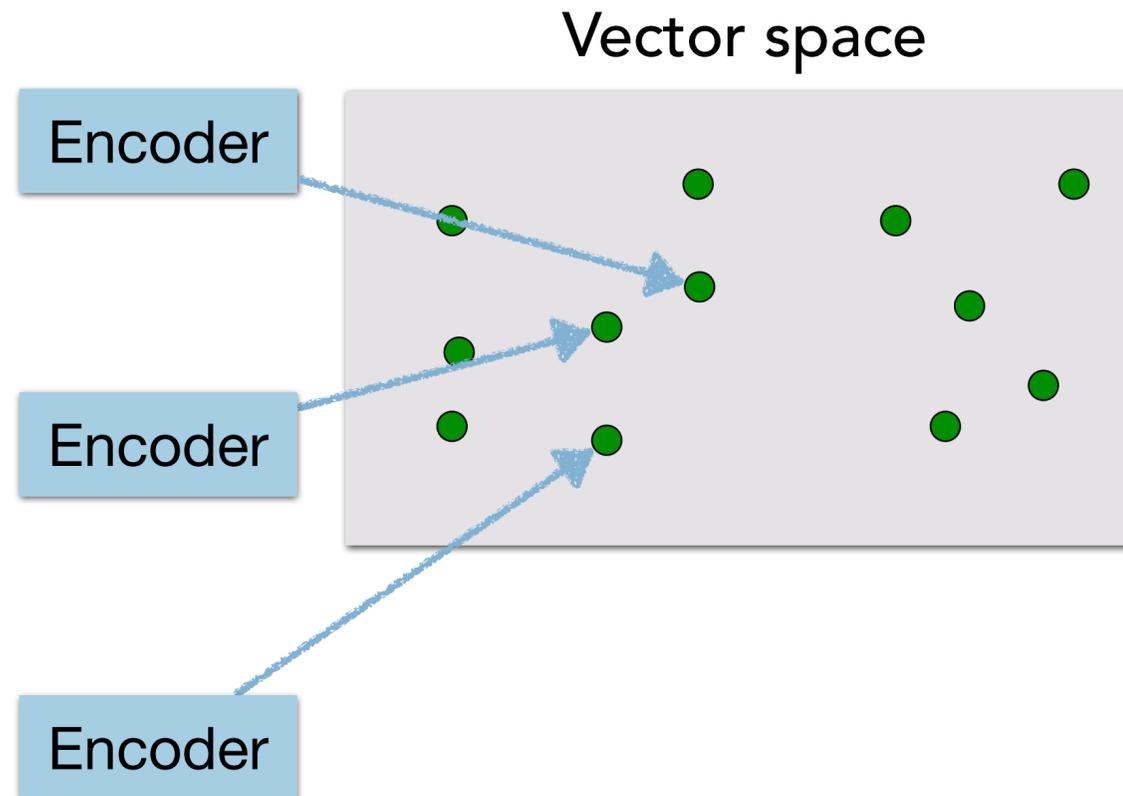
- Lexical-based (sparse) retrieval
- Neural (dense) retrieval

Recap: Neural (dense) retrieval

As Harry shouted,
"Expelliarmus!"
Voldemort cried,
"Avada Kedavra!" A

just as a jet of
red light blasted
from Harry's –
they met in midair

How can a jet of
water be powerful
enough to cut
through steel?



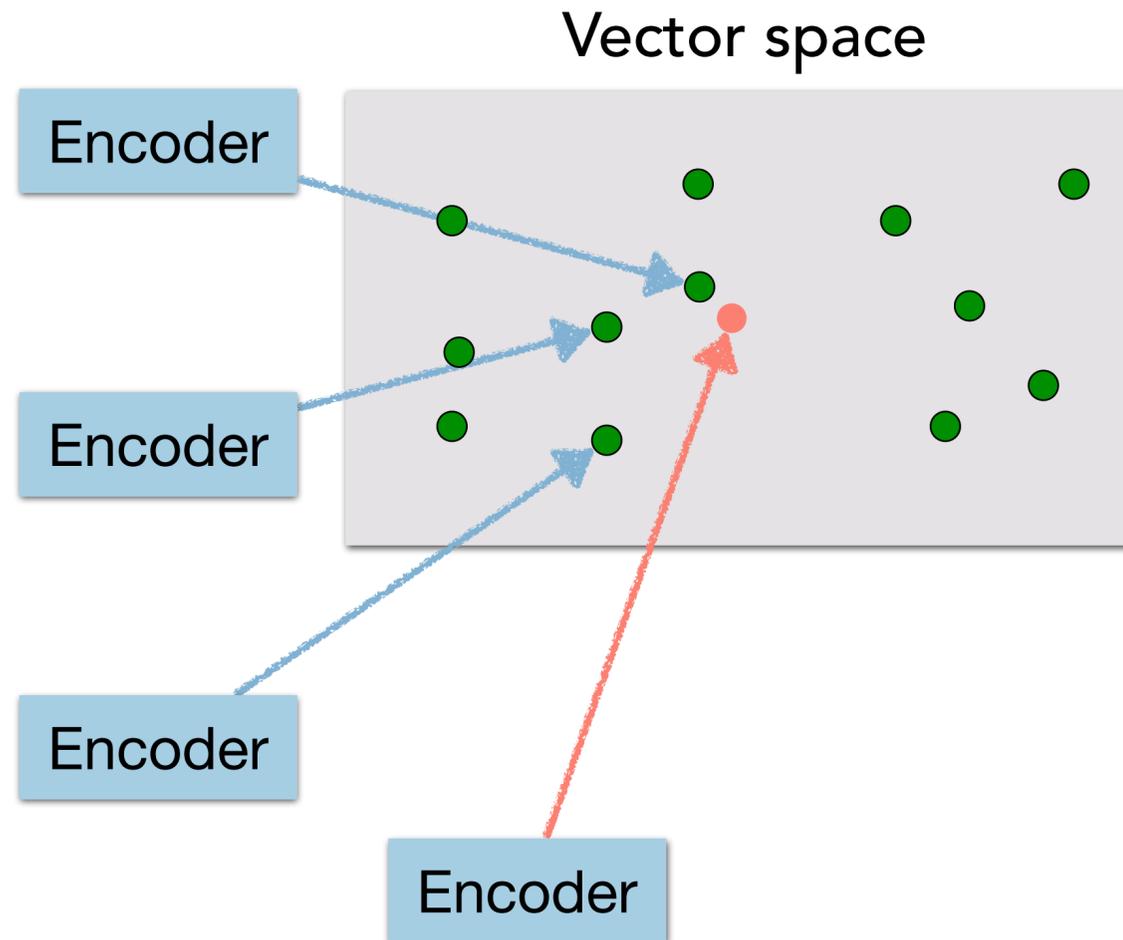
$$\mathbf{z}_i = \text{Enc}(z_i) \in \mathbb{R}^h \quad (1 \leq i \leq N)$$

Recap: Neural (dense) retrieval

As Harry shouted,
"Expelliarmus!"
Voldemort cried,
"Avada Kedavra!" A

just as a jet of
red light blasted
from Harry's –
they met in midair

How can a jet of
water be powerful
enough to cut
through steel?



$$\mathbf{z}_i = \text{Enc}(z_i) \in \mathbb{R}^h \quad (1 \leq i \leq N)$$

$$\mathbf{x} = \text{Enc}(x) \in \mathbb{R}^h$$

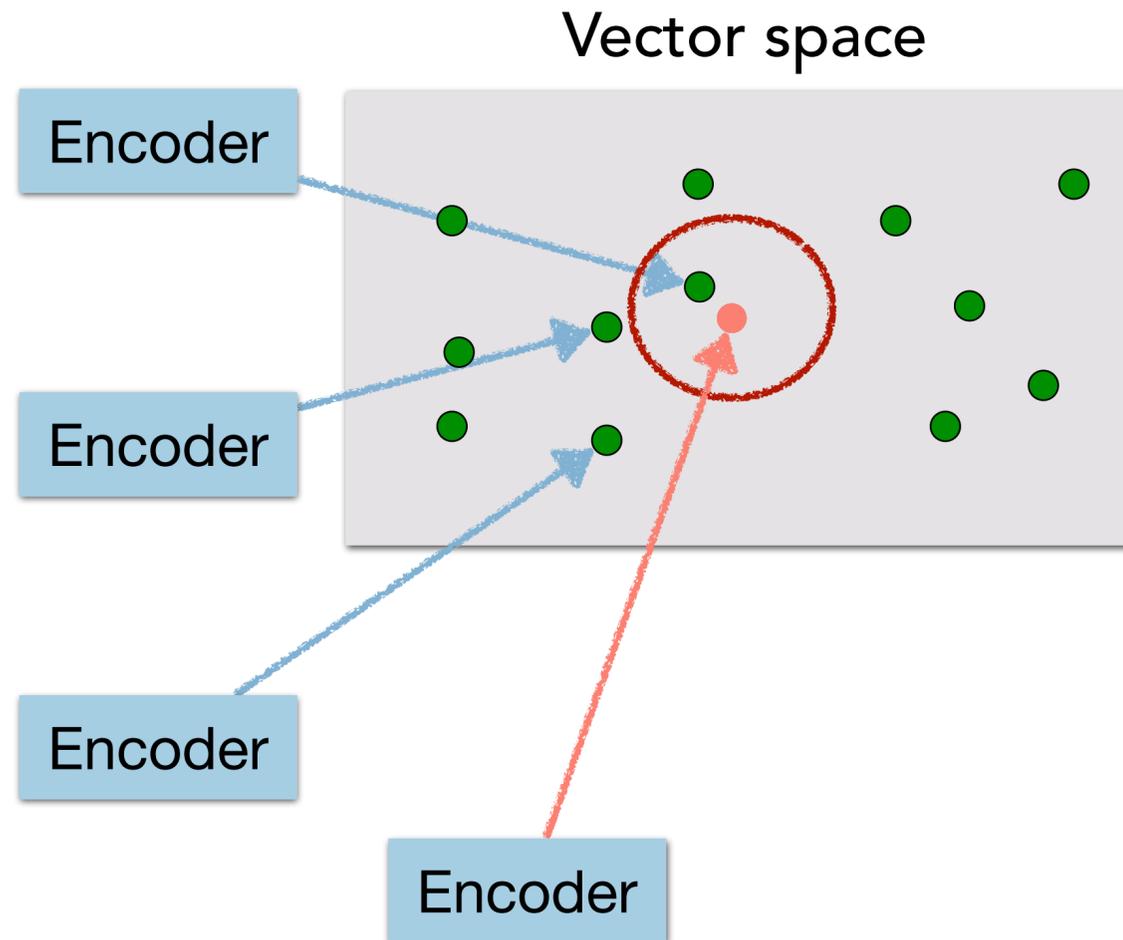
Question: What color is the Killing Curse in Harry Potter?

