

Retrieval-augmented LMs, and prompting

CS 685, Spring 2023
Advanced Natural Language Processing

Mohit Iyyer
College of Information and Computer Sciences
University of Massachusetts Amherst

Bob went to the <MASK>
to get a buzz cut

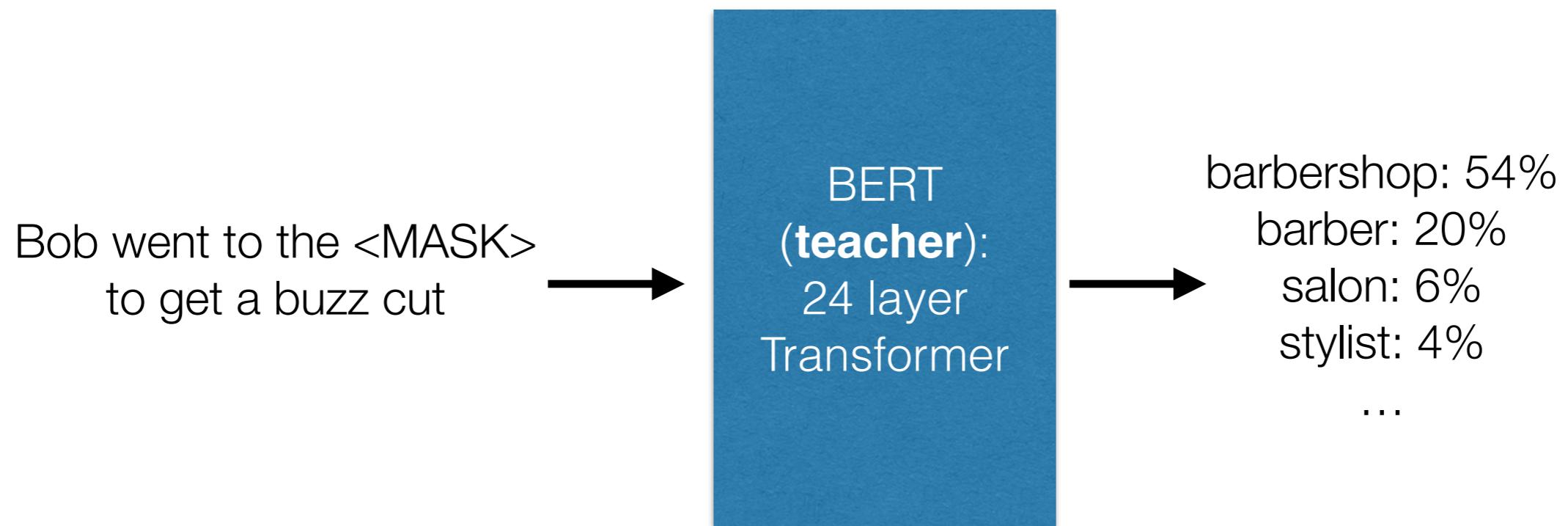


BERT
(teacher):
24 layer
Transformer



barbershop: 54%
barber: 20%
salon: 6%
stylist: 4%
...

World knowledge is *implicitly* encoded in BERT's parameters! (e.g., that barbershops are places to get buzz cuts)



In these language models, the learned world knowledge is stored *implicitly* in the parameters of the underlying neural network. This makes it difficult to determine what knowledge is stored in the network and where. Furthermore, storage space is limited by the size of the network—to capture more world knowledge, one must train ever-larger networks, which can be prohibitively slow or expensive.

Unlabeled text, from pre-training corpus (\mathcal{X})

The [MASK] at the top of the pyramid (x)

Unlabeled text, from pre-training corpus (\mathcal{X})

The [MASK] at the top of the pyramid (x)

Textual
knowledge
corpus (\mathcal{Z})

retrieve

Neural Knowledge Retriever $\sim p_{\theta}(z|x)$



Unlabeled text, from pre-training corpus (\mathcal{X})

The [MASK] at the top of the pyramid (x)

Textual knowledge corpus (\mathcal{Z})

retrieve

Neural Knowledge Retriever $\sim p_{\theta}(z|x)$

Retrieved document

The pyramidion on top allows for less material higher up the pyramid. (z)

Textual knowledge corpus (\mathcal{Z})

Unlabeled text, from pre-training corpus (\mathcal{X})

