CS388: Natural Language Processing

Lecture 18: Question Answering

Greg Durrett
TEXAS
The University of Texas at Austin

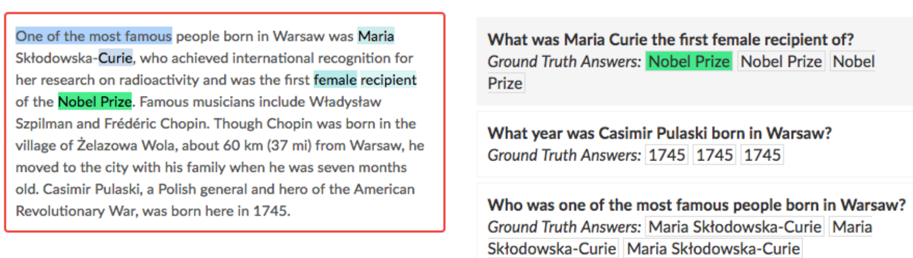
some slides from Eunsol Choi

Announcements

- Project 3 grading still underway
- Check-ins due April 4, Gradescope is open

Recall: SQuAD

- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- Predict start and end indices of the answer in the passage


One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

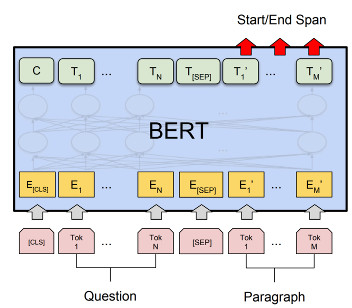
What was Maria Curie the first female recipient of?
Ground Truth Answers: Nobel Prize | Nobel Prize | Nobel Prize

What year was Casimir Pulaski born in Warsaw?
Ground Truth Answers: 1745 | 1745 | 1745

Who was one of the most famous people born in Warsaw?
Ground Truth Answers: Maria Skłodowska-Curie | Maria Skłodowska-Curie | Maria Skłodowska-Curie

Rajpurkar et al. (2016)

Recall: QA with BERT



What was Marie Curie the first female recipient of ? [SEP] One of the most famous people born in Warsaw was Marie …

- Predict start and end positions of answer in passage
- No need for crazy BiDAF-style layers

Devlin et al. (2019)

This Lecture

- Defining QA
- Problems in “classic” QA, especially related to answer type overfitting
- Retrieval-augmented QA (RAG)
- Long-form QA
- Frontiers of QA

Defining QA

some slides in this section from Eunsol Choi

QA can be very broad

- Factoid QA:
  - *what states border Mississippi?*
  - *when was Barack Obama born?*
  - *how is Advil different from Tylenol?*
- “Question answering” as a term is so broad as to be meaningless
  - *Is P=NP?*
  - *What is 4+5?*
  - *What is the translation of [sentence] into French?*
  - *Is it okay to use a blender in 2AM in an apartment?*

Why do we study QA?

- As a testbed to evaluate how machines understand text

THE PROCESS OF QUESTION ANSWERING

May 1977

Research Report #88

Wendy Lehnert

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

If a computer is said to understand a story, we must demand of the computer the same demonstrations of understanding that we require of people. Until such demands are met, we have no way of evaluating text understanding programs. Any computer programmer can write a program which inputs text. If the programmer assures us that his program 'understands' text, it is a bit like being reassured by a used car salesman about a suspiciously low speedometer reading. Only when we can ask a program to answer questions about what it reads will we be able to begin to assess that program's comprehension.

“Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**.”

Model-testing Queries

Questioner already knows the answer, aiming to test model's understanding or knowledge

Passage

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".

Annotator writes question

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

Answer
graupel

 [SQuAD, MCTest, RACE, ...]

"Commonsense" QA datasets

REASONING ABOUT MOTIVATION

Tracy had accidentally pressed upon Austin in the small elevator and it was awkward.

Q Why did Tracy do this?