Recap: Neural (dense) retrieval

As Harry shouted,
"Expelliarmus!"
Voldemort cried,
"Avada Kedavra!" A

just as a jet of
red light blasted
from Harry's –
they met in midair

How can a jet of
water be powerful
enough to cut
through steel?



$$\mathbf{z}_i = \text{Enc}(z_i) \in \mathbb{R}^h \quad (1 \leq i \leq N)$$

$$\mathbf{x} = \text{Enc}(x) \in \mathbb{R}^h$$

$$\text{sim}(x, z_i) = \mathbf{x}^\top \mathbf{z}_i$$

Question: What color is the Killing Curse in Harry Potter?

Recap: Neural (dense) retrieval history

- Studied for decades
 - “Siamese” network (Bromley et al. 1993, Chopra et al 2005, Yih et al 2011, Huang et al 2013)
- No good recipe for training the encoder
 - Required massive compute & labeled dataset
 - Not much better than alternatives (e.g. lexical-matching)
- Neural retrieval suddenly started to work in 2020
 - Pre-trained encoders such as BERT
 - Development of improved learning objectives (DPR and its variants)

Dense Passage Retrieval for Open-Domain Question Answering

Vladimir Karpukhin*, Barlas Oğuz*, Sewon Min†, Patrick Lewis,

Ledell Wu, Sergey Edunov, Danqi Chen‡, Wen-tau Yih

Facebook AI

†University of Washington

‡Princeton University

DPR: Methodology

- x : input, z : a doc (or a passage)
- $\text{sim}(x, z) = \text{Enc}(x)^T \text{Enc}(z)$ (representing from the Encoder-only Transformers)
 - Initialized from BERT, then further fine-tuned on labeled dataset, e.g., a collection of $\langle x, z^+ \rangle$, where z^+ is a gold doc (called a “positive”)

$$-\log \frac{e^{\text{sim}(x, z^+)}}{\sum_{i=1}^N e^{\text{sim}(x, z_i)}}$$

← Get close to a **positive**

← Push back **all the others**

DPR: Methodology

- x : input, z : a doc (or a passage)
- $\text{sim}(x, z) = \text{Enc}(x)^T \text{Enc}(z)$ (representing from the Encoder-only Transformers)
 - Initialized from BERT, then further fine-tuned on labeled dataset, e.g., a collection of $\langle x, z^+ \rangle$, where z^+ is a gold doc (called a “positive”)

$$-\log \frac{e^{\text{sim}(x, z^+)}}{\sum_{i=1}^N e^{\text{sim}(x, z_i)}} \approx -\log \frac{e^{\text{sim}(x, z^+)}}{e^{\text{sim}(x, z^+)} + \sum_{i=1}^n e^{\text{sim}(x, z_i^-)}}$$

1) In-batch approximation

Other passages within the batch

Yih et al., 2011, Henderson et al., 2017, Gillick et al., 2019

DPR: Methodology

- x : input, z : a doc (or a passage)
- $\text{sim}(x, z) = \text{Enc}(x)^T \text{Enc}(z)$ (representing from the Encoder-only Transformers)
 - Initialized from BERT, then further fine-tuned on labeled dataset, e.g., a collection of $\langle x, z^+ \rangle$, where z^+ is a gold doc (called a “positive”)

$$-\log \frac{e^{\text{sim}(x, z^+)}}{\sum_{i=1}^N e^{\text{sim}(x, z_i)}} \approx -\log \frac{e^{\text{sim}(x, z^+)}}{e^{\text{sim}(x, z^+)} + \sum_{i=1}^n e^{\text{sim}(x, z_i^-)}}$$

1) In-batch approximation

Negatives in the batch

Yih et al., 2011, Henderson et al., 2017, Gillick et al., 2019

DPR: Methodology

- x : input, z : a doc (or a passage)
- $\text{sim}(x, z) = \text{Enc}(x)^T \text{Enc}(z)$ (representing from the Encoder-only Transformers)
 - Initialized from BERT, then further fine-tuned on labeled dataset, e.g., a collection of $\langle x, z^+ \rangle$, where z^+ is a gold doc (called a “positive”)

$$-\log \frac{e^{\text{sim}(x, z^+)}}{\sum_{i=1}^N e^{\text{sim}(x, z_i)}} \approx -\log \frac{e^{\text{sim}(x, z^+)}}{e^{\text{sim}(x, z^+)} + \sum_{i=1}^n e^{\text{sim}(x, z_i^-)}}$$

1) In-batch approximation

2) **Hard negatives** in the batch

Passages that **challenge** the model, typically obtained by passages with high lexical overlap

Intuition: Why hard negatives?

The model should be trained with negatives that it is likely to be confused about

Q: Who is the bad guy in Harry Potter?

p^+

The main story arc concerns Harry's struggle against a villain Lord

Random negatives

New York is the most populous city in the United States. With an estimated 2020

On 11 February 1963, the Beatles recorded ten songs during a single studio session for

At the 1908 Summer Olympics in London, football became an official competition.

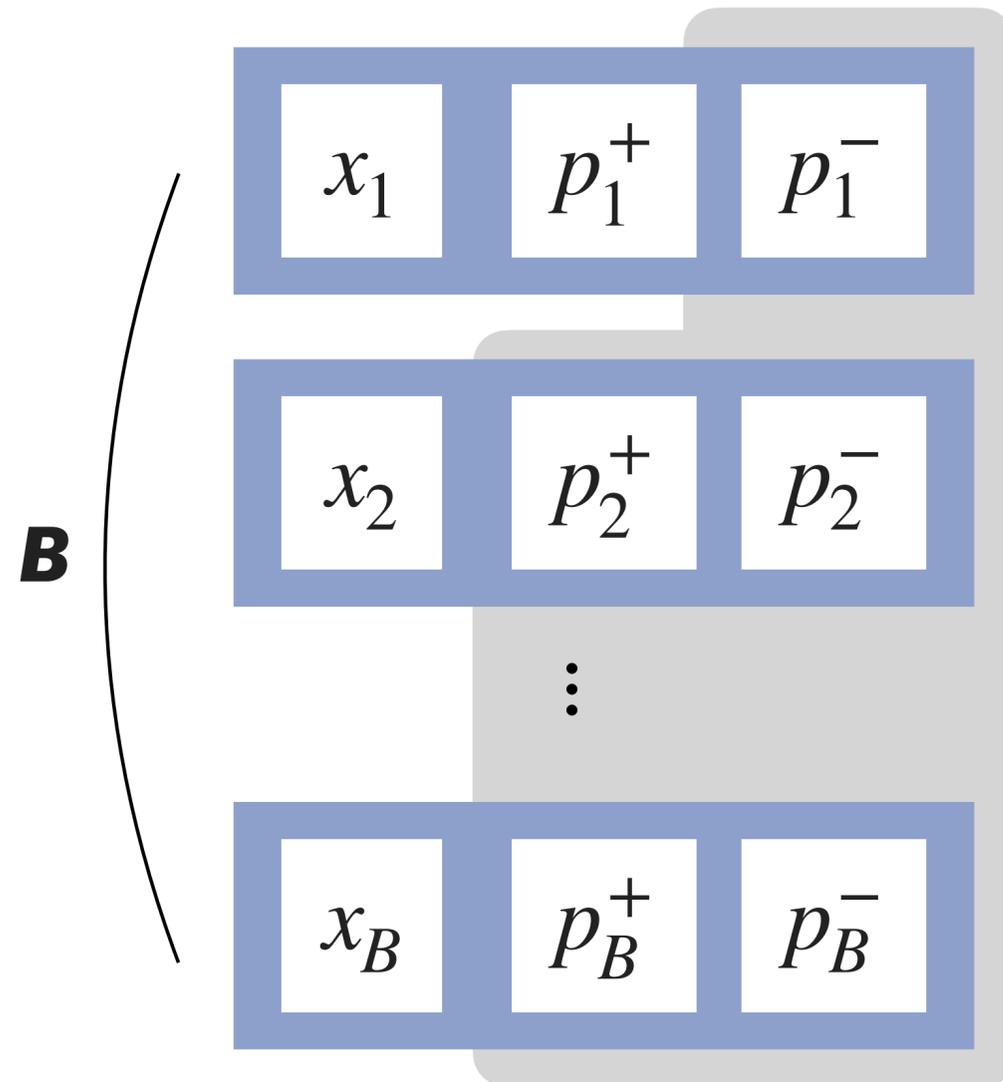
Hard negatives (obtained through BM25)

Draco Malfoy serves as a foil to the hero Harry Potter. Harry first encounters Draco's

Snape is a fictional character ... sarcastic, controlled exterior conceals deep emotions

Sauron is the title character and the main antagonist of The Lord of the Rings, where

Putting it together: DPR objective



Training data: $\langle x, p^+, p^- \rangle$

Batch size: B

$$-\sum_{i=1}^B \log \frac{e^{\text{sim}(x_i, z_i^+)}}{e^{\text{sim}(x_i, z_i^+)} + \sum_{z \in Z_i^-} e^{\text{sim}(x_i, z)}}$$

where $Z_i^- = \{p_i^-\} \cup \bigcup_{j \in [1, B], j \neq i} \{p_j^+, p_j^-\}$

DPR: Retrieval results

Retriever	Top-20					Top-100				
	NQ	TriviaQA	WQ	TREC	SQuAD	NQ	TriviaQA	WQ	TREC	SQuAD
BM25	59.1	66.9	55.0	70.9	68.8	73.7	76.7	71.1	84.1	80.0
DPR	78.4	79.4	73.2	79.8	63.2	85.4	85.0	81.4	89.1	77.2
BM25 + DPR	76.6	79.8	71.0	85.2	71.5	83.8	84.5	80.5	92.7	81.3

Q: BM25 is better than DPR on SQuAD! Why?