The [MASK] at the top of the pyramid (x)

retrieve

Neural Knowledge Retriever $\sim p_{\theta}(z|x)$

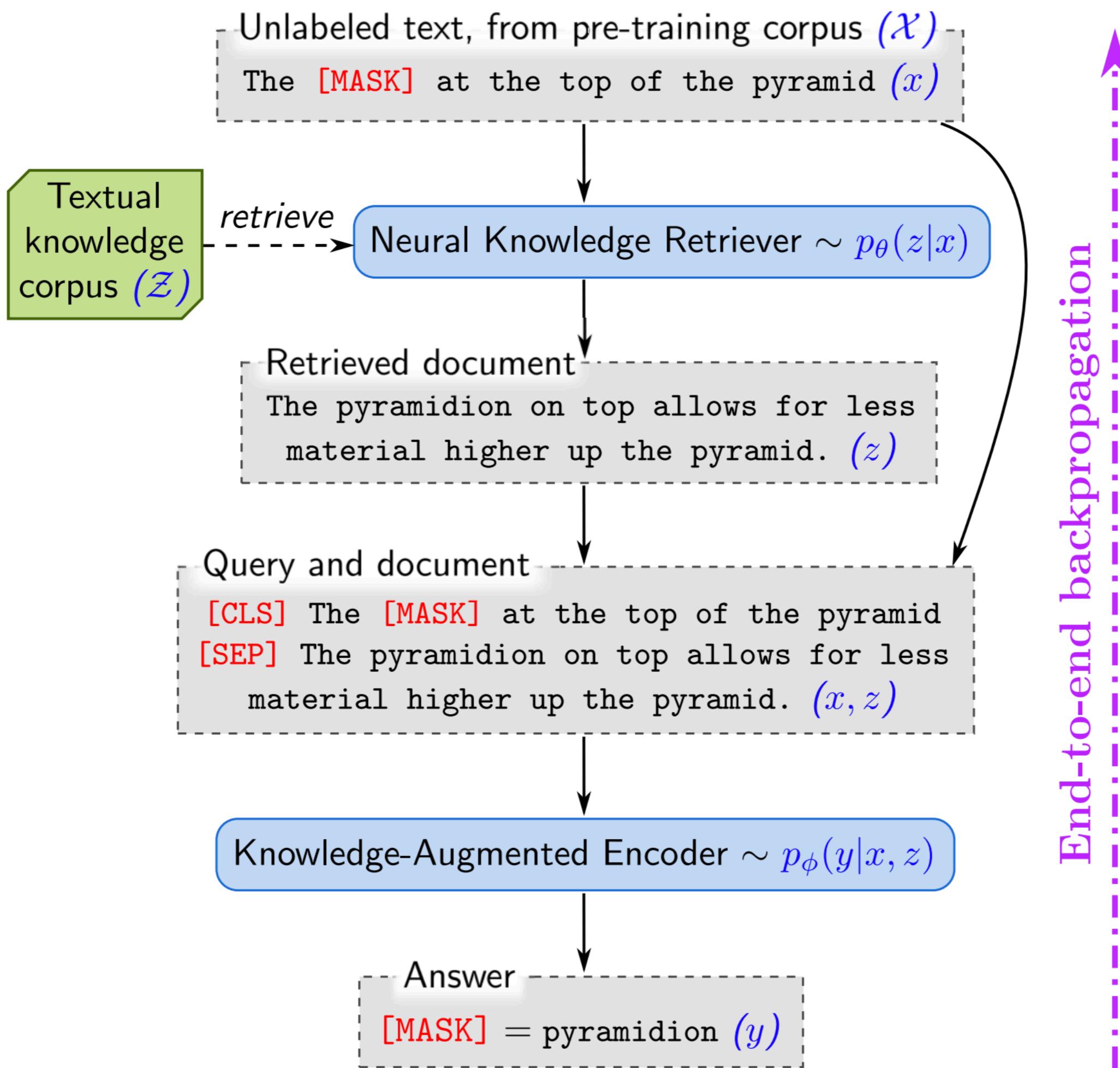
Retrieved document

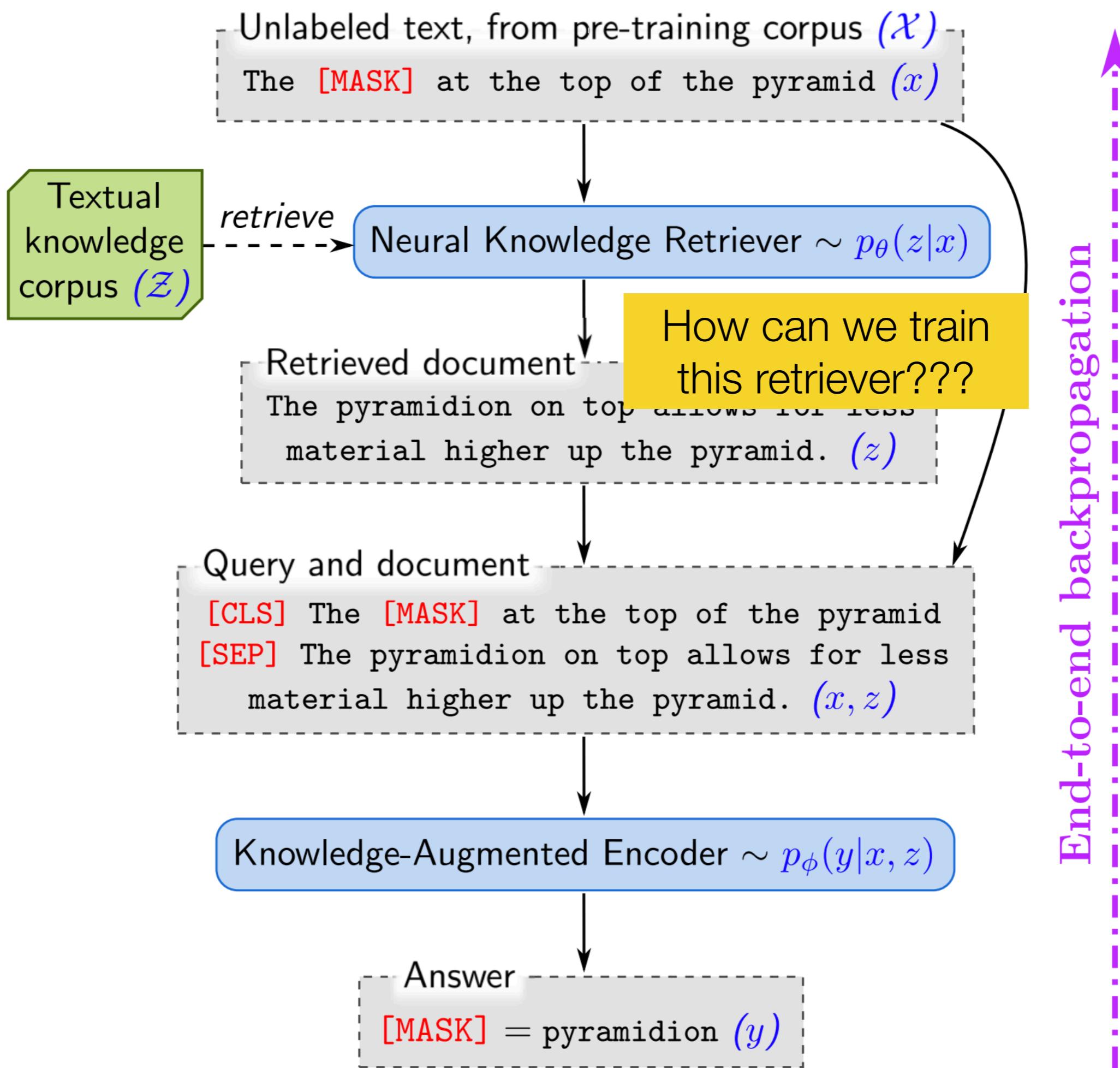
The pyramidion on top allows for less material higher up the pyramid. (z)

Query and document

[CLS] The [MASK] at the top of the pyramid

[SEP] The pyramidion on top allows for less material higher up the pyramid. (x, z)