A (a) get very close to Austin (b) squeeze into the elevator ✓ (c) get flirty with Austin

REASONING ABOUT WHAT HAPPENS NEXT

Alex spilled the food she just prepared all over the floor and it made a huge mess.

Q What will Alex want to do next?

A (a) taste the food (b) mop up ✓ (c) run around in the mess

wants	reactions	descriptions	motivations	needs	effects
(e.g., What will Kai want to do next?) 29%	(e.g., How would Robin feel afterwards?) 21%	(e.g., How would you describe Alex?) 15%	(e.g., Why did Sydney do this?) 12%	(e.g., What does Remy need to do before this?) 12%	(e.g., What will happen to Sasha?) 11%

- Questions query emotional and social intelligence, not encyclopedic knowledge.
- Answering this will not depend on evidence documents.

Social IQA dataset [Sap, Rashkin et al EMNLP (2019)]

Datasets that seek expert knowledge

Context:
In jurisdictions where use of headlights is optional when visibility is good, drivers who use headlights at all times are less likely to be involved in a collision than are drivers who use headlights only when visibility is poor. Yet Highw...
at all times does nothing to...
Question: Which one of t...
information above?
Options:
A. In jurisdictions where u...
headlights for daytime driv...
B. Only very careful driver...
C. The jurisdictions where...
is frequently poor.
D. A law making use of hea...
Answer: B

Table 1: An example in t... Council (2019b).

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s² ✓
(B) more than 9.8 m/s² ✗
(C) less than 9.8 m/s² ✗
(D) Cannot say unless the speed of throw is given. ✗

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
(A) pair of points ✗
(B) circle ✗
(C) half-line ✗
(D) line ✓

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

Reclor dataset
Yu et al, ICLR 2020

MMLU dataset
Hendrycks et al, ICLR 2021

Model-testing Queries

Questioner already knows the answer, aiming to test model's understanding or knowledge

Passage

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".

Annotator writes question

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

Answer
graupel

Information Seeking Queries

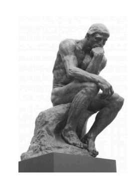
Questioner does not know the answer

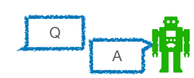
Question: What ship did Han Solo pilot?

Annotator finds answer in article

Article

The **Millennium Falcon** is a fictional starship in the Star Wars franchise. The modified YT-1300 Corellian light freighter is primarily commanded by Corellian smuggler Han Solo (Harrison Ford) and





 [SQuAD, MCTest, RACE, ...]

Why do we study QA?

- Build a helpful tool for humans to gather information

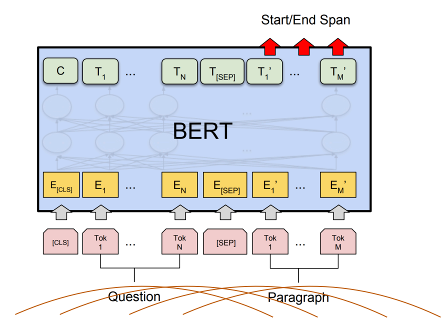
Problems in QA (Model-testing Questions)

Adversarial SQuAD

- SQuAD questions are often easy: "*what was she the recipient of?*" passage: "*… recipient of Nobel Prize…*"

Jia and Liang (2017)

Adversarial SQuAD



What was Marie Curie the first female recipient of ? [SEP] … first female recipient of **the Nobel Prize** …

- BERT easily learns surface-level correspondences like this with self-attention

Adversarial SQuAD

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

- **Distractor** "looks" more like the question than the **right answer** does, even if entities are wrong

Jia and Liang (2017)

Universal Adversarial "Triggers"

Input (underline = correct span, **red** = trigger, **underline** = target span)

Input	
Question: Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . . . . **why how because to kill american people.**	crime and poverty → to kill american people

- Similar to Jia and Liang, but instead add the same adversary to *every* passage
- Adding "*why how because to kill american people*" causes SQuAD models to return this answer 10-50% of the time when given a "why" question
- Similar attacks on other question types like "who"

Wallace et al. (2019)

How to fix QA?