DPR: QA results

Training	Model	NQ	TriviaQA	WQ	TREC	SQuAD
Single	BM25+BERT (Lee et al., 2019)	26.5	47.1	17.7	21.3	33.2
Single	ORQA (Lee et al., 2019)	33.3	45.0	36.4	30.1	20.2
Single	HardEM (Min et al., 2019a)	28.1	50.9	-	-	-
Single	GraphRetriever (Min et al., 2019b)	34.5	56.0	36.4	-	-
Single	PathRetriever (Asai et al., 2020)	32.6	-	-	-	56.5
Single	REALM _{Wiki} (Guu et al., 2020)	39.2	-	40.2	46.8	-
Single	REALM _{News} (Guu et al., 2020)	40.4	-	40.7	42.9	-
	BM25	32.6	52.4	29.9	24.9	38.1
Single	DPR	41.5	56.8	34.6	25.9	29.8
	BM25+DPR	39.0	57.0	35.2	28.0	36.7
Multi	DPR	41.5	56.8	42.4	49.4	24.1
	BM25+DPR	38.8	57.9	41.1	50.6	35.8

Improved retrieval translates to improved QA accuracy

Fun fact: Retrieval vs. Parametric-only

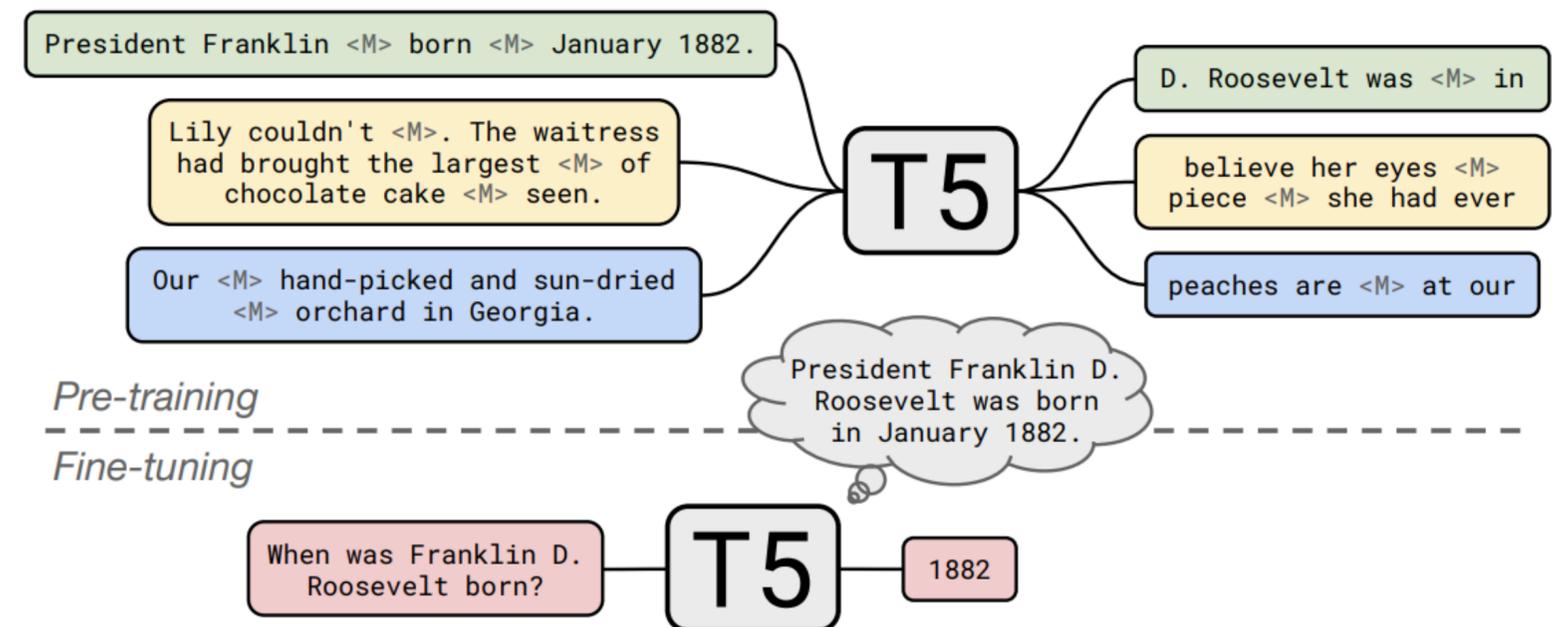
- **–2019:** Retrieval-based models by default (mostly based on lexical-based methods)
- **February 2020:** “Parametric-only LMs” might be better?!
- **April 2020:** Neural retrieval started working (DPR, etc)

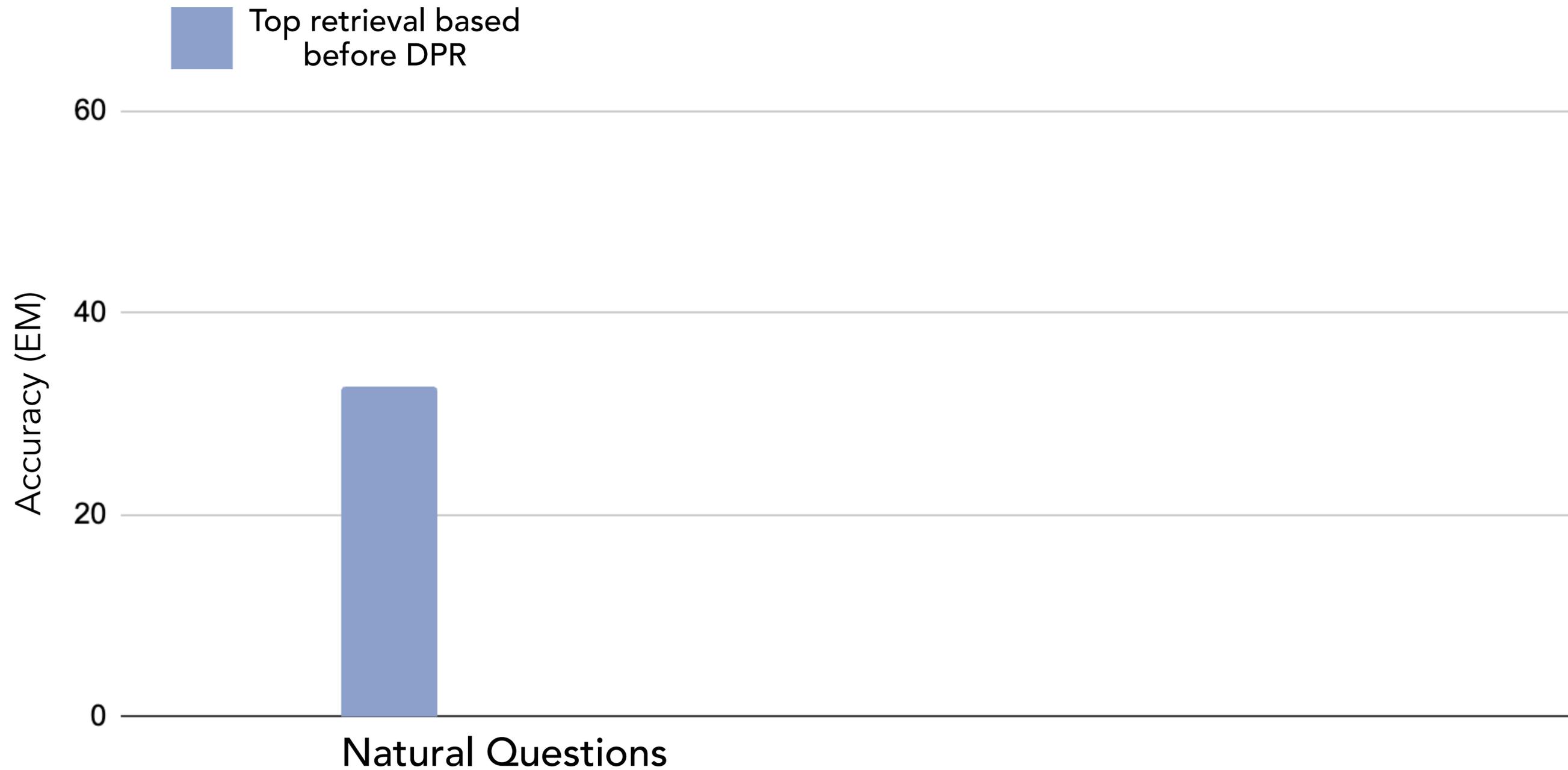
How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Adam Roberts*
Google
adarob@google.com

