

# Advances in Open Domain Long Form Question Answering

Comparison of AI21's Jurassic-X, and open source systems

A report by Amnon Lotenberg  
Sisense AIR team



**ELMo**

**ULMFiT**

Transformer

Bidirectional LM

**GPT**

Larger model  
More data

**GPT-2**

Defense



**Grover**



**BERT**

Cross-lingual

Multi-task  
+ Generation

**XLNet**  
**UDify**

**MT-DNN**

Knowledge distillation

**MT-DNN<sub>KD</sub>**

**MASS**  
**UniLM**

Span prediction  
Remove NSP

Longer time  
Remove NSP  
More data

**SpanBERT**

**RoBERTa**

Permutation LM  
Transformer-XL  
More data

**XLNet**

+Knowledge Graph



**ERNIE**  
**(Tsinghua)**

Neural entity linker

**KnowBERT**

Cross-modal

Whole Word Masking

**VideoBERT**  
**CBT**  
**ViLBERT**  
**VisualBERT**  
**B2T2**  
**Unicoder-VL**  
**LXMBERT**  
**VL-BERT**



**ERNIE (Baidu)**  
**BERT-wwm**

## Core Tasks

*Covered in  
Chapters 3–7*



Text  
Classification



Information  
Extraction



Conversational  
Agent



Information  
Retrieval



Question  
Answering Systems

## General Applications

*Covered in  
Chapters 4–7*



Spam  
Classification



Calendar Event  
Extraction



Personal  
Assistants



Search  
Engines

**JEOPARDY!**

Jeopardy!

## Industry Specific

*Covered in  
Chapters 8–10*



Social Media  
Analysis



Retail Catalog  
Extraction



Health Records  
Analysis

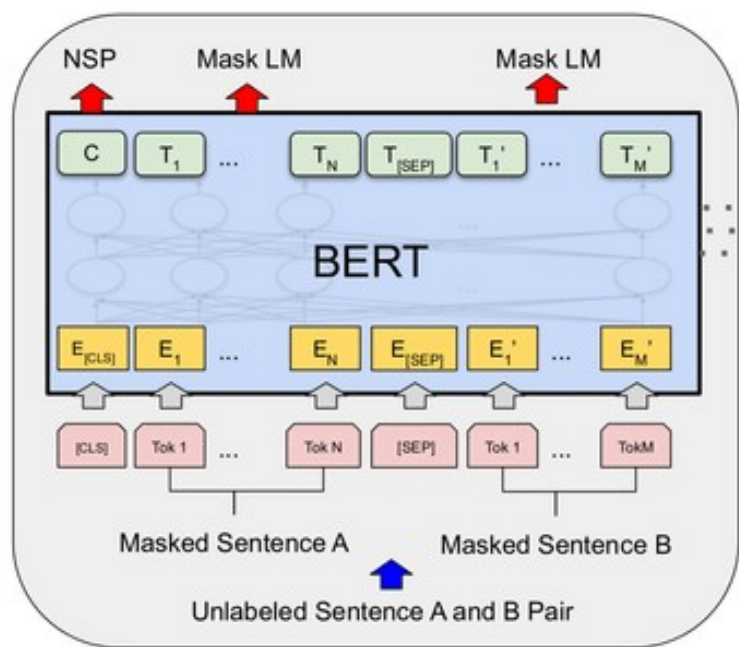


Financial  
Analysis

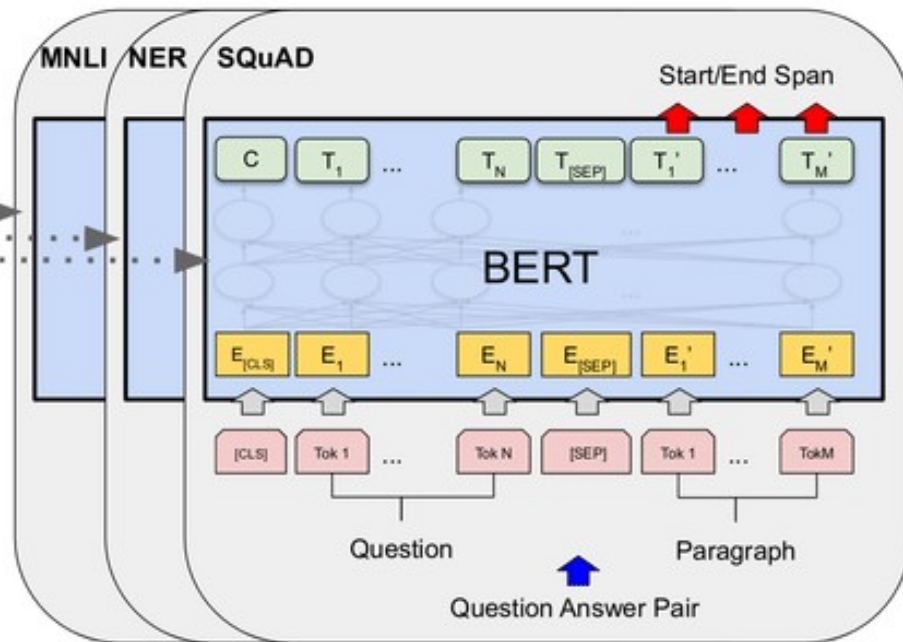


Legal Entity  
Extraction

# Transfer Learning

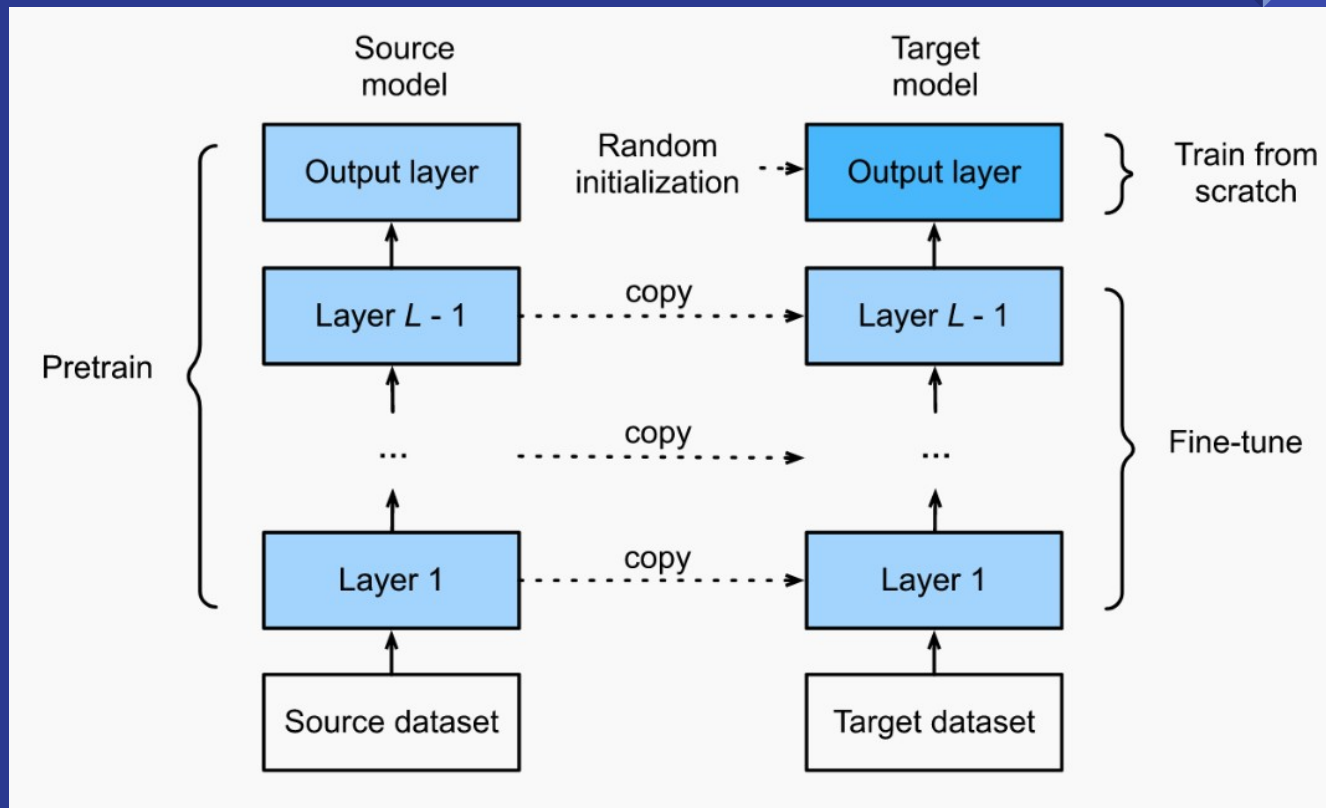


Pre-training



Fine-Tuning

# Transfer Learning



# Long Form Question Answering

Open Domain Question Answering

# Long Form Question Answering with ELI5

What would you like to ask? ---- select <MY QUESTION> to enter a new query

How do people make chocolate?