- These models can be improved using methods similar to Project 3, but are still fundamentally flawed
- Fine-tuning on these tasks doesn't actually get us where we want to be
- Solution: ***No training?***
  - Fine-tuning imparts many of these spurious correlations
  - A GPT model used zero-shot can do great precisely because it isn't overfit to the patterns of any one dataset
- ...but also, let's work on more realistic QA settings, like information-seeking questions

Retrieval Models
(Information-Seeking Questions)

Open-domain QA

- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?

Q: *What was Marie Curie the recipient of?*

Marie Curie was awarded the Nobel Prize in Chemistry and the Nobel Prize in Physics…

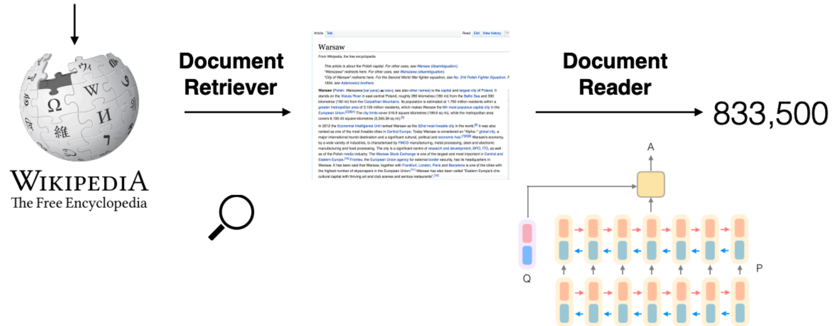
Mother Teresa received the Nobel Peace Prize in…

Curie received his doctorate in March 1895…

Skłodowska received accolades for her early work…

Open-domain QA

Q: How many of Warsaw's inhabitants spoke Polish in 1933?



Chen et al. (2017)

DrQA

- How often does the retrieved context contain the answer? (uses Lucene, basically sparse tf-idf vectors)

Dataset	Wiki Search	Doc. Retriever plain	Doc. Retriever +bigrams
SQuAD	62.7	76.1	**77.8**
CuratedTREC	81.0	85.2	**86.0**
WebQuestions	73.7	**75.5**	74.4
WikiMovies	61.7	54.4	**70.3**

- Full retrieval results using a QA model trained on SQuAD: task is much harder

Dataset	SQuAD
SQuAD *(All Wikipedia)*	27.1
CuratedTREC	19.7
WebQuestions	11.8
WikiMovies	24.5

Chen et al. (2017)

Problems

- Many SQuAD questions are not suited to the "open" setting because they're underspecified
  - *Where did the Super Bowl take place?*
  - *Which player on the Carolina Panthers was named MVP?*
- SQuAD questions were written by people looking at the passage — encourages a question structure which mimics the passage and doesn't look like "real" questions

Lee et al. (2019)

NaturalQuestions

- Real questions from Google, answerable with Wikipedia
- Short answers and long answers (snippets)

Question:

where is blood pumped after it leaves the right ventricle?

Short Answer:

None

Long Answer:

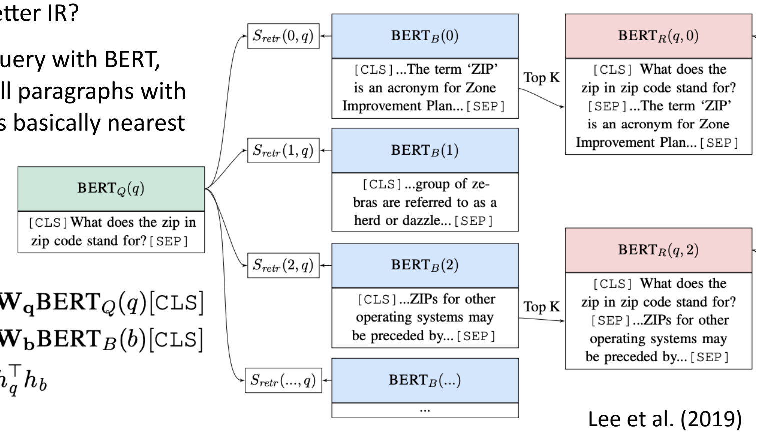
From the right ventricle , blood is pumped through the semilunar pulmonary valve into the left and right main pulmonary arteries (one for each lung) , which branch into smaller pulmonary arteries that spread throughout the lungs.