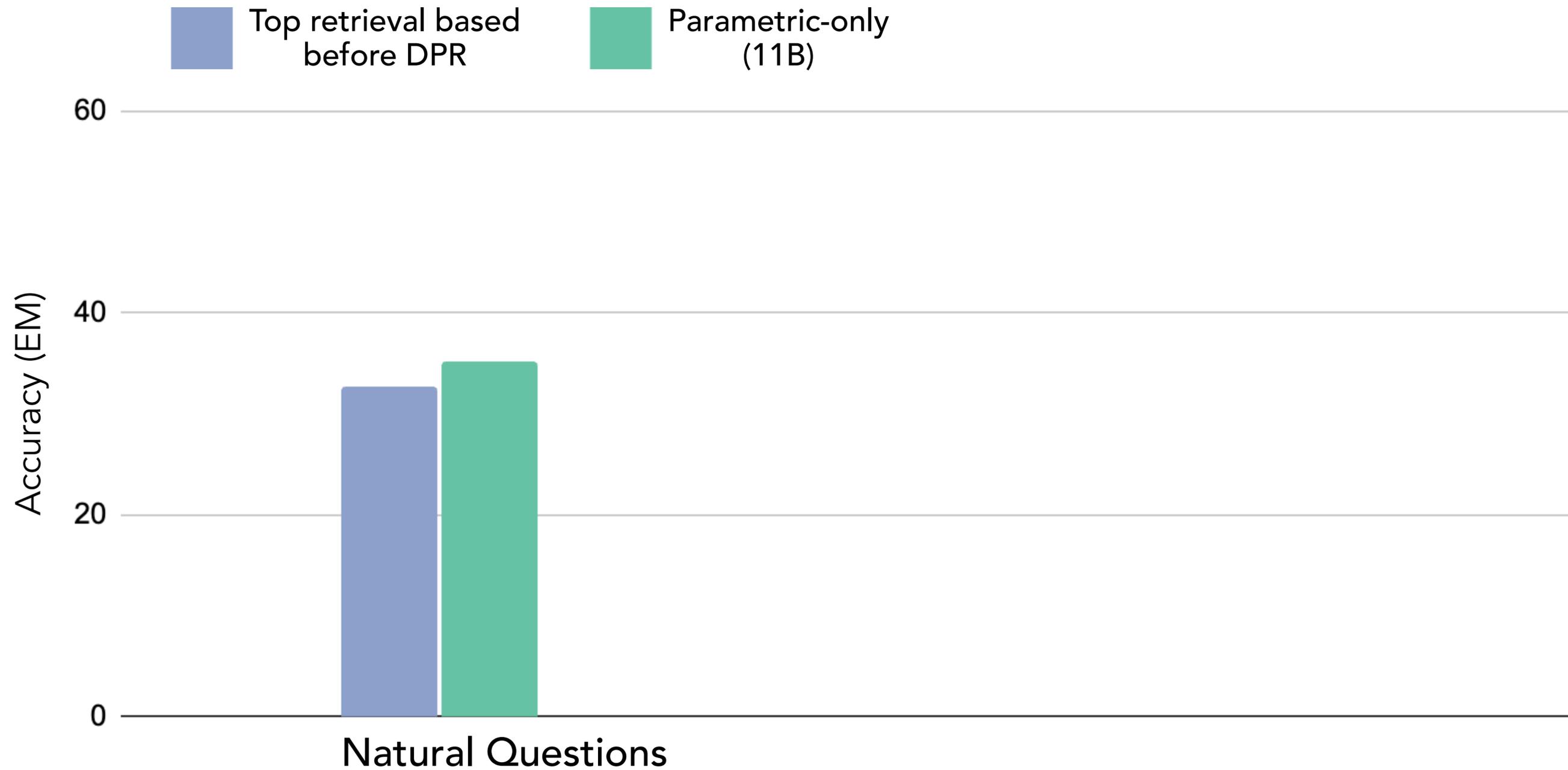
Colin Raffel*
Google
craffel@gmail.com

Noam Shazeer
Google
noam@google.com

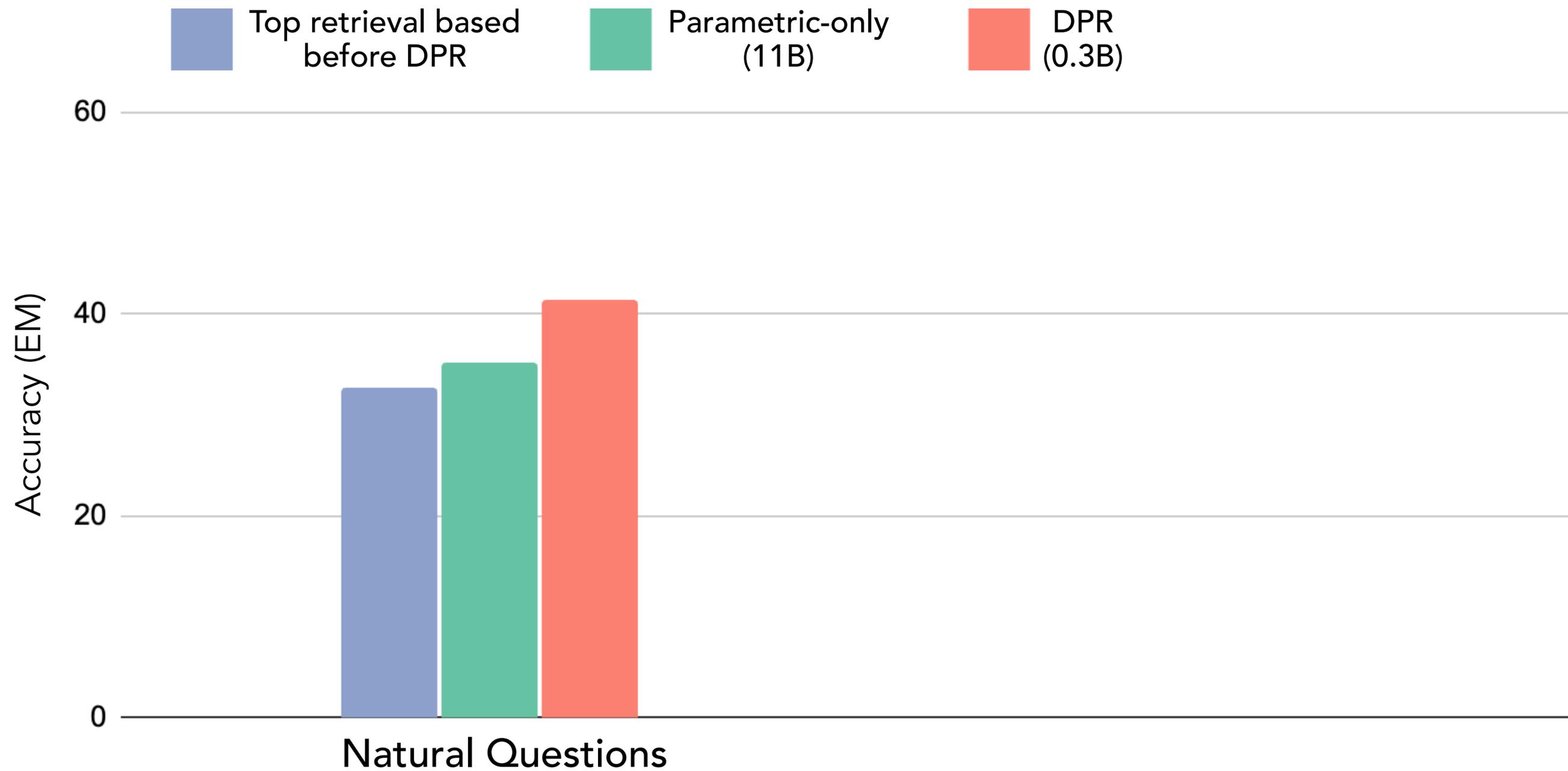




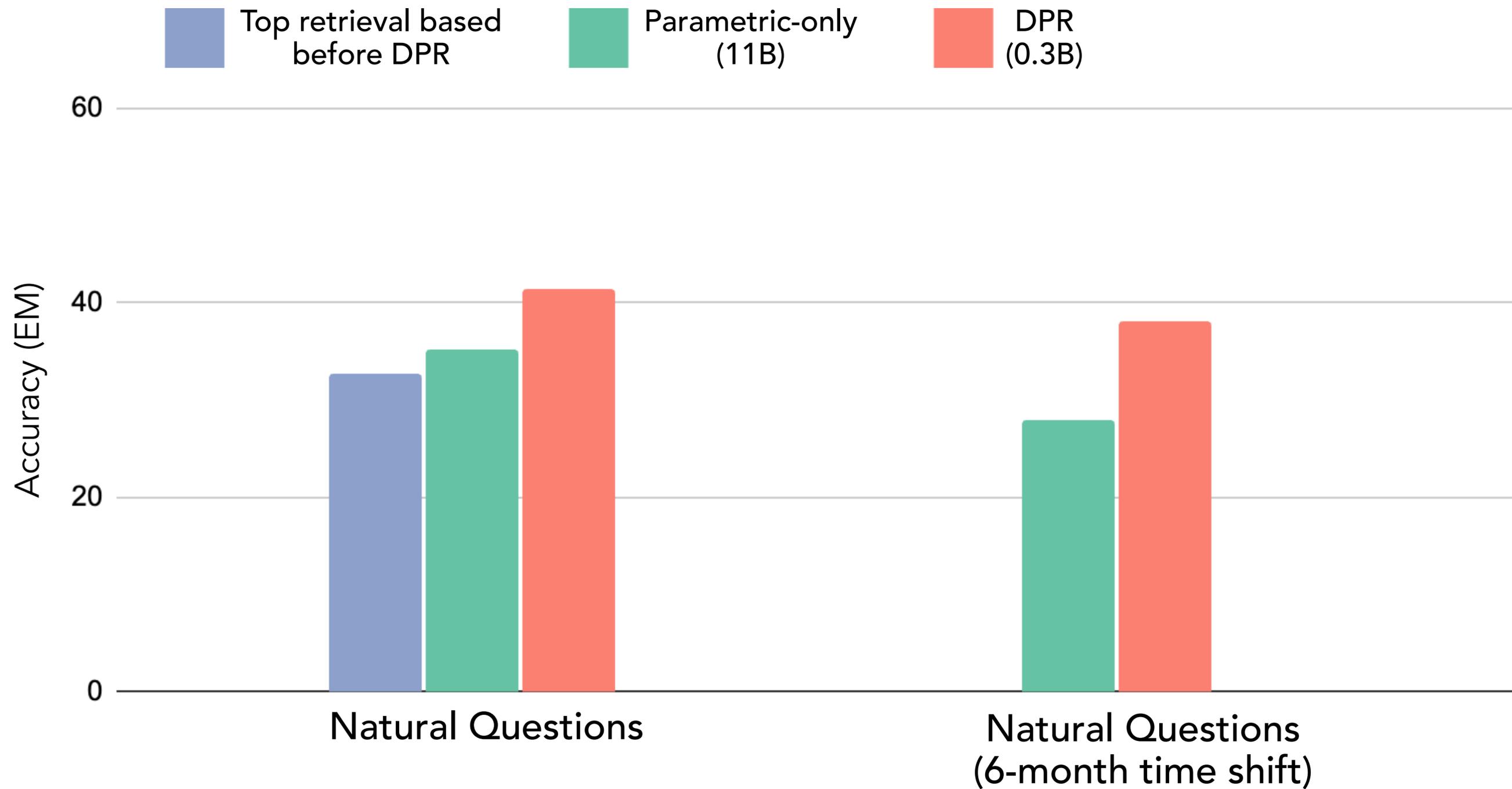
Independent evaluation from NeurIPS 2020 Competition (Min et al. PMLR 2021)