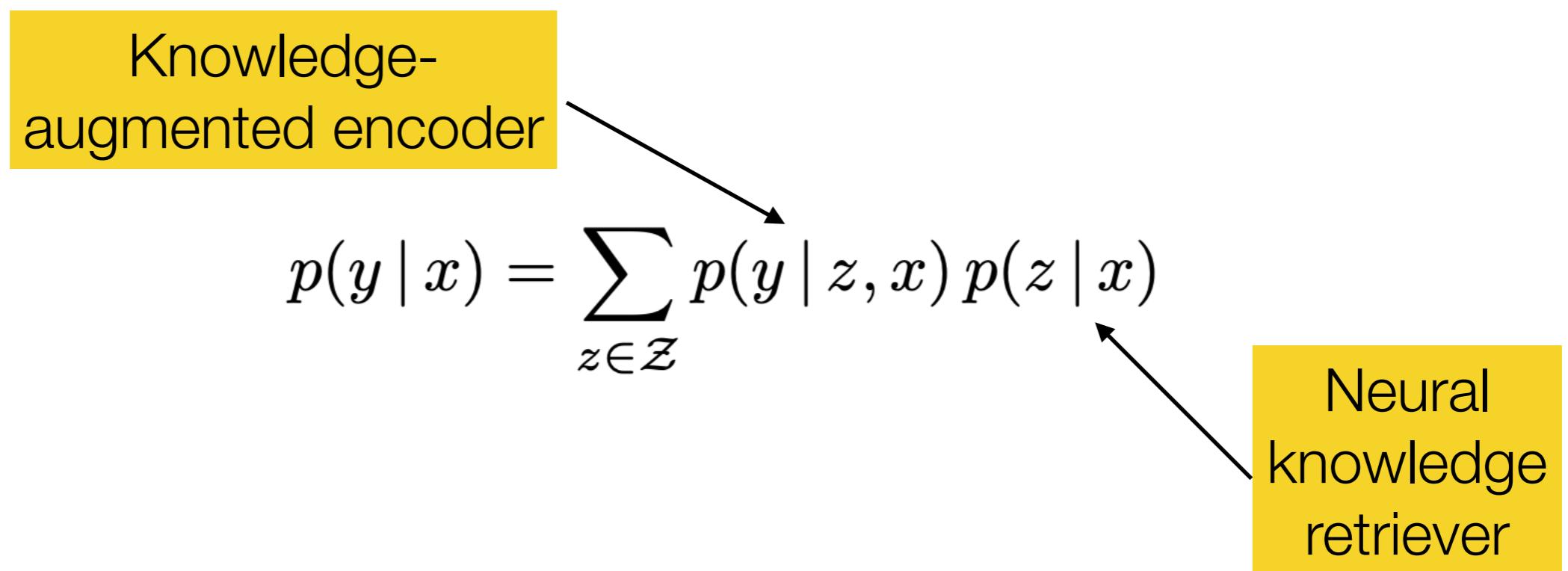




REALM decomposes $p(y | x)$ into two steps: *retrieve*, then *predict*. Given an input x , we first retrieve possibly helpful documents z from a knowledge corpus \mathcal{Z} . We model this as a sample from the distribution $p(z | x)$. Then, we condition on both the retrieved z and the original input x to generate the output y —modeled as $p(y | z, x)$. To obtain the overall likelihood of generating y , we treat z as a latent variable and marginalize over all possible documents z , yielding

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

REALM decomposes $p(y | x)$ into two steps: *retrieve*, then *predict*. Given an input x , we first retrieve possibly helpful documents z from a knowledge corpus \mathcal{Z} . We model this as a sample from the distribution $p(z | x)$. Then, we condition on both the retrieved z and the original input x to generate the output y —modeled as $p(y | z, x)$. To obtain the overall likelihood of generating y , we treat z as a latent variable and marginalize over all possible documents z , yielding



Knowledge Retriever The retriever is defined using a dense inner product model:

$$p(z|x) = \frac{\exp f(x, z)}{\sum_{z'} \exp f(x, z')},$$
$$f(x, z) = \text{Embed}_{\text{input}}(x)^\top \text{Embed}_{\text{doc}}(z),$$

where $\text{Embed}_{\text{input}}$ and $\text{Embed}_{\text{doc}}$ are embedding functions that map x and z respectively to d -dimensional vectors. The *relevance score* $f(x, z)$ between x and z is defined as the inner product of the vector embeddings. The retrieval distribution is the softmax over all relevance scores.