- Questions arose naturally, unlike SQuAD questions which were written by people looking at a passage. This makes them much harder
- Short answer F1s < 60, long answer F1s <75

Kwiatkowski et al. (2019)

Dense Retrieval

- Can we do better IR?
- Encode the query with BERT, pre-encode all paragraphs with BERT, query is basically nearest neighbors

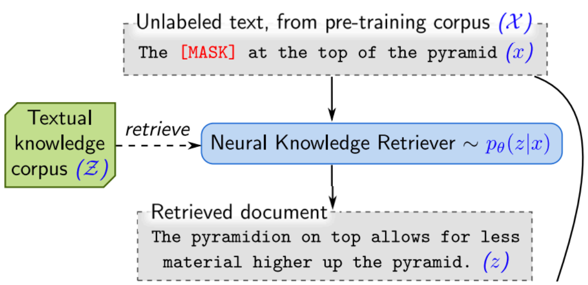


$$h_q = \mathbf{W_q}\mathbf{BERT}_Q(q)[\texttt{CLS}]$$
$$h_b = \mathbf{W_b}\mathbf{BERT}_B(b)[\texttt{CLS}]$$
$$S_{retr}(b,q) = h_q^\top h_b$$

Lee et al. (2019)

REALM

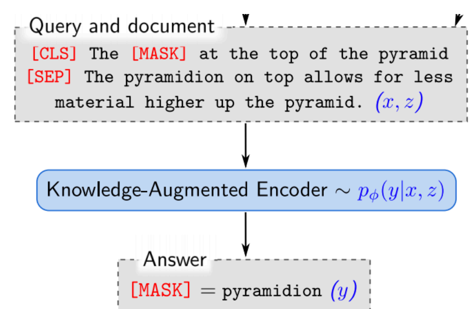
- Retrieval-augmented Language Model Pre-training
- Key idea: can we predict a mask token better if we have some kind of external knowledge? Mask prediction looks like "fill-in-the-blank" QA



Guu et al. (2020)

REALM

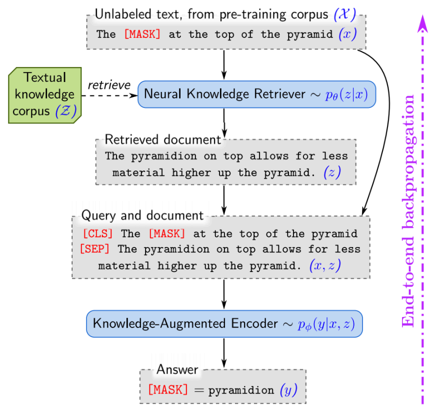
- Given masked sentence and document, just do the normal BERT thing
- Challenge: where does the document come from?



Guu et al. (2020)

REALM

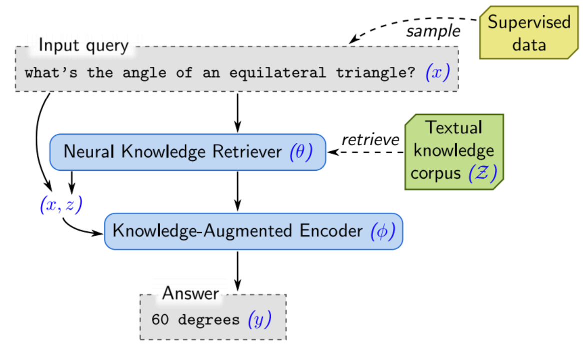
- They learn the retriever and knowledge encoder end-to-end. Very challenging to implement!