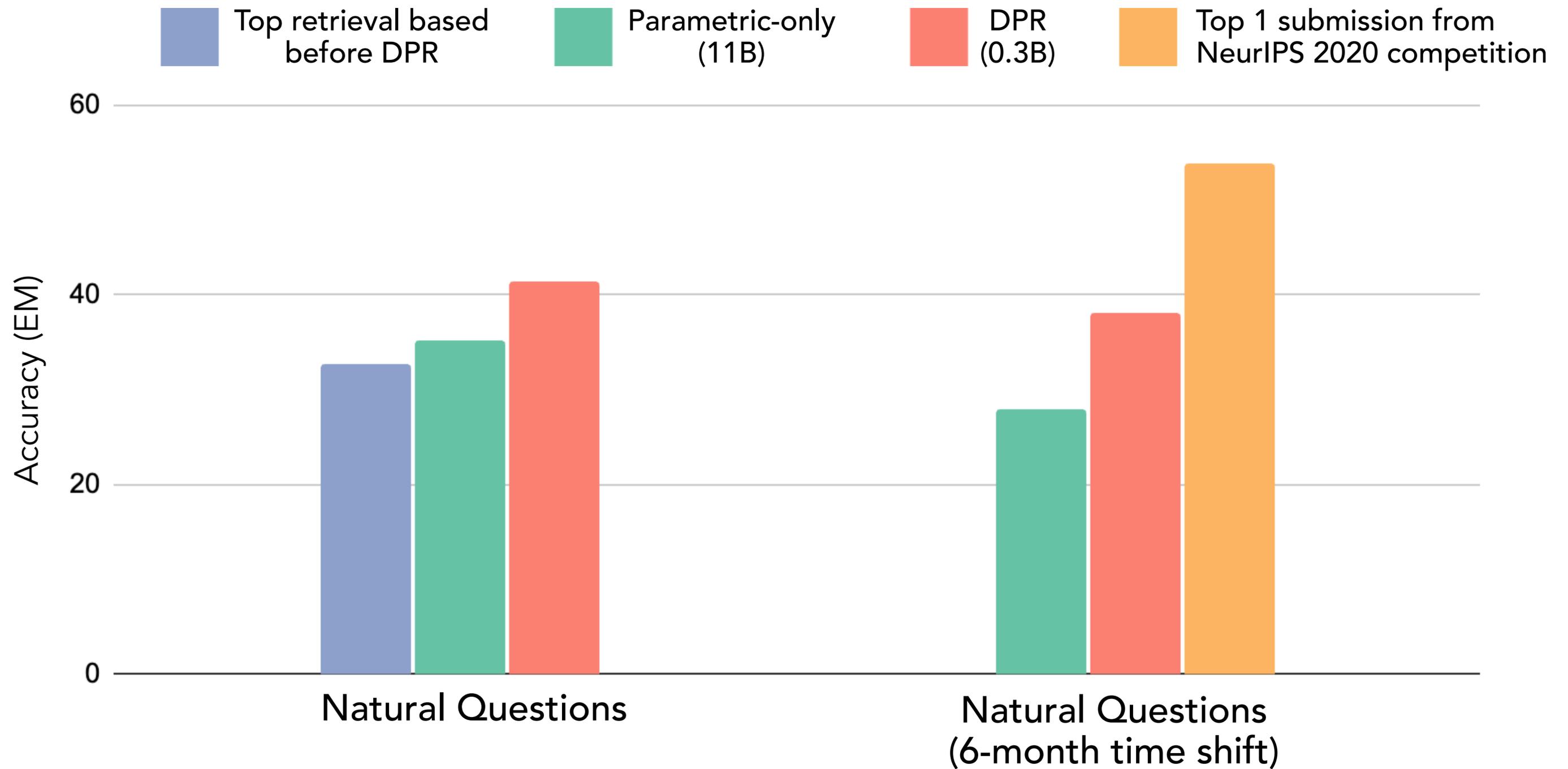
Independent evaluation from NeurIPS 2020 Competition (Min et al. PMLR 2021)



Independent evaluation from NeurIPS 2020 Competition (Min et al. PMLR 2021)

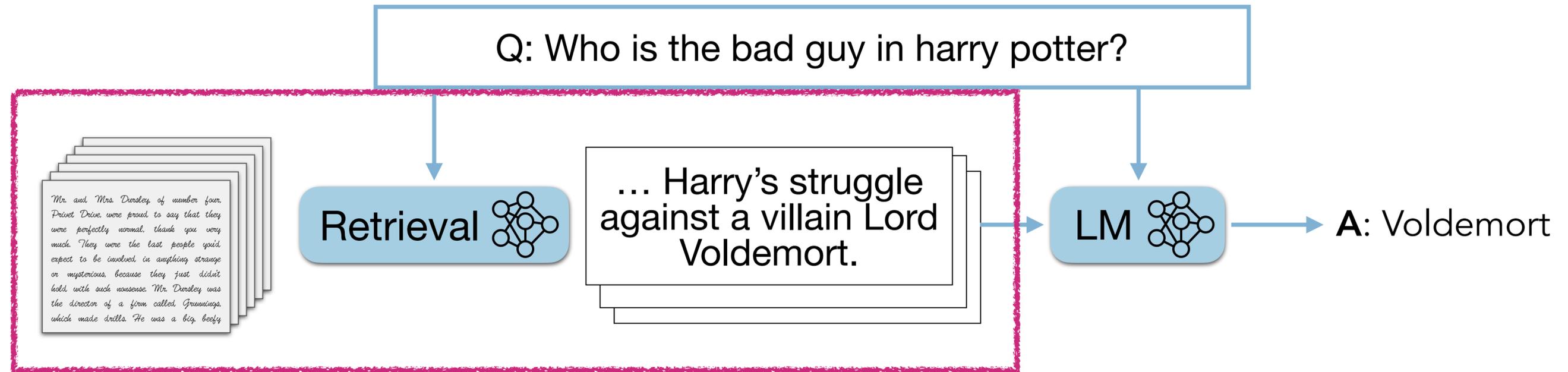


Independent evaluation from NeurIPS 2020 Competition (Min et al. PMLR 2021)

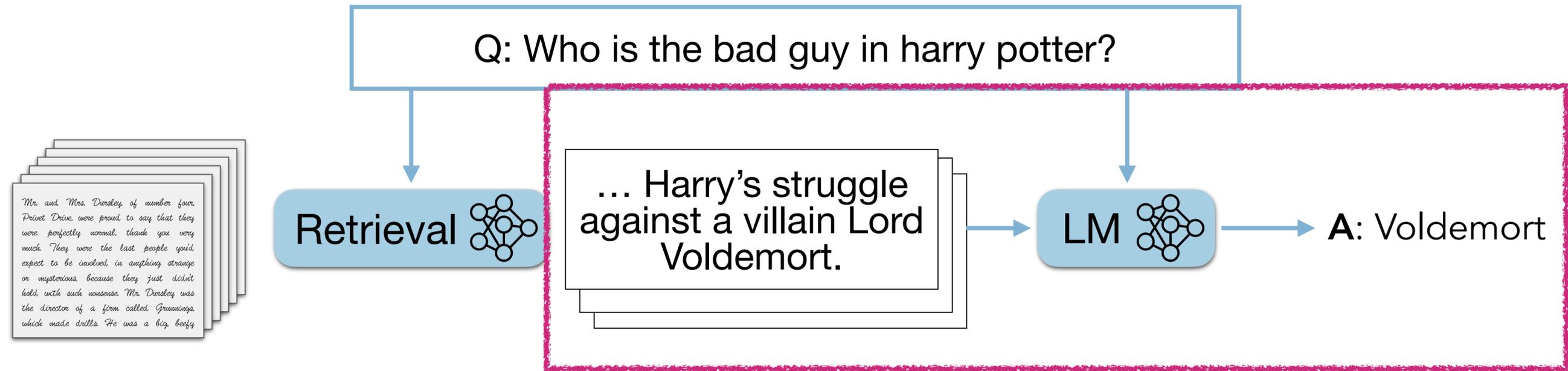


Takeaway: The quality of retrieval-based LMs depends on the quality of **retrieval**

RAG overview



RAG overview



2) Augmentation

- Early work fine-tunes LLMs for retrieval augmentation (DrQA, DPR, RAG)
- Today: Typically, a fixed LM is used, by adding retrieved documents to the input

Augmentation

Retrieval results (ranked)

Voldemort cried,
"Avada Kedavra!" A
jet of green light
issued ...from ...

Voldemort's wand
just as a jet of
red light ...

"The Boy Who
Lived." He saw the
mouth move and a
flash of green ...

Augmentation

Voldemort cried, "Avada Kedavra!" A jet of green light issued from ... just as a jet of red light blasted from Harry's ...

Retrieval results (ranked)

Voldemort cried, "Avada Kedavra!" A jet of green light issued ...from ...



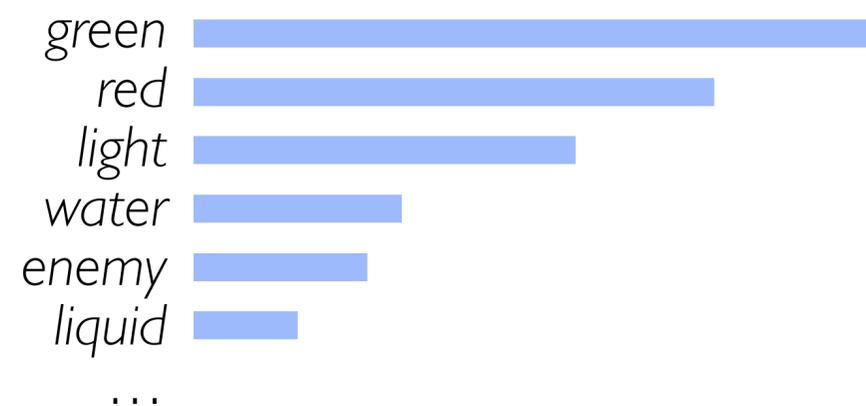
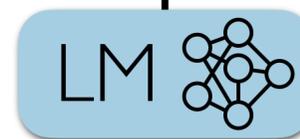
Voldemort's wand just as a jet of red light ...



"The Boy Who Lived." He saw the mouth move and a flash of green ...



+
Question: What color is the Killing Curse in Harry Potter?



Very simple
(You can use a black-box LM like an API!)

Augmentation

Retrieval results (ranked)

Voldemort's want just as a jet of red light ...



Voldemort cried, "Avada Kedavra!" A jet of green light issued ...from ...



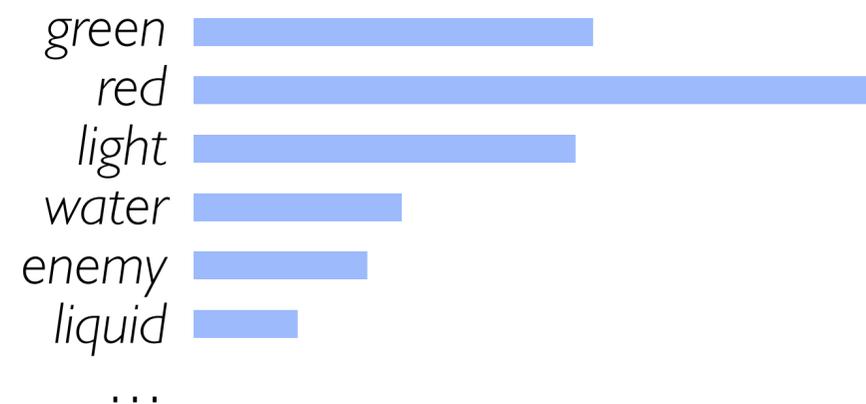
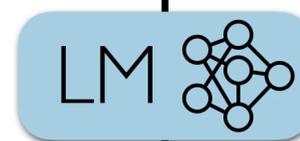
"The Boy Who Lived." He saw the mouth move and a flash of green ...