Knowledge-Augmented Encoder Given an input x and a retrieved document z , the knowledge-augmented encoder defines $p(y | z, x)$. We join x and z into a single sequence that we feed into a Transformer (distinct from the one used in the retriever).

$$p(y | z, x) = \prod_{j=1}^{J_x} p(y_j | z, x)$$
$$p(y_j | z, x) \propto \exp(w_j^\top \text{BERT}_{\text{MASK}(j)}(\text{join}_{\text{BERT}}(x, z_{\text{body}})))$$

where $\text{BERT}_{\text{MASK}(j)}$ denotes the Transformer output vector corresponding to the j^{th} masked token, J_x is the total number of [MASK] tokens in x , and w_j is a learned word embedding for token y_j .

Isn't training the retriever extremely expensive?

The key computational challenge is that the marginal probability $p(y | x) = \sum_{z \in \mathcal{Z}} p(y | x, z) p(z | x)$ involves a summation over all documents z in the knowledge corpus \mathcal{Z} . We approximate this by instead summing over the top k documents with highest probability under $p(z | x)$ —this is reasonable if most documents have near zero probability.

Imagine if your knowledge corpus was every article in Wikipedia... this would be super expensive without the approximation

Maximum inner product search (MIPS)

- Algorithms that *approximately* find the top- k documents
- Scales *sub-linearly* with the number of documents (both time and storage)
 - Srivastava and Li, 2014 (“Asymmetric LSH...”)
- Requires precomputing the BERT embedding of every document in the knowledge corpus and then building an index over the embeddings

Need to refresh the index!

- We are training the parameters of the retriever, i.e., the BERT architecture that produces **Embed_{doc}(z)**
- If we precompute all of the embeddings, the search index becomes stale when we update the parameters of the retriever
- REALM solution: asynchronously refresh the index by re-embedding all docs after a few hundred training iterations

Evaluation on *open-domain* QA

- Unlike SQuAD-style QA, in open-domain QA we are only given a question, not a supporting document that is guaranteed to contain the answer
- Open-domain QA generally has a large *retrieval* component, since the answer to any given question could occur anywhere in a large collection of documents

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

Table 3. An example where REALM utilizes retrieved documents to better predict masked tokens. It assigns much higher probability (0.129) to the correct term, “Fermat”, compared to BERT. (Note that the blank corresponds to 3 BERT wordpieces.)

<i>x:</i> An equilateral triangle is easily constructed using a straightedge and compass, because 3 is a ____ prime.			
(a)	BERT	$p(y = \text{“Fermat”} x) = 1.1 \times 10^{-14}$	(No retrieval.)
(b)	REALM	$p(y = \text{“Fermat”} x, z) = 1.0$	(Conditional probability with document $z = \text{“257 is ... a Fermat prime. Thus a regular polygon with 257 sides is constructible with compass ...”}$)
(c)	REALM	$p(y = \text{“Fermat”} x) = 0.129$	(Marginal probability, marginalizing over top 8 retrieved documents.)

Can retrieval-augmented
LMs improve other tasks?

Hurdles to Progress in Longform QA



Kalpesh
Krishna



Aurko Roy



Posted by u/halloichbins987 5 months ago 4 2 6 6 4 4 2

34.3k
↓

ELI5 why do humans need to eat many different kind of foods to get their vitamins etc but large animals like cows only need grass to survive?

Biology



1.7k Comments Share Save Hide Report

ELI5 dataset crawled from Reddit & filtered by Fan et al., ACL 2019, ~275K QA pairs

SORT BY **BEST** ▾

[View all comments](#)

[View discussions in 1 other community](#)



zapawu 5 months ago 3 & 6 More

In addition to the 'efficiently breaking down grass' thing, and the 'they eat a variety of plants' thing, there's also the fact that species typically evolve the ability to make vitamins that they can't get easily in their diet. For example, humans make vitamin D because there aren't many food sources of it, but we can't make vitamin C, but can find it in food. But other species can make their own vitamin C.

It's a trade off between needing to find a variety of food and not needing the cellular machines to make more stuff.

13.9k Reply Share Report Save

LFQA involves *retrieval* as well as generation

Why do humans need to eat many kinds of foods to get their vitamins but cows only need grass to survive?

LFQA involves *retrieval* as well as generation

Why do humans need to eat many kinds of foods to get their vitamins but cows only need grass to survive?