Go!

## The model generated answer is:

There are a lot of different ways, but the most common way is to grind up the cocoa beans and put it in a bowl with a little bit of butter. Then you heat the butter and let it melt. The melted butter is used to make the cocoa butter, and the melted butter makes the chocolate.

## The model is drawing information from the following Wikipedia passages:

01 - **Article:** The Licktators

Section: [Foundation](#)

02 - **Article:** Askinosie Chocolate

Section: [Products](#) & [Sustainable Nutrition Programs](#)

03 - **Article:** Outline of chocolate

Section: [Outline of chocolate](#)

04 - **Article:** Cocoa bean

Section: [Chocolate production](#)

05 - **Article:** Parrozzo

Section: [Ingredients and preparation](#)

06 - **Article:** Bristol Port Railway and Pier

Section: [Rolling stock](#)

07 - **Article:** Somebody's Mother's Chocolate Sauce

Section: [History](#) & [Products](#)

08 - **Article:** Ferrero-Küsschen

03 - **Article:** Outline of chocolate

Section: [Outline of chocolate](#)

04 - **Article:** Cocoa bean

Section: [Chocolate production](#)

05 - **Article:** Parrozzo

Section: [Ingredients and preparation](#)

06 - **Article:** Bristol Port Railway and Pier

Section: [Rolling stock](#)

07 - **Article:** Somebody's Mother's Chocolate Sauce

Section: [History](#) & [Products](#)

08 - **Article:** Ferrero-Küsschen

Section: [Ingredients](#) & [Sales](#)

09 - **Article:** Breakaway (biscuit)

Section: [Ingredients](#) & [Advertising](#)

10 - **Article:** Sugar (software)

Section: [Software components](#) & [History](#)

## The most similar question in the ELI5 training set was:

How is white chocolate made?

Its answers were:

1. White chocolate isn't really chocolate but it has a byproduct of cocoa processing. It's made with cocoa butter, sugar and milk solids.
2. Chocolate is a combination of several things including cocoa solids, cocoa butter, milk solids, sugar, fat, and a few other things. It's the cocoa solids which give it the chocolate color and taste.

White chocolate is all of that, *except* for the cocoa solids. It's not truly chocolate, but a chocolate derivative.

# LMs still struggle with LFQA

1. Big DBs/KBs are too vast to train on
2. Constantly updating
3. Lack of access to proprietary information sources
4. The *huge* LMs require long-context fine tuning in order to be used as readers
5. Even humans struggle with QA





Tamara  
Harvey

## Resolved Question

[Show me another »](#)

# I'm 10 and I havent started my period yet, am I pregnant?

Please help I'm really worried. I havent had sex but I kissed a boy last week and I think I tasted his sperm. i don't know what to do. I havent told anybody

7 months ago

[Report Abuse](#)



SteamGea...

## Best Answer - Chosen by Asker

Yes.

7 months ago

[Report Abuse](#)

21 people rated this as good

Asker's Rating: ★★★★★

Thanks. I think I'm going to have an abortion so its ok.

Action Bar:

9

Interesting! ▼

Email

Comment (5)

Save ▼

# Outline

1. Problem definition and challenges
2. Datasets and benchmarks for ODQA/LFQA
3. ELI5-based system
4. Jurassic-X
5. recap

# Datasets



- + 270K question/answer pairs from the Reddit forum, "Explain Like I'm Five" (ELI5)
- does not include gold retrievals of supporting docs

**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. [...] Jellyfish do not have brains, and most barely have nervous systems. They have primitive nerve cells that help them orient themselves in the water and sense light and touch. [...] While they don't possess brains, the animals still have neurons that send all sorts of signals throughout their body. "It is not true that jellyfish have no central nervous systems. They have an unusual nervous system," [...]

**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.

# LFQA - an improvement on ELI5

- 226,147 train, 3,020 validation, and 10,000 test samples
- inc. more subreddits
- Less overlap between train/validation/test sets
- Answers tend more to be found in the context

<https://aclanthology.org/2021.naacl-main.393.pdf>

<https://towardsdatascience.com/long-form-qa-beyond-eli5-an-updated-dataset-and-approach-319cb841aabb>

# Natural Questions

by Google

±300K examples of

- + A naturally occurring question
- + A wiki page which *might* hold the answer
- + human annotated answers, in the form of passages from Wikipedia

Has open book and closed book versions

Open book is simpler than ELI5 in the sense that you get the wiki page (~100 passages)

Also suffers from substantial train /

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

## Tokyo Imperial Palace

From Wikipedia, the free encyclopedia

Coordinates: 35°36′25″N 139°45′21″E﻿ / ﻿

The **Tokyo Imperial Palace** (皇居 Kōkyō, literally "Imperial Residence") is the primary residence of the **Emperor of Japan**. It is a large park-like area located in the Chiyoda ward of Tokyo and contains buildings including the main palace (宮殿 Kyūden), the private residences of the Imperial Family, an archive, museums and administrative offices. It is built on the site of the old Edo Castle. The total area including the gardens is 1.15 square kilometres (0.44 sq mi)<sup>[1]</sup>. During the height of the 1980s Japanese property bubble, the palace grounds were valued by some to be more than the value of all of the real estate in the state of California.<sup>[2][3]</sup>

**Contents** [hide]




- History
  - 1.1 Edo castle
  - 1.2 Old palace
  - 1.3 Present palace
- Grounds
  - 2.1 Kyūden
  - 2.2 Fukiage Garden
- East Garden
  - 3.1 Tōkagakudō (Music Hall)
  - 3.2 Ninomaru Garden
  - 3.3 Suwa no Chaya
- Kitanomaru
- Kōkyō-gaien
- Gallery
- References
- External links

**long answer**

**History**

Panorama of the Imperial Palace in Tokyo

Main gate to the Imperial Palace



Question:

who lives in the imperial palace  
in tokyo?