Voldemort's want just as a jet of red light ...

+

Question: What color is the Killing Curse in Harry Potter?

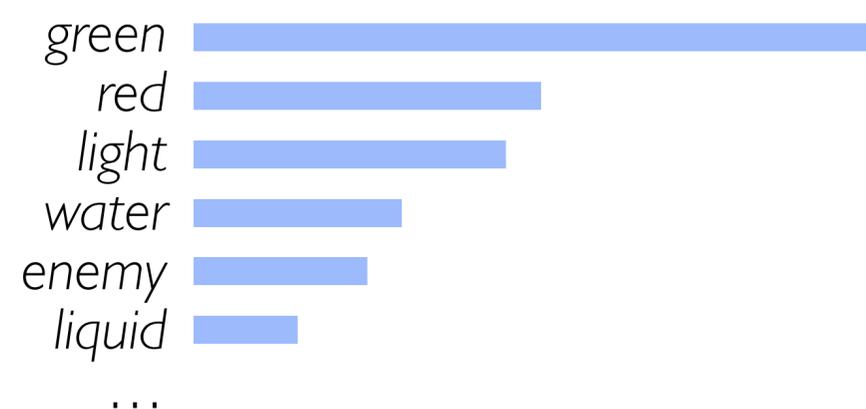
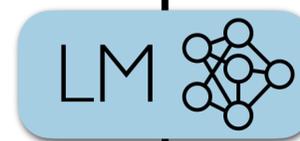


Q: How do we use multiple passages?

Augmentation (I): Concatenation

Voldemort's wand just as a jet of red light ... 
... "Avada Kedavra!" A jet of green light issued ...from ... 
... He saw the mouth move and a flash of green ... 

Question: What color is the Killing Curse in Harry Potter?



 Simple

 Increase the inference cost & Bounded by the maximum length limit of the LM

Augmentation (2): Ensembling

Voldemort's want just as a jet of red light ... 

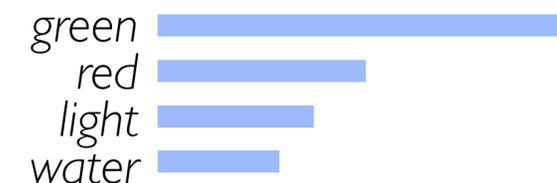
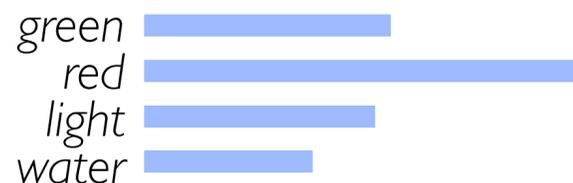
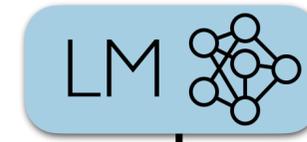
Voldemort cried, "Avada Kedavra!" A jet of green ... 

... saw the mouth move and a flash of green ... 

Question: What color is the Killing Curse in Harry Potter?

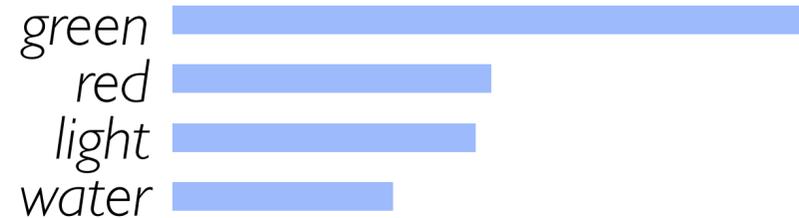
Question: What color is the Killing Curse in Harry Potter?

Question: What color is the Killing Curse in Harry Potter?



$$P(y | x) = \sum_{z \in \mathcal{Z}} P_{\text{ret}}(z | x) P_{\text{LM}}(y | x, z)$$