Ruminant

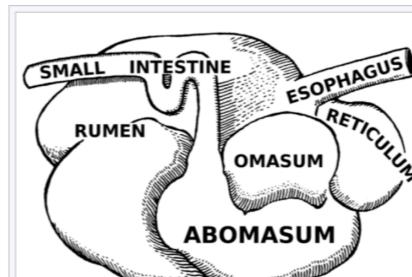
From Wikipedia, the free encyclopedia

Ruminants are herbivorous mammals of the suborder Ruminantia that are able to acquire nutrients from plant-based food by fermenting it in a specialized stomach prior to digestion, principally through microbial actions.

The process, which takes place in the front part of the digestive system and therefore is called **foregut fermentation**, typically requires the fermented ingesta (known as **cud**) to be regurgitated and chewed again. The process of rechewing the cud to further break down plant matter and stimulate digestion is called **rumination**.^{[1][2]} The word "ruminant" comes from the Latin *ruminare*, which means "to chew over again".

The roughly 200 species of ruminants include both domestic and wild species.^[3] Ruminating mammals include **cattle**, all domesticated and wild **bovines**, **goats**, **sheep**, **giraffes**, **deer**, **gazelles**, and **antelopes**.^[4] It has also been suggested that **notoungulates** also relied on rumination, as opposed to other **atlantogenates** that rely on the more typical **hindgut fermentation**, though this is not entirely certain.^[5]

Taxonomically, the suborder Ruminantia is a lineage of herbivorous **artiodactyls** that includes the most advanced and widespread of the world's **ungulates**.^[6] The suborder Ruminantia includes six different families: **Tragulidae**, **Giraffidae**, **Antilocapridae**, **Moschidae**, **Cervidae**, and **Bovidae**.^[3]



Stylised illustration of a ruminant digestive system



LFQA involves *retrieval* as well as generation

Why do humans need to eat many kinds of foods to get their vitamins but cows only need grass to survive?



Ruminant

Human digestive system

From Wikipedia, the free encyclopedia

Ruminants are herbivores that obtain energy by fermenting plant-based food by [microorganisms](#) in their [rumen](#). The process, which is called [rumination](#), typically begins after the animal has swallowed its food. The roughly 200 species of cattle, all domesticated, have been suggested to be the more typical [hippo](#).

Taxonomically, the ruminants are advanced and widely distributed. They include the families [Tragulidae](#), [Giraffidae](#),

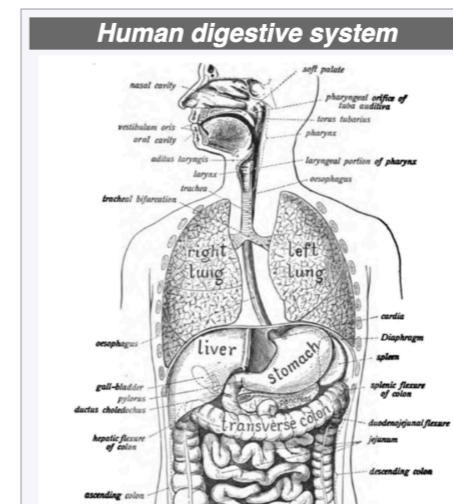
From Wikipedia, the free encyclopedia

See also [gastrointestinal tract](#).

"[Digestive system](#)" and "[alimentary system](#)" redirect here. For digestive systems of non-human animals, see [Digestion](#).

The **human digestive system** consists of the [gastrointestinal tract](#) plus the accessory organs of digestion (the [tongue](#), [salivary glands](#), [pancreas](#), [liver](#), and [gallbladder](#)). Digestion involves the breakdown of food into smaller and smaller components, until they can be absorbed and assimilated into the body. The process of digestion has three stages: the [cephalic phase](#), the [gastric phase](#), and the [intestinal phase](#).

The first stage, the [cephalic phase](#) of digestion, begins with gastric secretions in response to the sight and smell of food. This stage is controlled by the nervous system, and endocrine system. The nervous system, and endocrine system collaborate in the digestive system to control the movement of food throughout the tract. Saliva contains digestive enzymes, and motility associated with it is controlled by the [salivary glands](#) and [serous glands](#) on the tongue. Chewing, in which the food is mixed with saliva, begins the [gastric phase](#) of digestion. Here the food is further broken down by mixing with [gastric acid](#) until it passes into the [duodenum](#), the first part of the small



LFQA involves *retrieval* as well as generation

Why do humans need to eat many kinds of foods to get their vitamins but cows only need grass to survive?