Short Answer:

the Imperial Family

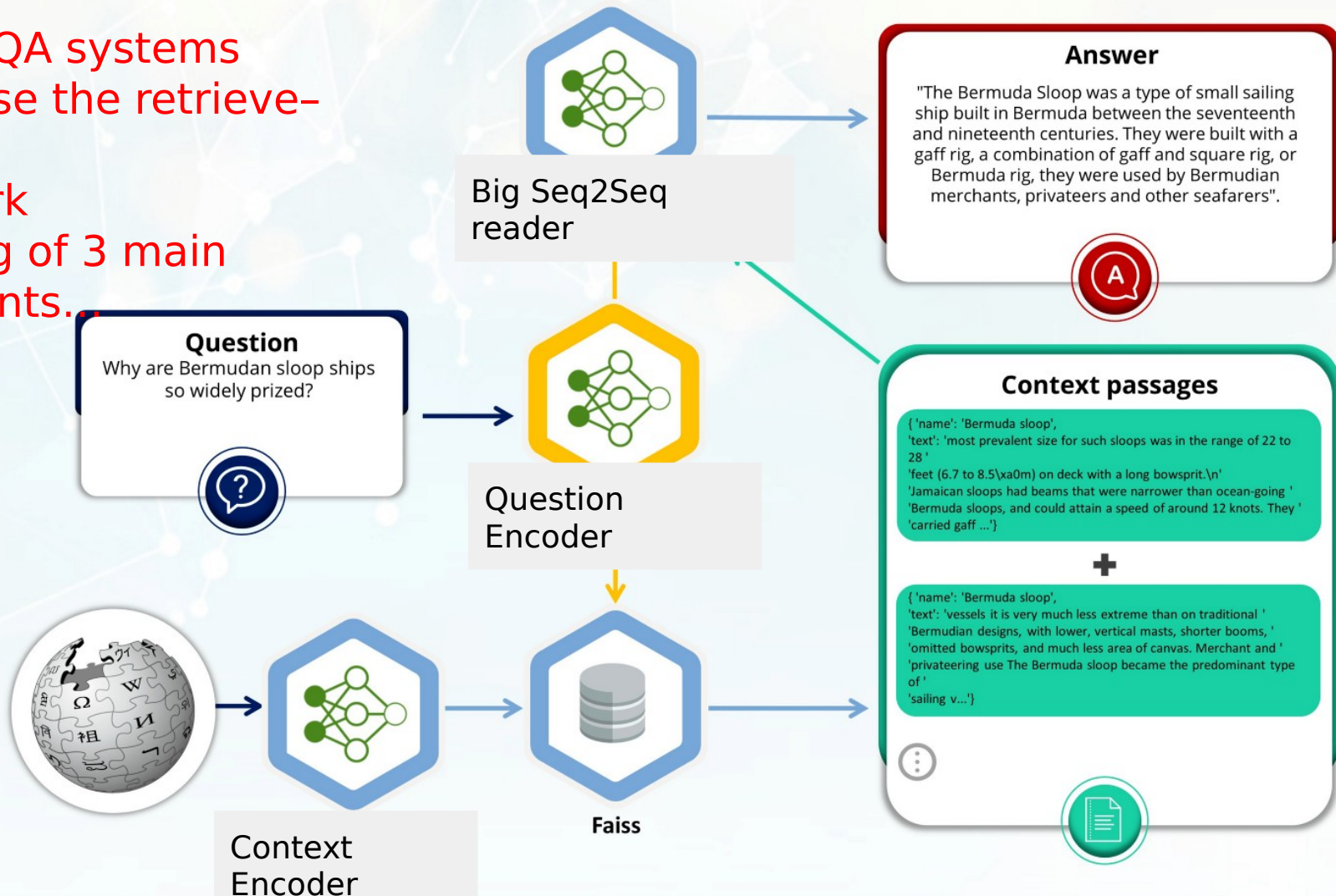
Long Answer:

The Tokyo Imperial Palace (皇居, Kōkyō, literally "Imperial Residence") is the primary residence of the Emperor of Japan. It is a large park-like area located in the Chiyoda ward of Tokyo and contains buildings including the main palace (宮殿, Kyūden), the private residences of the Imperial Family, an archive, museums and administrative offices.

# An anatomy of an LFQA System

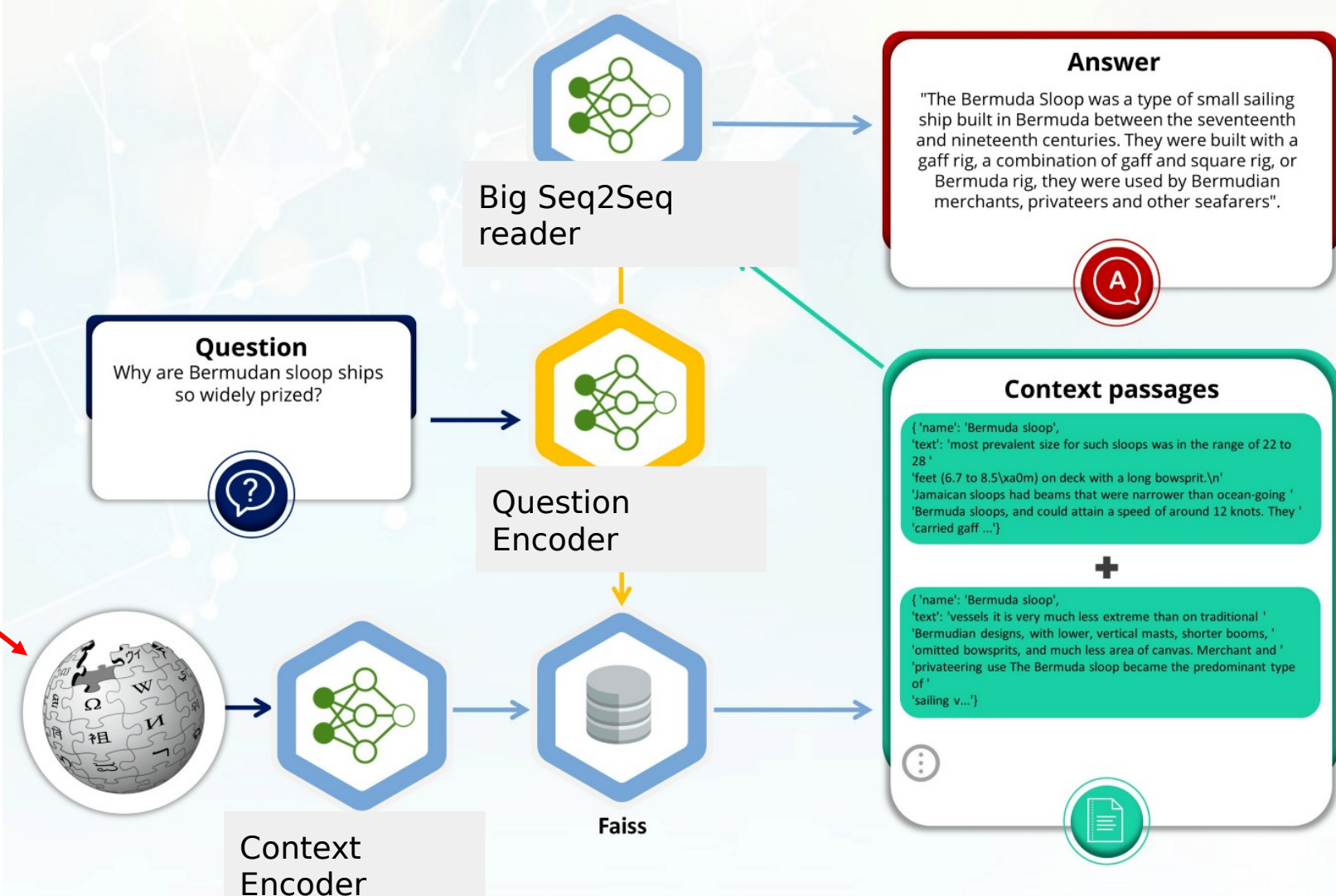
Based on open sourced work of Yacine Jernite and the ELI5/LFQA teams working with HuggingFace

LFQA/ODQA systems usually use the retrieve-read framework consisting of 3 main components.