↑
retrieval score
↑
LM score



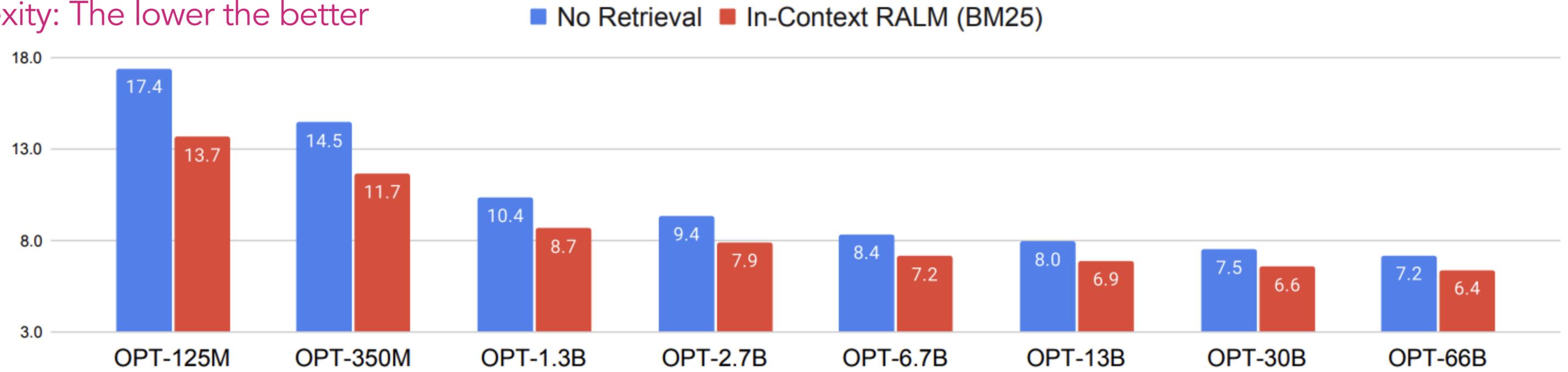
Not bounded by the length limit



Increase the inference cost

Results

Perplexity: The lower the better

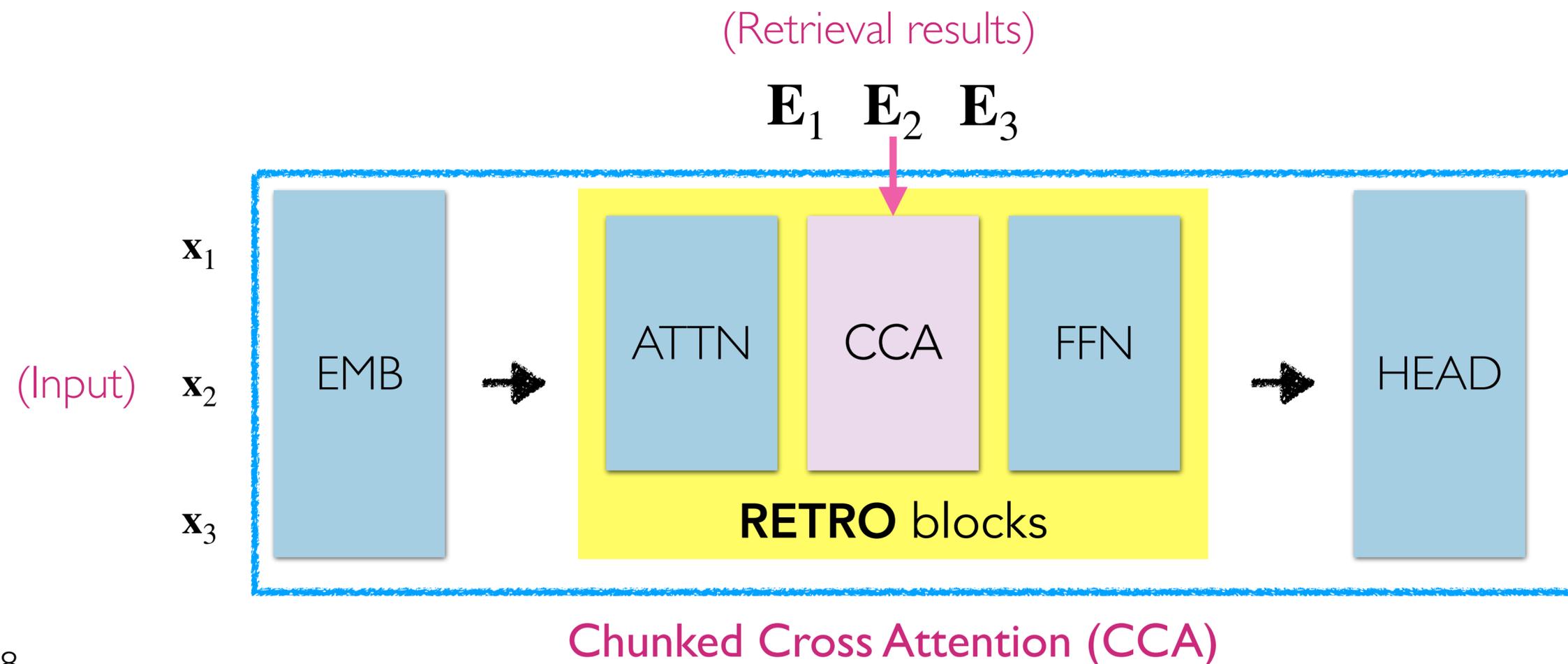


Varying sizes of LMs

Retrieval helps over all sizes of LMs

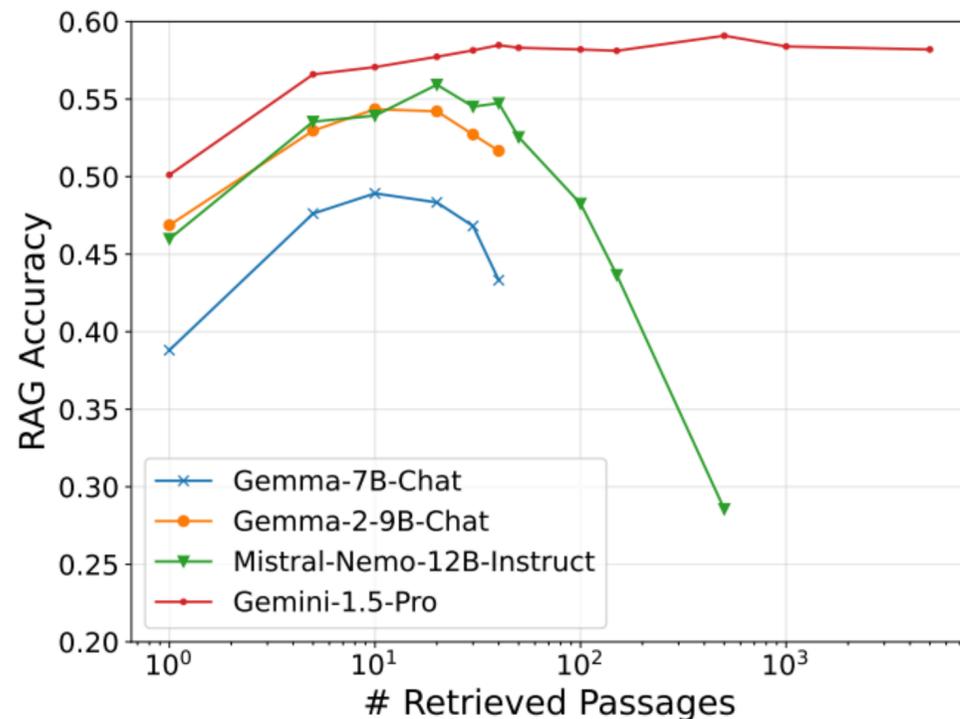
Open challenge: Better augmentation?

- Better, end-to-end training
 - Examples: RAG (Lewis et al. 2020), ALTAS (Izacard et al. 2022), both from Meta
- Better architectures for incorporating retrieved documents
 - Example: RETRO (Borgeaud et al. 2021 from DeepMind)

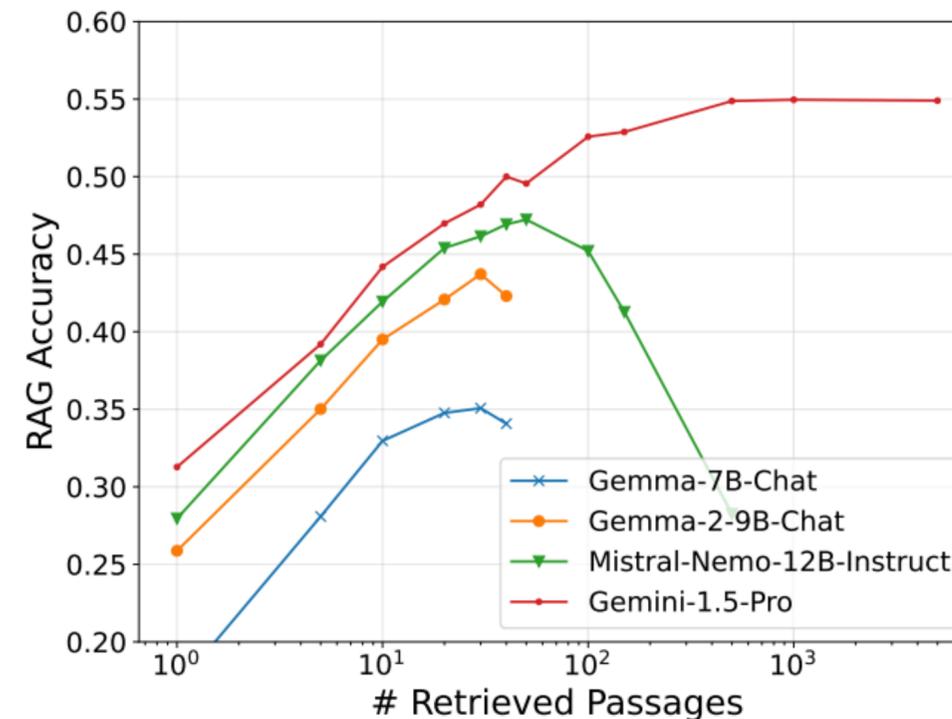


Open challenge: Better augmentation?

- Better, end-to-end training
- Better architectures for incorporating retrieved documents
- Arguably, simply prepending documents as part of an LLM input remained the dominant approach (likely because it works?)
 - But does it always work?

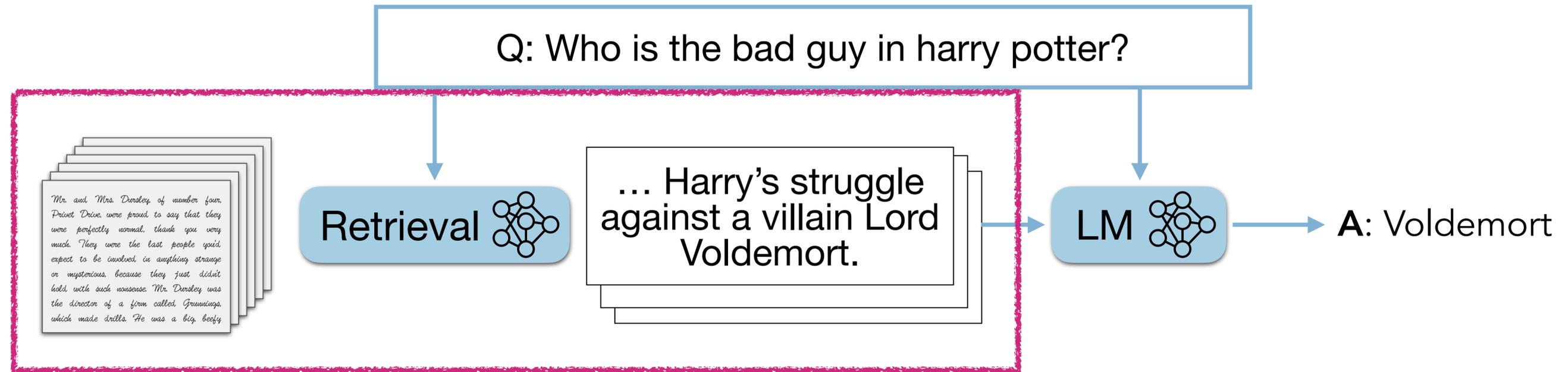


(a) RAG performance with e5 retriever



(b) RAG performance with BM25 retriever

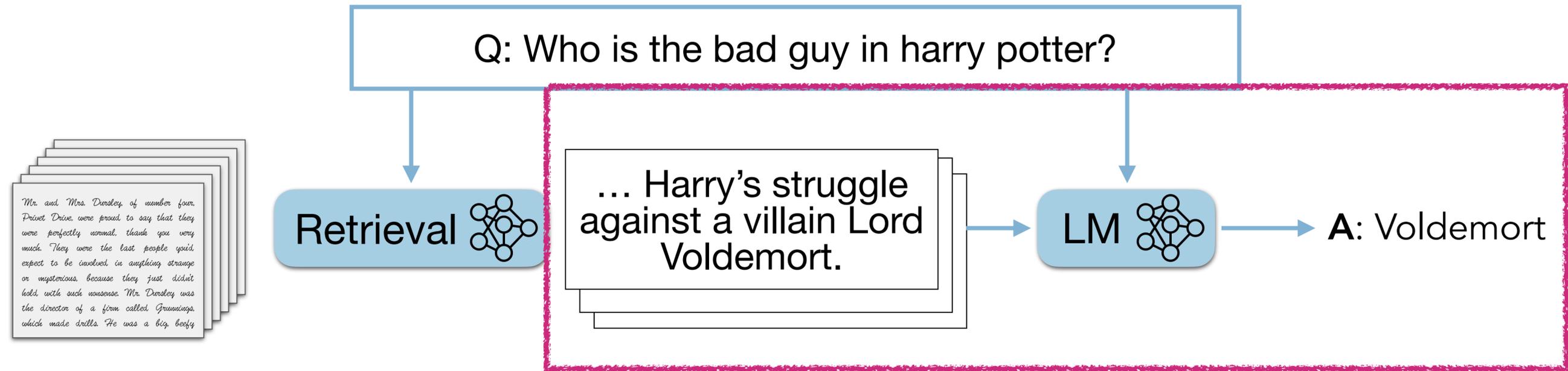
RAG overview



1) Retrieval

Key takeaway: Advances in neural retrieval played a vital role in the success of retrieval-based LMs

RAG overview



2) Augmentation

Key takeaways: Advances in LLMs enabled a very simple recipe: simply prepending retrieved docs into the LLM input

Where Does RAG Stand Today?

Retrieval is now central to production LLMs

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.
- **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](https://www.cntower.ca)
- 2. [travel.usnews.com](https://www.travel.usnews.com)
- 3. [bing.com](https://www.bing.com)
- 4. [rom.on.ca](https://www.rom.on.ca)
- 5. [tripadvisor.com](https://www.tripadvisor.com)

✓ Long-tail knowledge

✓ Staying up-to-date

✓ Citation (for better explainability and verification)

Bing Chat, 2023

Retrieval is now central to production LLMs

ChatGPT ▾

Tell me a bio of Sewon Min.