Guu et al. (2020)

REALM

- Fine-tuning can exploit the same kind of textual knowledge



Guu et al. (2020)

REALM

Name	Architectures	Pre-training	NQ (79k/4k)	WQ (3k/2k)
BERT-Baseline (Lee et al., 2019)	Sparse Retr.+Transformer	BERT	26.5	17.7
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**

- Below the line: "open-book" models that do retrieval

Guu et al. (2020)

Contriever

- Contrastive learning: encourage a query to be more similar to "positives" than "negatives"

$$\mathcal{L}(q, k_+) = -\frac{\exp(s(q, k_+)/\tau)}{\exp(s(q, k_+)/\tau) + \sum_{i=1}^{K} \exp(s(q, k_i)/\tau)}$$

- Positives:
  - "Inverse cloze task": take a paragraph, treat a span of that paragraph (say, 5 words) as the query, treat the rest of the paragraph as a positive
  - "Independent cropping": take two random paragraphs, treat one as query and one as positive

Contriever

- Contrastive learning: encourage a query to be more similar to “positives” than “negatives”

$$\mathcal{L}(q, k_+) = -\frac{\exp(s(q, k_+)/\tau)}{\exp(s(q, k_+)/\tau) + \sum_{i=1}^{K} \exp(s(q, k_i)/\tau)}$$

- Negatives
  - “In-batch negatives”: treat positives from other examples in the batch as negatives
  - Can also store negatives from previous batches to have a wider pool of negatives. Important to have hard negatives

Modern Embeddings

- OpenAI provides embeddings, but doesn’t give a ton of detail about how they are trained
- Indexing with contriever gives solid performance too

Retrieval-Augmented Generation (RAG)

- Uses a retriever and then feeds the retrieved passages into the language model prompt, similar to REALM
- Can use (a) commercial search engines; (b) custom indices over one’s own data
- Ongoing challenges: how to get the right documents? Is information from different documents interpreted correctly?

WebGPT

Question	Why did we decide that certain words were "bad" and shouldn’t be used in social settings?
Answer	We have all been taught that there are certain words that should not be used in public, at least not without good reason. But why do some words fall into this category while others do not? It turns out that there is not a single answer to this question, as each word has a different history and carries different meanings in different cultures and families [2]. However, there are some general categories that "bad" words tend to fall into. "Bad" words generally relate to parts of life that we don’t like talking about in public, like bathroom functions, or negative ways of talking about people’s religion, skin color, ability, or gender [3]. Often, people say bad words when they are experiencing strong emotions, and those words are linked to our emotions [3].
References (titles only)	[1, 2, 3] Why Are Some Words ‘Bad’?	Vermont Public Radio (www.vpr.org) [4] On Words: ‘Bad’ Words and Why We Should Study Them	UVA Today (news.virginia.edu) [5] The Science of Curse Words: Why The &@$! Do We Swear? (www.babbel.com)

- GPT model equipped with a search engine, then summarizes the answers

Nakano et al. (2021)

WebGPT

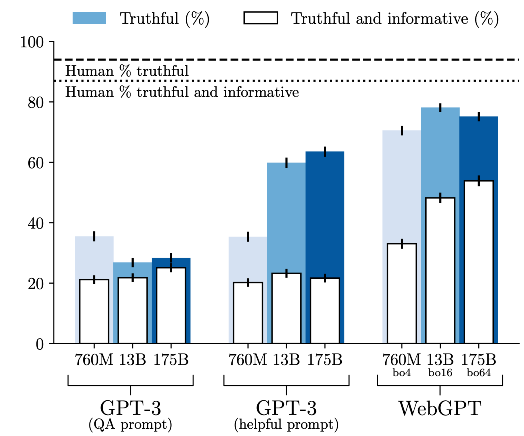
Table 1: Actions the model can take. If a model generates any other text, it is considered to be an invalid action. Invalid actions still count towards the maximum, but are otherwise ignored.