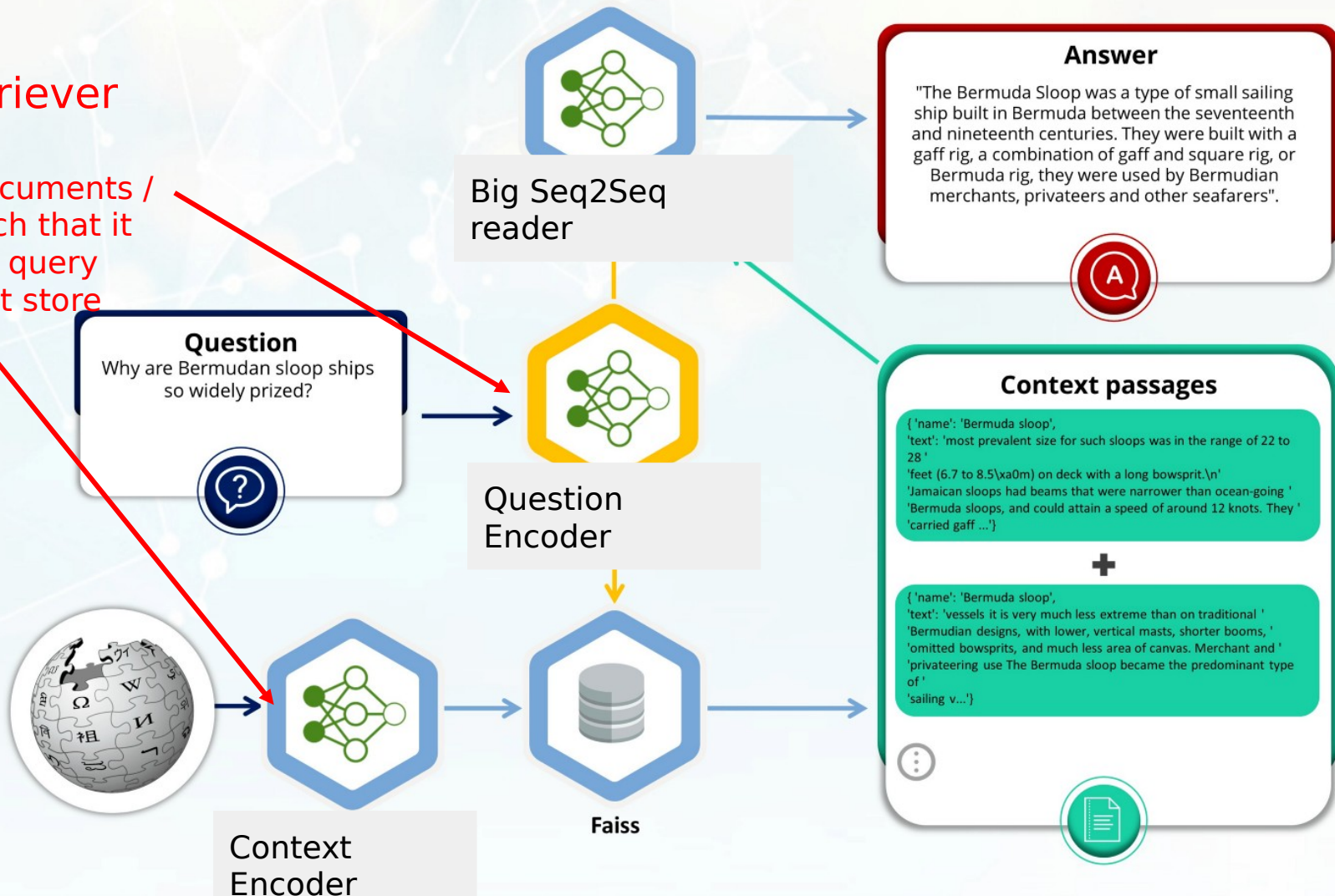


Document Store

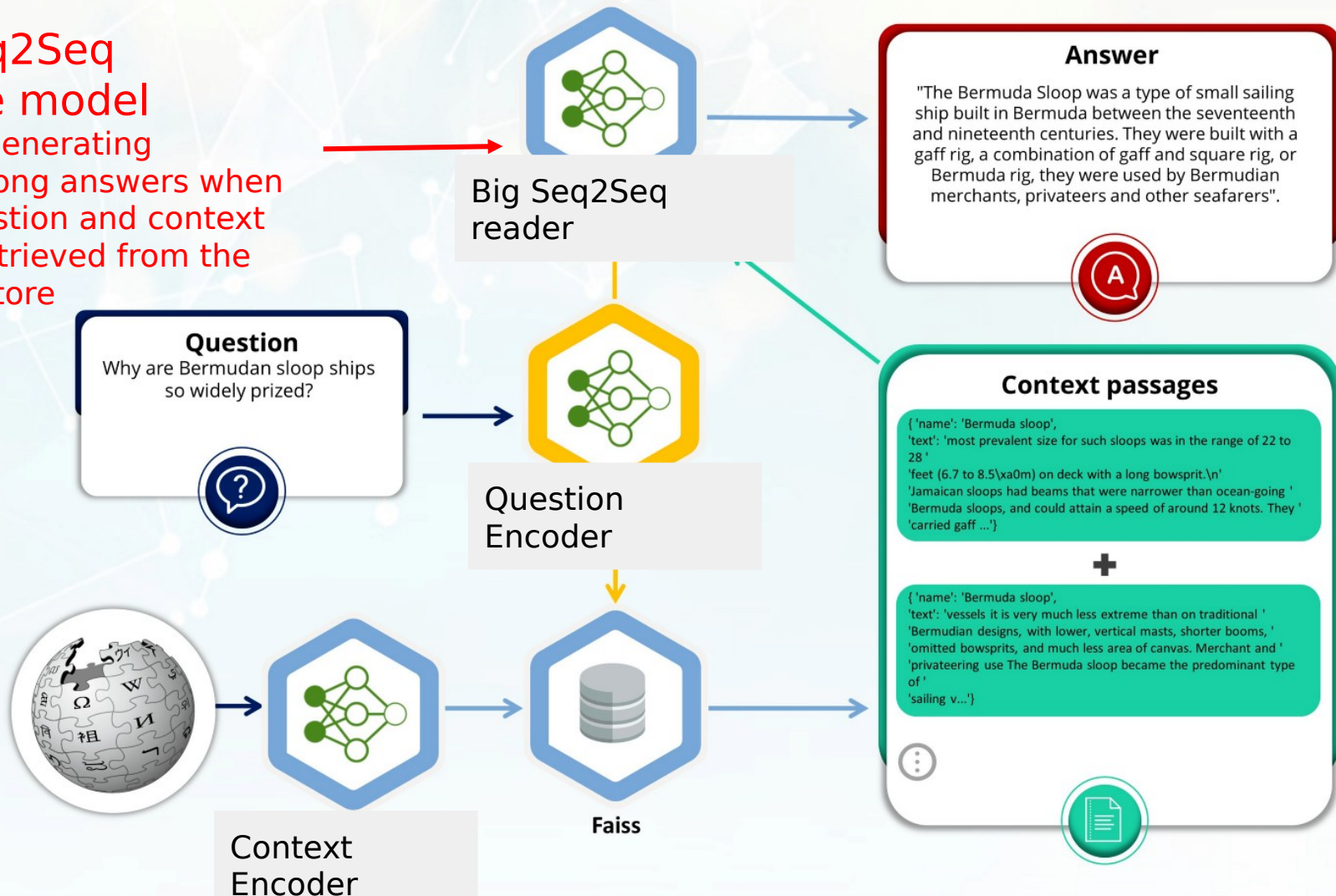


## Small Retriever model(s)

to encode documents / questions such that it is possible to query the document store



A big Seq2Seq language model capable of generating paragraph-long answers when given a question and context passages retrieved from the document store