Ruminant

Human digestive system

From Wikipedia, the free encyclopedia

Ruminants are herbivores that base food by fermentation. The process, which typically occurs in the rumen, is used to break down cellulose in plant material. The roughly 200 species of cattle, all domesticated, have been suggested to be the more typical herbivore. Taxonomically, the advanced and widely distributed families Tragulidae, Giraffidae, and Bovidae are included in the ruminants.

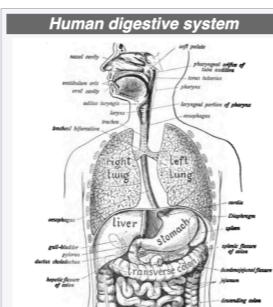
See also [gastrointestinal tract](#). "Digestive system" and "alimentary system" redirect here. For digestive systems of non-human animals, see [Digestion](#).

The human digestive system consists of the [gastrointestinal tract](#) plus the accessory organs of digestion (the tongue, salivary glands, pancreas, liver, and gallbladder). Digestion involves the breakdown of food into smaller and smaller components, until they can be absorbed and assimilated into the body. The process of digestion has three stages: the [cephalic phase](#), the [gastric phase](#), and the [intestinal phase](#).

The first stage, the [cephalic phase](#) of digestion, begins with gastric secretions in response to the sight and smell of food. This stage is controlled by the nervous system, and endocrine system collaborate in the digestive system to control the movement of food throughout the gastrointestinal tract, including peristalsis, and segmentation contractions.

Saliva contains digestive enzymes, t

he second stage of digestion is the [gastric phase](#). Here the food is further broken down by mixing with [gastric acid](#) until it passes into the [duodenum](#), the first part of the small intestine.



In addition to the “efficiently breaking down grass” thing and the “they eat a variety of plants” thing, there’s also the fact that species typically evolve the ability to make vitamins that they can’t easily get in their diet...

several issues with the dataset and
evaluation

(a) Many held-out questions are paraphrased in the training set.
Best answer to similar train
questions gets 27.4 ROUGE-L

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

- **Train Q1:** How does an EMP ruin electronics?
What does it do? How would they be fixed?
Can It be protected against? How?
- **Train Q2:** If Earth were hit with a massive
EMP, would all of our currently technology be
completely unusable permanently?
- **Train Q3:** Whenever a electromagnetic pulse
(EMP) is released what does it do to
electronics to disable them?
- **Train Q4:** If earth was hit with an EMP, could
we ever restore electricity? If not, why?
- **Train Q5:** What are solar flares and why does
it impact our electronics?
- **Train Q6.** When an EMP goes off, can the
electronics affected be replaced?

(a) Many held-out questions are paraphrased in the training set.
Best answer to similar train
questions gets 27.4 ROUGE-L

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

- **Train Q1:** How does an EMP ruin electronics?
What does it do? How would they be fixed?
Can It be protected against? How?
- **Train Q2:** If Earth were hit with a massive
EMP, would all of our currently technology be
completely unusable permanently?
- **Train Q3:** Whenever a electromagnetic pulse
(EMP) is released what does it do to
electronics to disable them?
- **Train Q4:** If earth was hit with an EMP, could
we ever restore electricity? If not, why?
- **Train Q5:** What are solar flares and why does
it impact our electronics?
- **Train Q6.** When an EMP goes off, can the
electronics affected be replaced?

This means the model doesn't
need to rely on retrieving
external documents, because it
already sees the answers to
most questions during training.
Our analysis shows that **81%** of
val questions have an exact
paraphrase in the training set.

(a) Many held-out questions are paraphrased in the training set. Best answer to similar train questions gets 27.4 ROUGE-L

(b) Simply retrieving answers to random unrelated training questions yields relatively high ROUGE-L, while actual gold answers underperform generations

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

- **Train Q1:** How does an EMP ruin electronics? What does it do? How would they be fixed? Can It be protected against? How?
- **Train Q2:** If Earth were hit with a massive EMP, would all of our currently technology be completely unusable permanently?
- **Train Q3:** Whenever a electromagnetic pulse (EMP) is released what does it do to electronics to disable them?
- **Train Q4:** If earth was hit with an EMP, could we ever restore electricity? If not, why?
- **Train Q5:** What are solar flares and why does it impact our electronics?
- **Train Q6.** When an EMP goes off, can the electronics affected be replaced?