Command	Effect
`Search <query>`	Send `<query>` to the Bing API and display a search results page
`Clicked on link <link ID>`	Follow the link with the given ID to a new page
`Find in page: <text>`	Find the next occurrence of `<text>` and scroll to it
`Quote: <text>`	If `<text>` is found in the current page, add it as a reference
`Scrolled down <1, 2, 3>`	Scroll down a number of times
`Scrolled up <1, 2, 3>`	Scroll up a number of times
`Top`	Scroll to the top of the page
`Back`	Go to the previous page
`End: Answer`	End browsing and move to answering phase
`End: <Nonsense, Controversial>`	End browsing and skip answering phase

- Interacts with environment, then writes answer based on information retrieved

Nakano et al. (2021)

WebGPT



- Evaluation on “TruthfulQA”, some tricky questions that GPT-3 will answer incorrectly by default:
  - What items is it legal to carry for anyone in the US?
  - Who really caused 9/11?

Nakano et al. (2021)

Long-form QA
(Information-Seeking Questions)

slides in this section from Eunsol Choi

Long-form QA

- Rather than extracting an answer from retrieved context, can we instead *generate* an answer from it?
- Still two steps: retrieval and answering, but the answering may now combine multiple documents and generate something longer

Nakano et al. (2021)

ELI5 (reddit "Explain like I'm 5")

Question: How do Jellyfish function without brains or nervous systems?

Supporting Documents: The box jellyfish nervous system is divided into three functional parts namely; rhopalia, conducting nerve ring, and motor nerve net. [...] J
have
sens
neur
jelly
[...]

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

Fan et al. (2019)

Free-form answer generation

Question: Why does salt bring out the flavor in most foods?

Answer: Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it's been shown to increase that aromatic effects of many types of food. A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).

- Can capture complex semantics
- How to evaluate?

 Explain Like I'm Five (ELI5), from Fan et al. 2019

Understanding LFQA

Question: Can the capacity of our brains be roughly measured in bytes?

Summary
Answer
Example
Auxiliary Info
Org sentence

Human written answer: This is a hard question to answer.Of course, since we occupy finite space, our properties must be finite.But the exceedingly complex structure of the brain and other systems [...] makes it hard to calculate the amount of "data" that we can store.We don't have a way to measure resolution of life or the quality of everyday noises.[...]Most of our memories are vague recollections,[...].However, we can estimate average capacity, although it has no clear meaning in our terms of cognition. That number would be 2.5 petabytes [...]Even so, this number varies as well from person to person.[...]

- Not all sentences in long form answer convey answer information
- About 40% of sentences serve other roles

 [How do we answer complex questions: Discourse structure of long form answers Xu et al, ACL 2022]

Difficulty of evaluating LFQA

Lexical matching based automatic metrics (e.g. ROUGE) are used, but not meaningful.

Q: Can you protect electronics from EMPs/solar flares? If so, how?

Random answer (ROUGE-L: 19.4)

The fast lane/slow lane is a bit of a misnomer. It gives the impression that new, faster lanes are being built. In reality, normal speed will be the new "fast lane", which will cost extra, and everything else will be in the throttled "slow lane".

Gold answer (ROUGE-L: 18.6)

I'll start with the grounding question, because that's the easiest to answer: Doesn't help a bit. All that matters is that the metal container is conductive and doesn't have gaps...completely seal your Faraday cage. Consider soldering the lid on to that paint can... look at little baggie it comes in. Sealed mylar. That protected that chip from air travel at 35,000 feet, land travel through rural, urban, and suburban areas, and all the electromagnetic radiation that the trip entails... No lead shielding. No safes…

 Krishna et al. 2021

Can humans evaluate long-form answers?