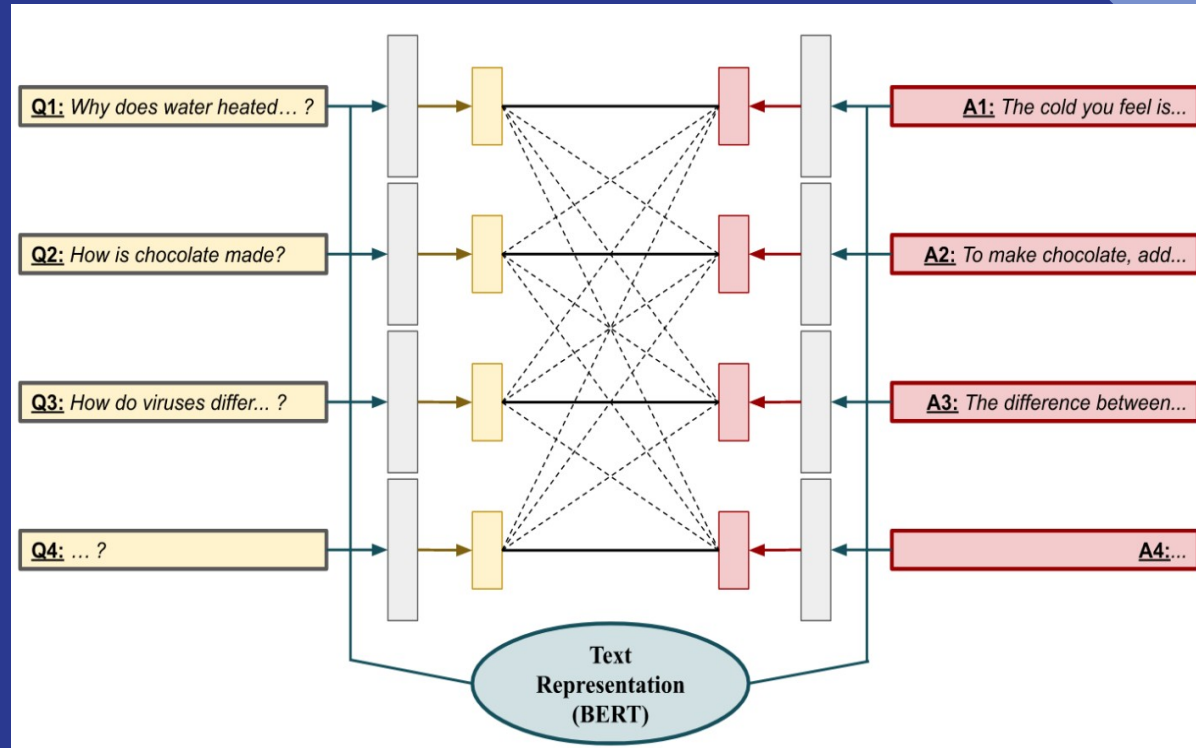
# Retriever *contrastive* training

If the Qs and As are similarly styled, you can use a single LM for both

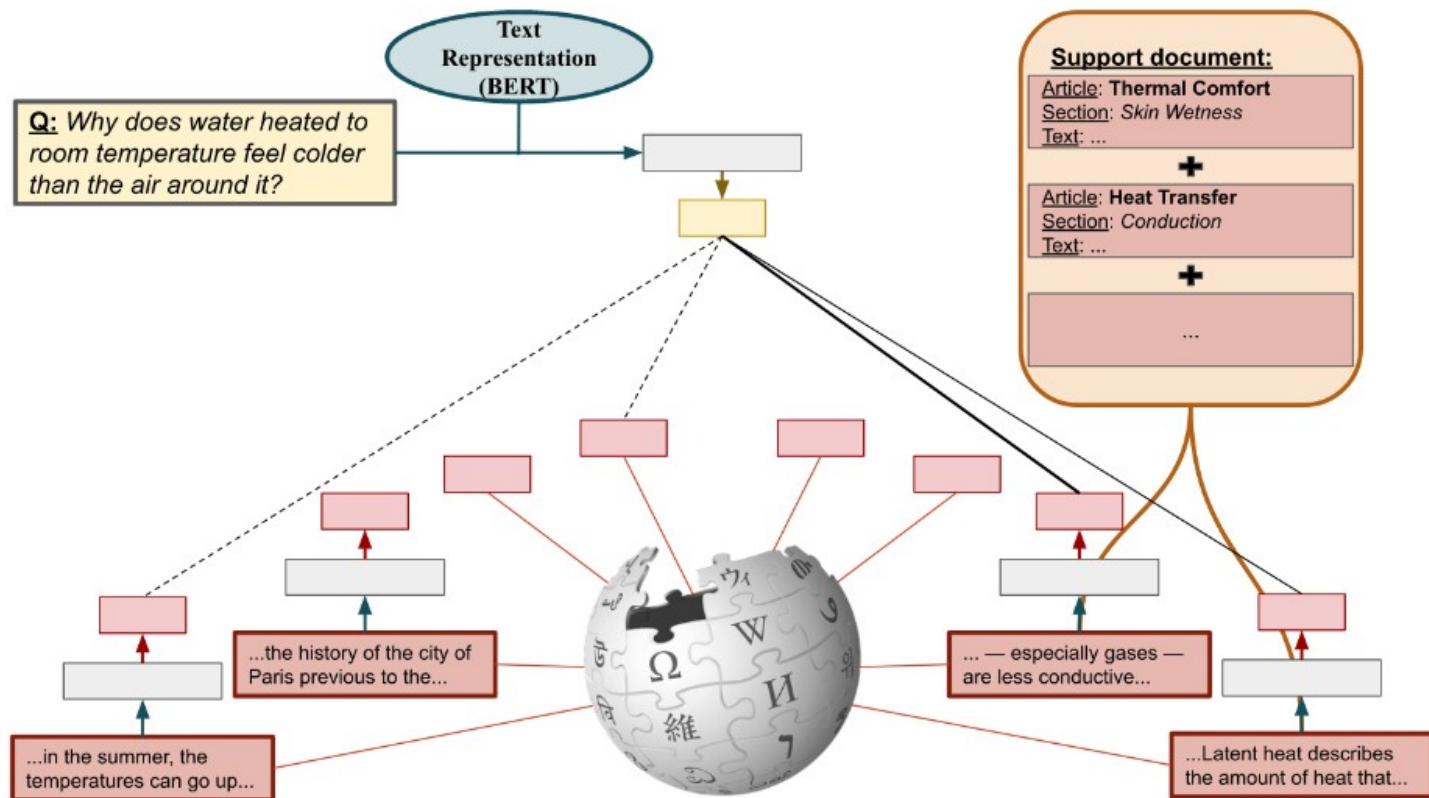
Fine tune a *small* LM, DistilBERT, ~100M parameters