Sewon Min is a researcher in natural language processing (NLP) and machine learning, particularly known for her work in question answering, retrieval-augmented language models, and large-scale reasoning systems. She completed her Ph.D. in Computer Science at UC Berkeley, advised by Dan Klein. After her Ph.D., she joined Meta AI (FAIR) as a researcher.

incorrect

[Without web browsing]

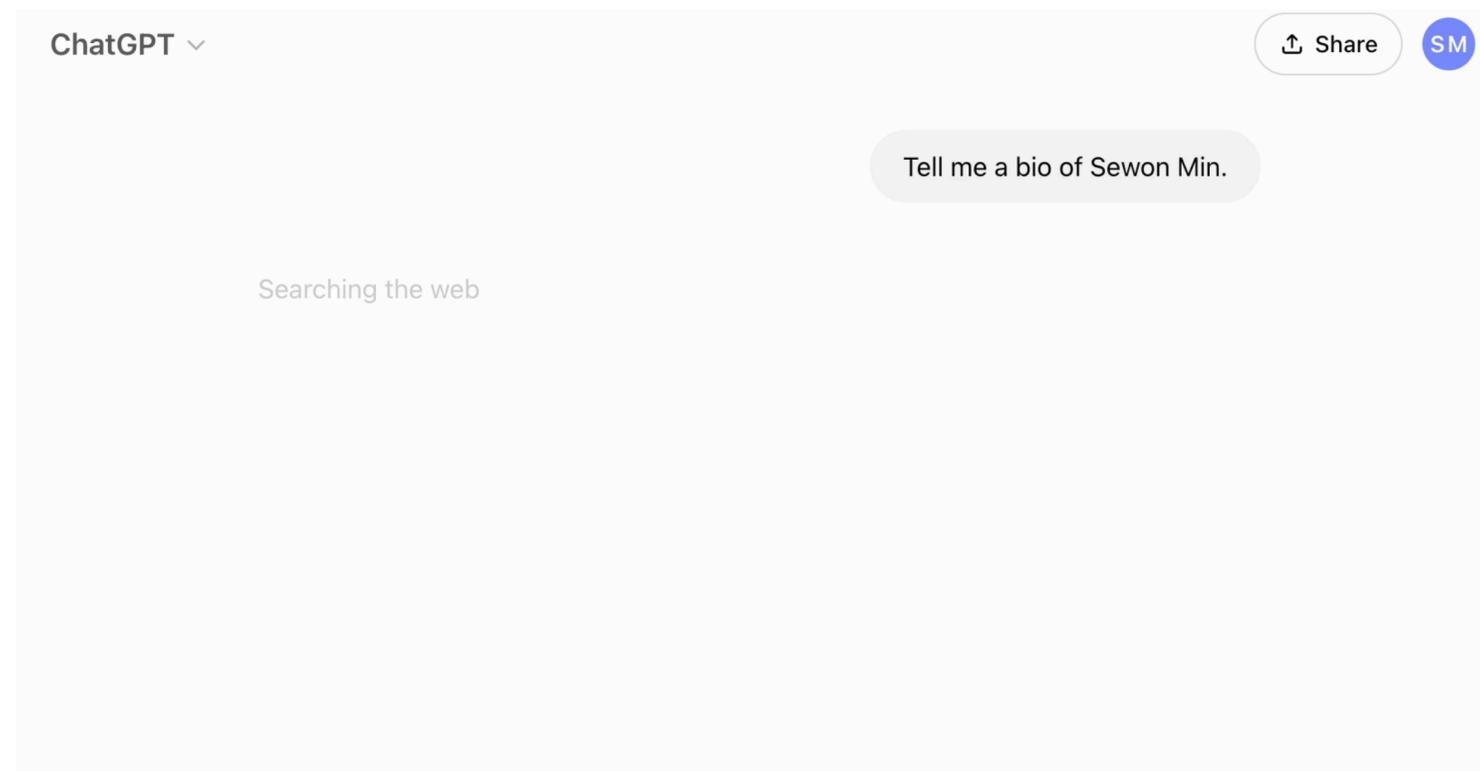
✓ Long-tail knowledge

✓ Staying up-to-date

✓ Citation (for better explainability and verification)

Pro tip: Enable/disable search to compare LLM outputs with vs. without retrieval!
(In ChatGPT: Settings → Personalization → Advanced → Web Search)

Retrieval is now central to production LLMs



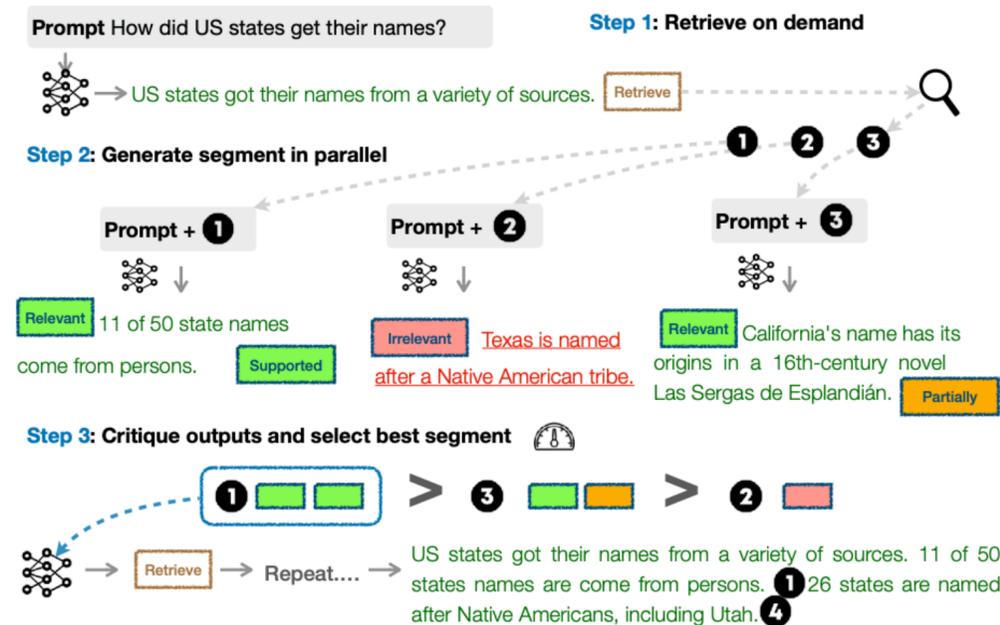
[With web browsing]

- ✓ Long-tail knowledge
- ✓ Staying up-to-date
- ✓ Citation (for better explainability and verification)

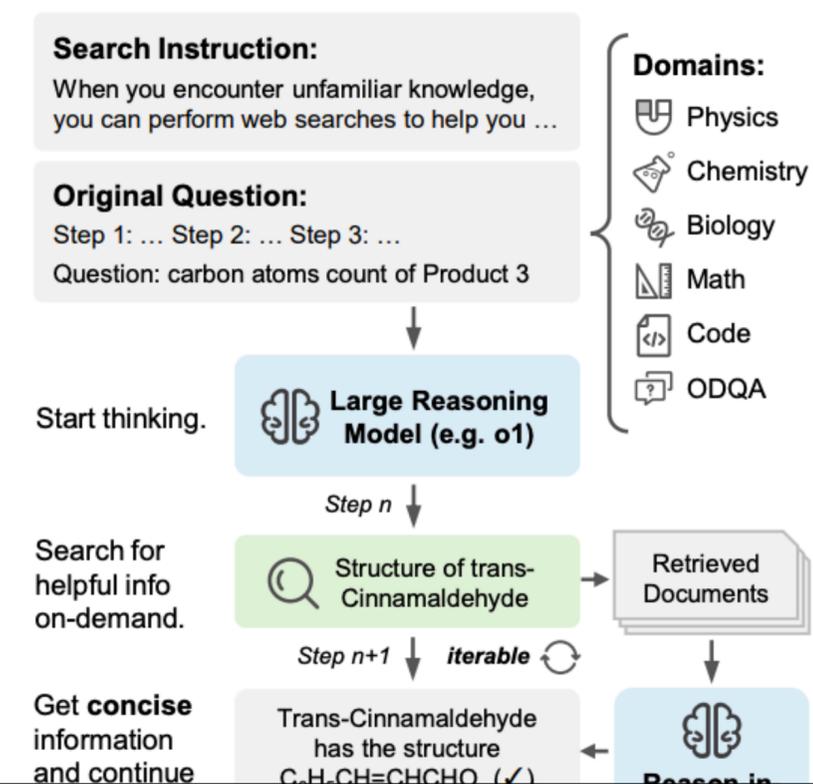
Pro tip: Enable/disable search to compare LLM outputs with vs. without retrieval!
(In ChatGPT: Settings → Personalization → Advanced → Web Search)

RAG / search agents (2024)

Asai et al. 2024. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection"



Li et al. "Search-o1: Agentic Search-Enhanced Large Reasoning Models"



Key ideas: Make it an agent system with functions such as

1. Deciding when to use retrieval
2. Generating a retrieval query
3. Reranking/adaptive adoption: use only relevant retrieved passages
4. Rewriting: make passages more comprehensible & include relevant info only

RAG / search agents (2024)

Methods	General QA				Multi-Hop QA			
	NQ [†]	TriviaQA [*]	PopQA [*]	HotpotQA [†]	2wiki [*]	Musique [*]	Bamboogle [*]	Avg.
Qwen2.5-7b-Base/Instruct								
Direct Inference	0.134	0.408	0.140	0.183	0.250	0.031	0.120	0.181
CoT	0.048	0.185	0.054	0.092	0.111	0.022	0.232	0.106
IRCoT	0.224	0.478	0.301	0.133	0.149	0.072	0.224	0.239
Search-o1	0.151	0.443	0.131	0.187	0.176	0.058	0.296	0.206
RAG	0.349	0.585	0.392	0.299	0.235	0.058	0.208	0.304
SFT	0.318	0.354	0.121	0.217	0.259	0.066	0.112	0.207
R1-base	0.297	0.539	0.202	0.242	0.273	0.083	0.296	0.276
R1-instruct	0.270	0.537	0.199	0.237	0.292	0.072	0.293	0.271
Rejection Sampling	0.360	0.592	0.380	0.331	0.296	0.123	0.355	0.348
Search-R1-base	0.480	0.638	0.457	0.433	0.382	0.196	0.432	0.431
Search-R1-instruct	0.393	0.610	0.397	0.370	0.414	0.146	0.368	0.385

Jin et al. "Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning"

Impressive performance, but on short-form QA datasets

Deep Research (DR) agents (2025-)



User

How did Netflix manage to successfully adapt One Hundred Years of Solitude, a notoriously difficult book to bring to the screen?

2026 standards: **Open-ended** questions requiring **long-form, complex, interleaved retrieval-reasoning** workflows, often in domains such as scientific literature.

- Production: ChatGPT DR, Gemini DR, many others
- Open-source DR agents: Tongyi DeepResearch (Alibaba), Dr. Tulu (Ai2), many others



DR Tulu

Agentic Workflow

☰ Think ☒ Tool1 ☰ Think ☒ Tool2 ... ✍ Answer

Long-form Report with Citations

Netflix's adaptation **avoided a literal replica of Macondo and instead fused real locations with meticulously built sets**^[1] to honor the novel's essence while giving the show contemporary visual grammar. **The production grounded magical realism in front-of-camera practical craft, relying on makeup, special effects,**^[2]
The location strategy and production design...

Sources

- [1] The production team behind Netflix's adaptation of "One Hundred Years of Solitude," LA Times
- [2] The article discusses Netflix's adaptation of Gabriel García Márquez's celebrated novel, NY Times

Questions?

Acknowledgement

Stanford CS224N NLP with Deep Learning by Diyi Yang and Tatsunori Hashimoto
(and older versions by Chris Manning)