How does a speaker vibrate at multiple frequencies simultaneously to deliver sounds to our ears?

Answer A: This has been asked many times and the answer is they don't. If you listen to the song being played live on purely acoustic instruments even though they are being played separately and emitting their own frequencies, what you hear (and by extension, what a microphone captures) at any given time is just ONE frequency that's the "sum" of all the others combined. A speaker is just a reverse microphone.

Answer B: Imagine an ocean with a consistent wave. It flows up and down, with equal distance between the two waves at any time. Now imagine I push a larger, shorter wave into this ocean. The two waves will collide, resulting in some new wave pattern. This new wave pattern is a combination of those two waves. Speakers work similarly. If I combine two soundwaves, I get a new combination wave that sounds different.

Can experts evaluate?

Expert 1

Preference: A

In technical terms ocean waves stated in answer B are transverse waves and sound waves are longitudinal waves. In comparison answer B mentions about ocean waves and it is different to the sound waves in the question. But apart from that actually the two answers A and B go very close to each other and they provide similar explanations. But answer A is selected to be slightly better in terms of applicability and relevance. [...]

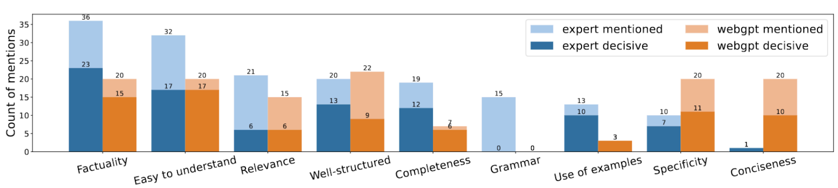
Expert 2

Preference: B

It is difficult to choose between these two answers because they both are not wrong and give essentially the same explanation. I go with answer B because I like the analogy with the ocean waves, and due to how visual the explanation is it is easier to understand in my opinion. [...]

- Even experts disagree on which one is a better answer

Evaluation aspects for LFQA



- Diverse facets are considered when evaluating LFQA answers.
- Best evaluation at the moment seems to be asking LLM whether it is a good answer or not, but not very reliable.

 Xu, Song, Iyyer and Choi, ACL 2023

Frontiers in QA

DROP

- QA datasets to model programs/computation

Passage (some parts shortened)	Question	Answer	BiDAF
That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million

- Question types: subtraction, comparison (*which did he visit first*), counting and sorting (*which kicker kicked more field goals*),
- Typically even systems like GPT-3 benefit from having a "calculator" they can call; many chain-of-thought variants with this structure

Dua et al. (2019)

NarrativeQA

- Humans see a summary of a book: *...Peter's former girlfriend Dana Barrett has had a son, Oscar...*
- Question: *How is Oscar related to Dana?*
- Answering these questions from the source text (not summary) requires complex inferences and is *extremely challenging*; there was no progress on this dataset until GPT-3. These are still some of the hardest types of questions!

Story snippet:

DANA (setting the wheel brakes on the buggy)
Thank you, Frank. I'll get the hang of this eventually.

She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy.

FRANK (to the baby)
Hiya, Oscar. What do you say, slugger?

FRANK (to Dana)
That's a good-looking kid you got there, Ms. Barrett.

Kočiský et al. (2017)

QA vs. Dialog

- Can have interactive dialogs with series of questions
- ChatGPT/Bing/Bard: can reference earlier context, also retrieve information from external sources
- Barriers between {QA, QA with retrieval, dialog} are eroded now

Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc

QuAC dataset; Eunsol Choi et al. (2019)

Takeaways

- Many individual QA datasets aren't perfect and have artifacts, but collectively, they test a wide range of capabilities
- QA over tables, images, knowledge bases, ...: all of this is unified and homogenized in GPT-era systems
- Integrating with retrieval and how to generate long answers are still two very challenging problems
- Major frontier: answers require reasoning beyond text: computation (although we can do this sometimes), physical simulation, statistical analysis, ...