Compare 1 batch of 512 Q+A pairs at a time

Output short vectors (128 dimensions), efficient for GPUs

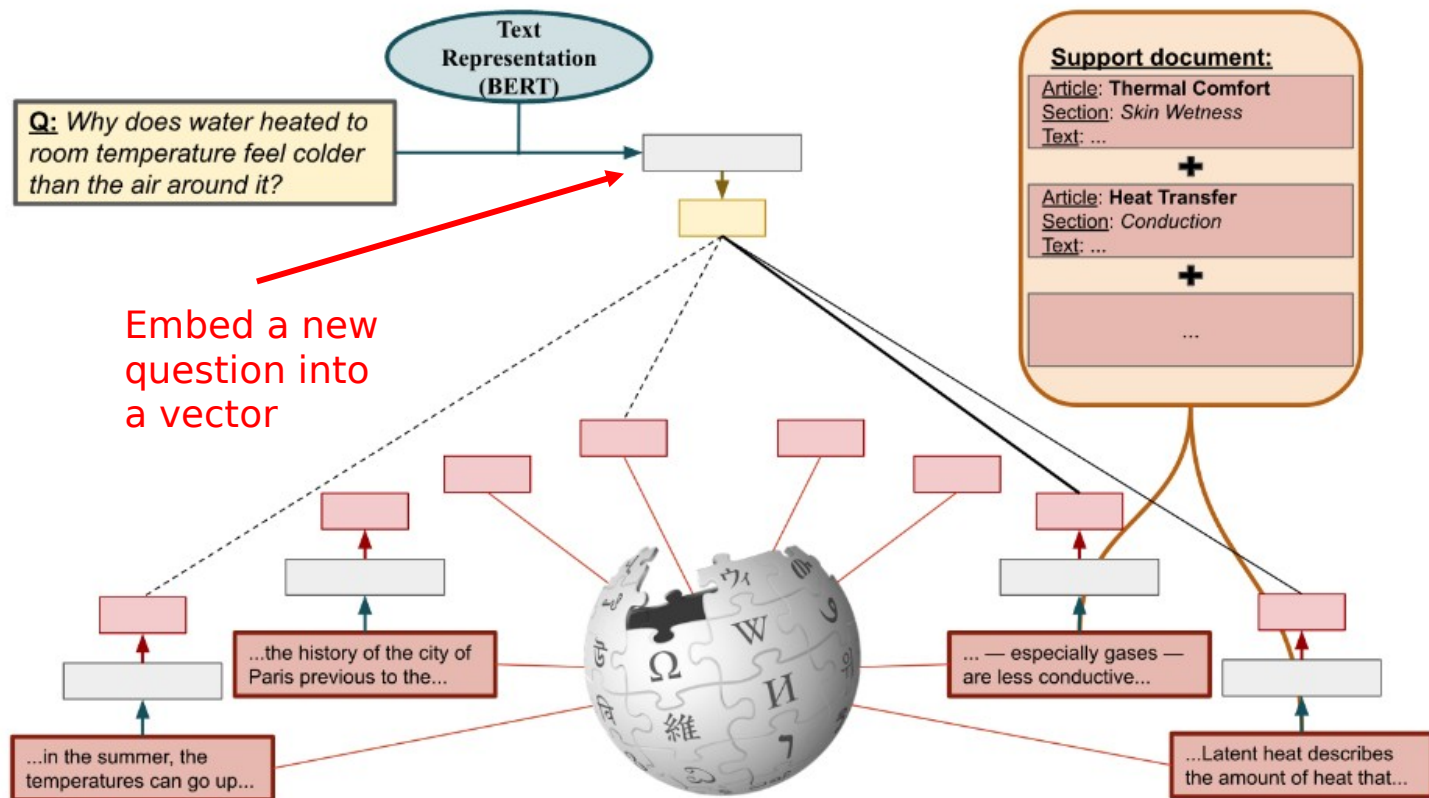


# The trained Retriever at work

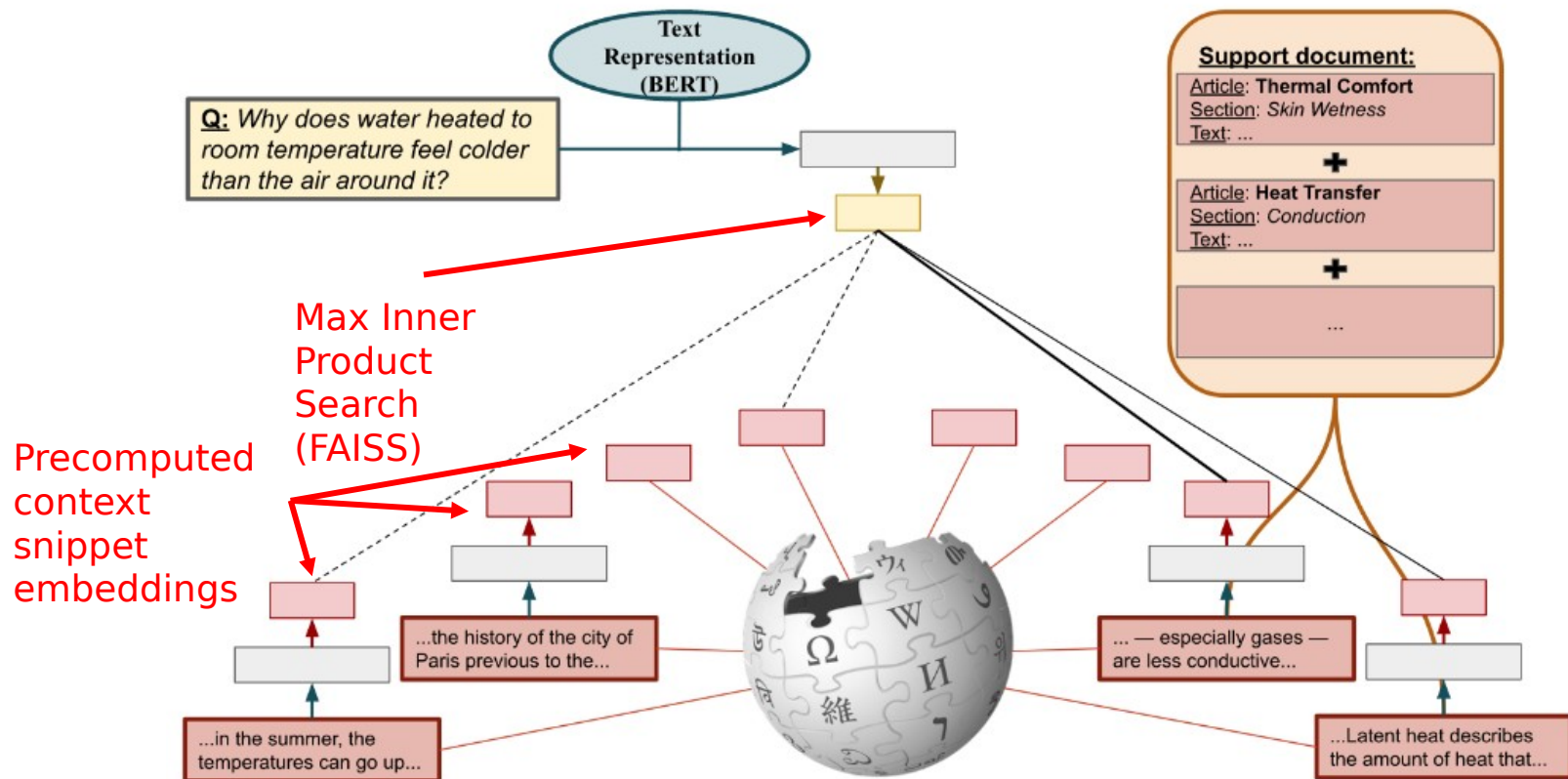




# The trained Retriever at work

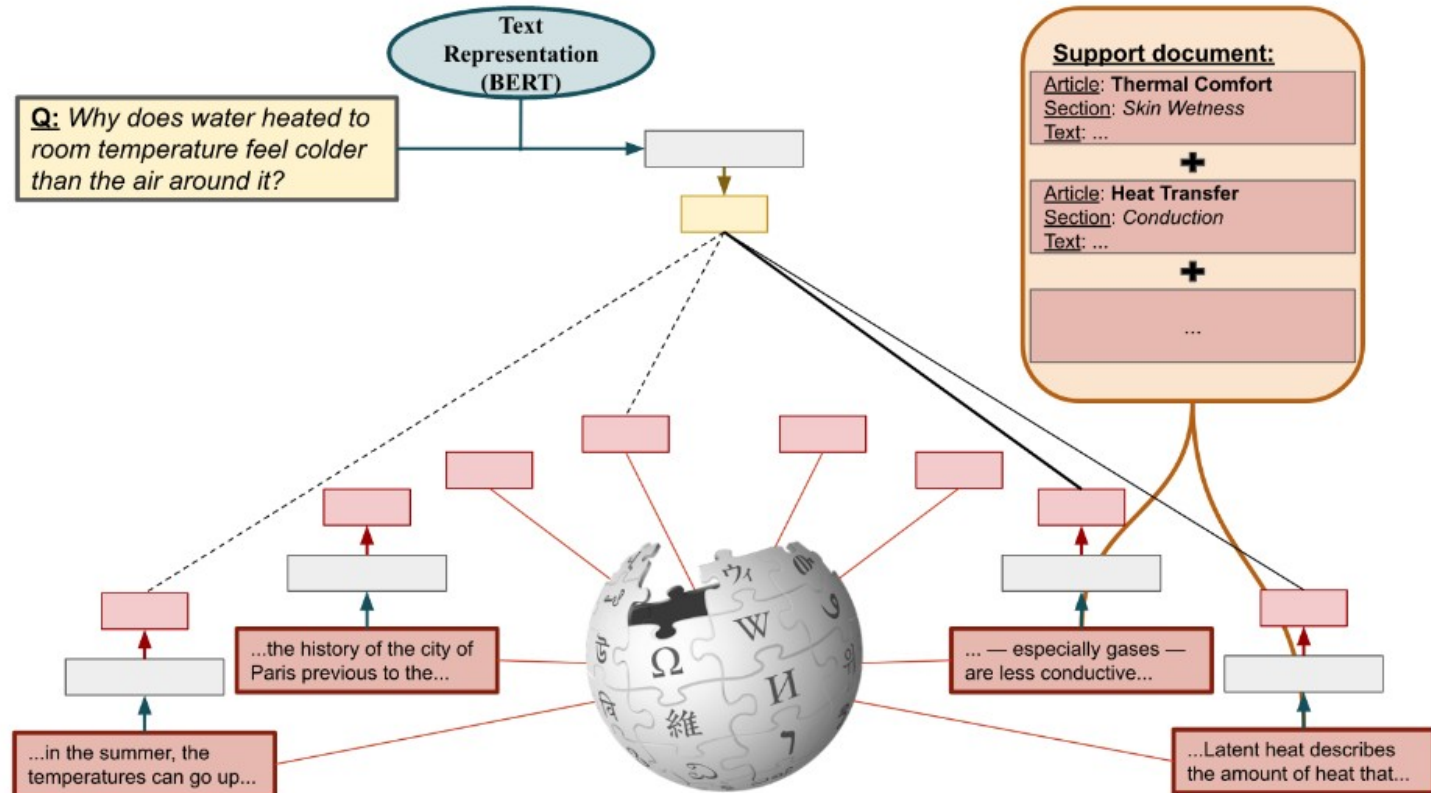


# The trained Retriever at work



# The trained Retriever at work

10 best  
passages





Article		Sections	Text
0	---	---	--- Why does water heated to room temperature feel colder than the air around it?
1	Heat transfer	Heat transfer in the human body & Evaporative cooling	when the skin is completely wet. The body continuously loses water by evaporation but the most significant amount of heat loss occurs during periods of increased physical activity. Evaporative cooling Evaporative cooling happens when water vapor is added to the surrounding air. The energy needed to evaporate the water is taken from the air in the form of sensible heat and converted into latent heat, while the air remains at a constant enthalpy. Latent heat describes the amount of heat that is needed to evaporate the liquid; this heat comes from the liquid itself and the surrounding gas and surfaces.
2	Johan Sandström	Sandström Theorem	at greater pressures. There is an ambiguity, however, as to the meaning of the terms 'heating' and 'cooling' in Sandstrom's theorem. So far, heating and cooling has always been interpreted in the literature as being associated with 'surface heating' and 'surface cooling' respectively. In real fluids, however, molecular and turbulent diffusion always cause internal heating/cooling even in absence of external heating/cooling, as long as the temperature of the fluid considered is non-uniform. As is well-known, molecular and turbulent diffusion tends to relax the system toward thermodynamic equilibrium, i.e., toward an isothermal state, which for a statically stable fluid, will warm up