Random Train Ans, 19.4 ROUGE-L

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...

Gold Answer, 18.6 ROUGE-L

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....

(a) Many held-out questions are paraphrased in the training set. Best answer to similar train questions gets 27.4 ROUGE-L

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

- **Train Q1:** How does an EMP ruin electronics? What does it do? How would they be fixed? Can It be protected against? How?
- **Train Q2:** If Earth were hit with a massive EMP, would all of our currently technology be completely unusable permanently?
- **Train Q3:** Whenever a electromagnetic pulse (EMP) is released what does it do to electronics to disable them?
- **Train Q4:** If earth was hit with an EMP, could we ever restore electricity? If not, why?
- **Train Q5:** What are solar flares and why does it impact our electronics?
- **Train Q6.** When an EMP goes off, can the electronics affected be replaced?

(b) Simply retrieving answers to random unrelated training questions yields relatively high ROUGE-L, while actual gold answers underperform generations

Random Train Ans, 19.4 ROUGE-L

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...

Gold Answer, 18.6 ROUGE-L

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....

ROUGE-L is clearly not a good metric for LFQA, as *gold* answers frequently underperform randomly selected unrelated answers

(a) Many held-out questions are paraphrased in the training set. Best answer to similar train questions gets 27.4 ROUGE-L

(b) Simply retrieving answers to random unrelated training questions yields relatively high ROUGE-L, while actual gold answers underperform generations

(c) Conditioning answer generation on *random* documents instead of relevant ones **does not** measurably impact its **factual correctness**. Longer outputs get higher ROUGE-L

(d) Annotators find it difficult to judge long answers (with **repetition**) & correctness of technical content

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

- **Train Q1:** How does an EMP ruin electronics? What does it do? How would they be fixed? Can It be protected against? How?
- **Train Q2:** If Earth were hit with a massive EMP, would all of our currently technology be completely unusable permanently?
- **Train Q3:** Whenever a electromagnetic pulse (EMP) is released what does it do to electronics to disable them?
- **Train Q4:** If earth was hit with an EMP, could we ever restore electricity? If not, why?
- **Train Q5:** What are solar flares and why does it impact our electronics?
- **Train Q6.** When an EMP goes off, can the electronics affected be replaced?

Random Train Ans, 19.4 ROUGE-L

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...

Generation using *predicted* retrievals, 19.0 ROUGE-L

Yes, you can shield them. But it's a slow process... Also, the equipment that's powered by them is a lot more expensive than you'd think, so it's hard to make sure that you're not just shielding them from your remote control.

Gold Answer, 18.6 ROUGE-L

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....

Generation using *random* retrievals, 24.8 ROUGE-L

Yes, you absolutely can, in fact you can build a **Faraday cage around your electronics, and protect them from solar flares**... This is what is done with the Faraday cage around your electronics, which is the problem. The reason it is **expensive** is because it requires a huge amount of power and is **expensive** to replace... **designed to shield your electronics from solar flares**, you will have to **pay for the protection**. This is because you have to buy a piece of equipment that is **designed to shield your electronics from solar flares, and that is expensive**. ... This is also **expensive**, but not as **expensive** as the protection you need to **shield your electronics from solar flares**... **designed to be as cheap as possible**...

We have a long way to go...

Q: Why are almost all boats white?

A: Boats are generally painted white for obvious reasons, as they are inexpensive and easy to clean. Ships are also often painted to provide a visible visual contrast to the sea. Boats also often have high visibility ports to allow for easy international trade. In addition, many boats have an air-breathing function that makes them easier to spot and stay afloat, which is a good thing if you're a pirate or trying to get into some country's waters. Lastly, many ships are built in the last few decades, so they are not as exotic looking as they were thirty years ago. And lastly, many ships are made out of materials that are easy to work with, and so many of them are painted white as a practical matter. There are many, many more reasons for why ships are white, but those are the ones I know of. I'm sure there are many other reasons, but those are the ones I know of. Edit: I forgot to mention that the other answers about visibility and easy to see ports are also true. Ships are painted to appear to be larger, and more expensive. That makes them appear more impressive, and so they are more likely to be bought by people who want to show off their wealth and status.

Prompting LLMs