3	Thermal equilibrium	Bodies prepared with separately uniform temperatures, then put into purely thermal communication with each other	are not in a relation of thermal equilibrium, heat will flow from the hotter to the colder, by whatever pathway, conductive or radiative, is available, and this flow will continue until thermal equilibrium is reached and then they will have the same temperature. One form of thermal equilibrium is radiative exchange equilibrium. Two bodies, each with its own uniform temperature, in solely radiative connection, no matter how far apart, or what partially obstructive, reflective, or refractive, obstacles lie in their path of radiative exchange, not moving relative to one another, will exchange thermal radiation, in net the hotter transferring energy to

# Reader: Generating Answers with a large Sequence-to-Sequence LMs

Fine tune BART-large

Input: concatenate a question and the supporting texts from the retriever

question: Why does water heated to room temperature feel colder than the air around it? context: \<P> when the skin is completely wet. The body continuously loses ... this heat comes from the liquid itself and the surrounding gas and surfaces. \<P> protected by a glass panel. Consequently, these types of collectors... Since heat loss due to convection cannot cross a vacuum, it forms an efficient isolation mechanism to keep heat inside the collector pipes. Since two flat \<P> ... \<P> changes. Conduction On... Fluids—especially gases—are less conductive. Thermal contact conductance is the study of heat conduction between solid bodies in contact. The process of heat transfer

Label is the gold answer

Loss func: minimize the perplexity of the gold answer

# Results

	ELI5		
	R1	R2	RL
Best Extractive	23.5	3.1	17.5
Language Model	27.8	4.7	23.1
Seq2Seq	28.3	5.1	22.8
Seq2Seq Multitask	28.9	5.4	23.1
BART	<b>30.6</b>	<b>6.2</b>	<b>24.3</b>

Yacine Jernite    27.3    5.5    25.8

# Demo

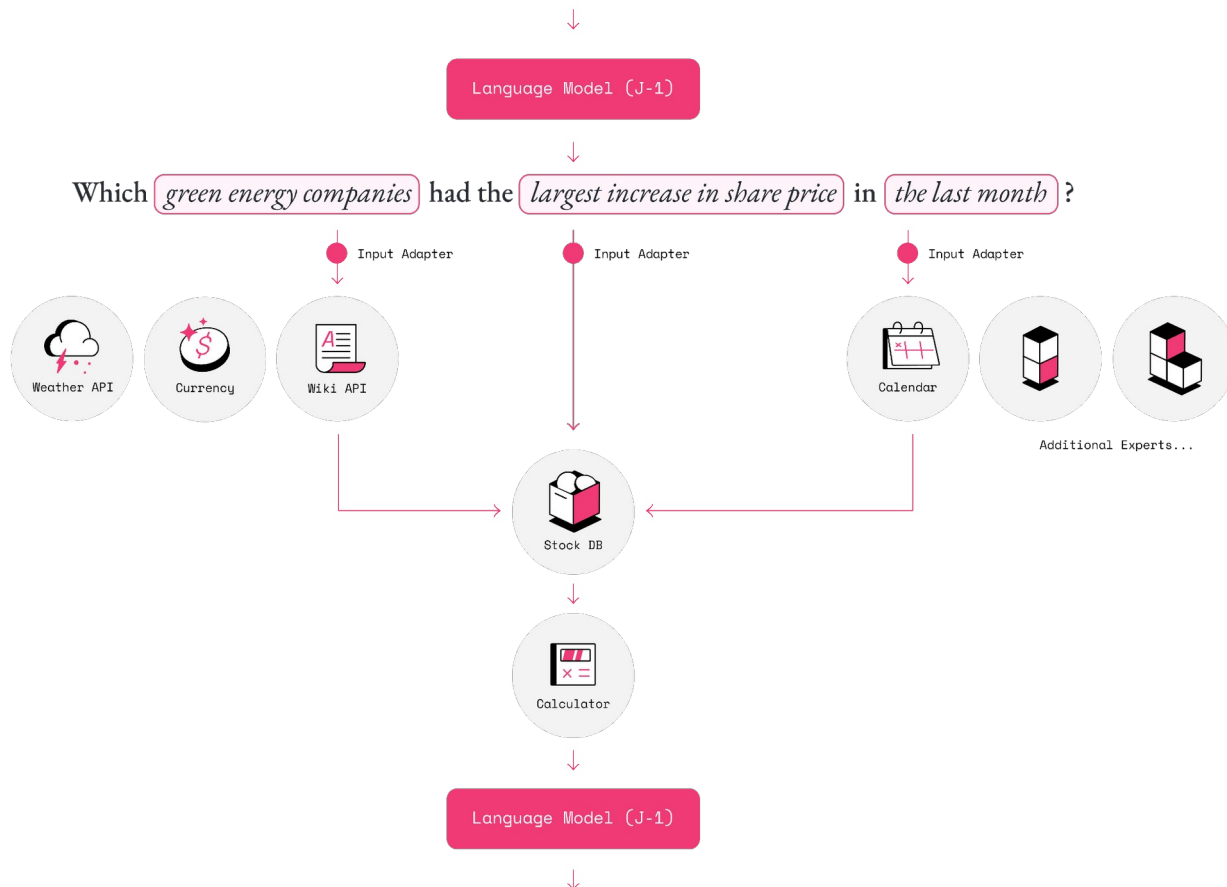
<https://huggingface.co/spaces/lfqa/lfqa>

# AI21's Jurassic-X



# AI

Which green energy companies had the largest increase in share price in the last month?



Windenergy and Tinergy grew the most over the past month, by over 12%

# AI21's Jurassic-X

QA function



Was Clinton ever elected president of the United States?

TO

Yes



GPT-3

No, Clinton was never elected as president of the United States.



Jurassic-1

No



Google

Clinton was elected president in the 1992 presidential elections...



Jurassic-X

Bill Clinton was elected president.



# AI21's Jurassic-X

LFQA function

“The best way these days to optimize performance for a given NLP task is usually to fine tune a pretrained LM...”





# AI21's Jurassic-X for LFQA

## Cons of fine tuning

“Catastrophic forgetting”:  
performance  
degrades on other  
tasks

Too expensive to fine  
tune for lots of tasks  
and domains

Need to fine tune just  
to increase the input

## J-X's Goals

Prevent forgetting,  
retain versatility

Extensibility

Cheap training

## Method

Frozen readers

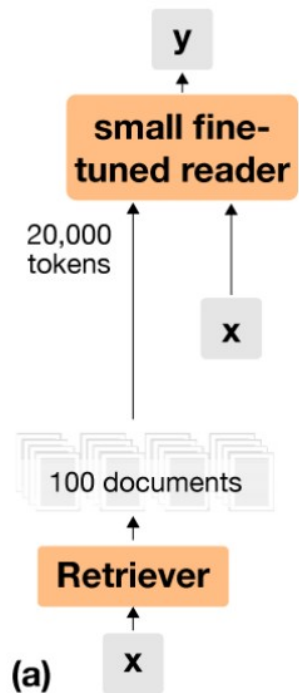
Recursive LMs

# AI21's Jurassic-X

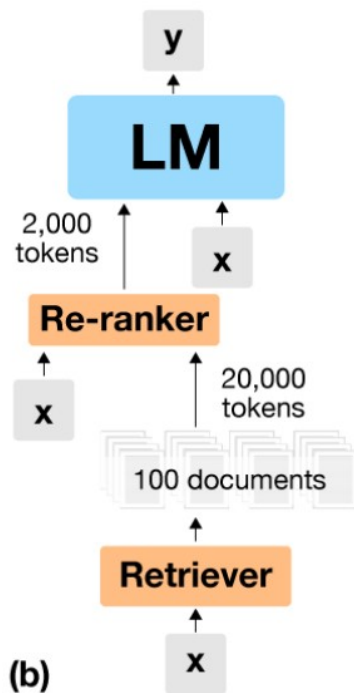
LFQA function

A FROZEN LM READER FOR OPEN-DOMAIN  
QUESTION ANSWERING  
In *open book* LFQA

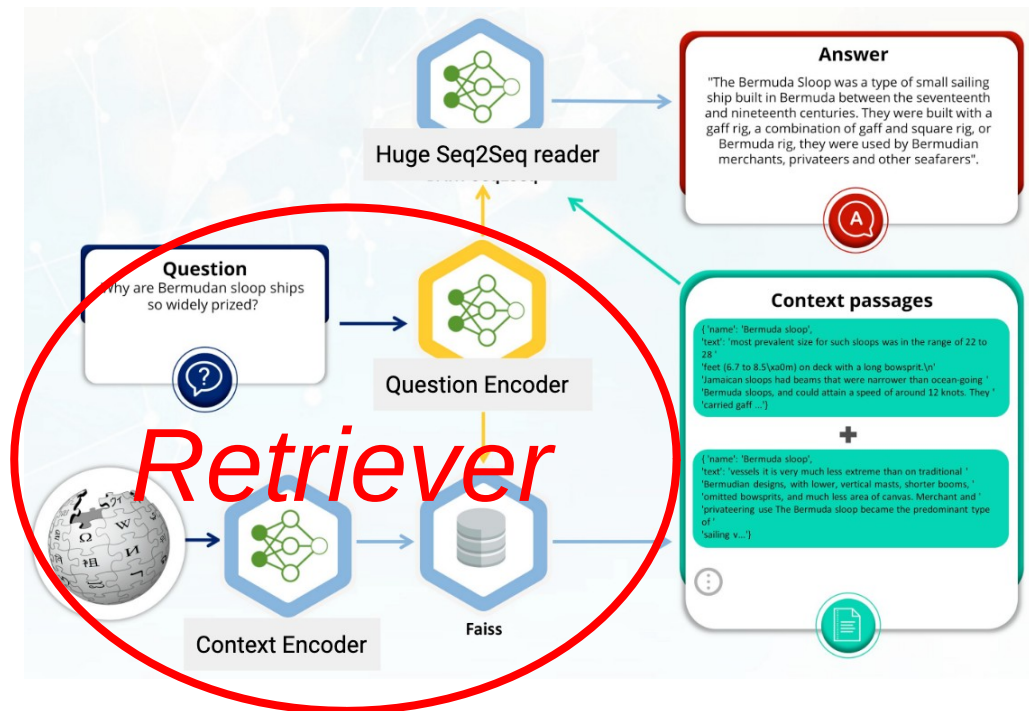


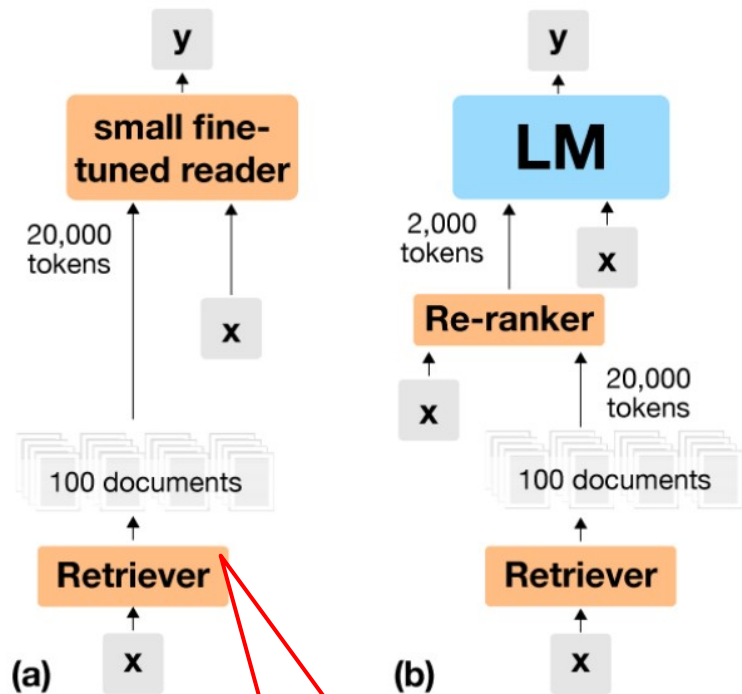


(a)

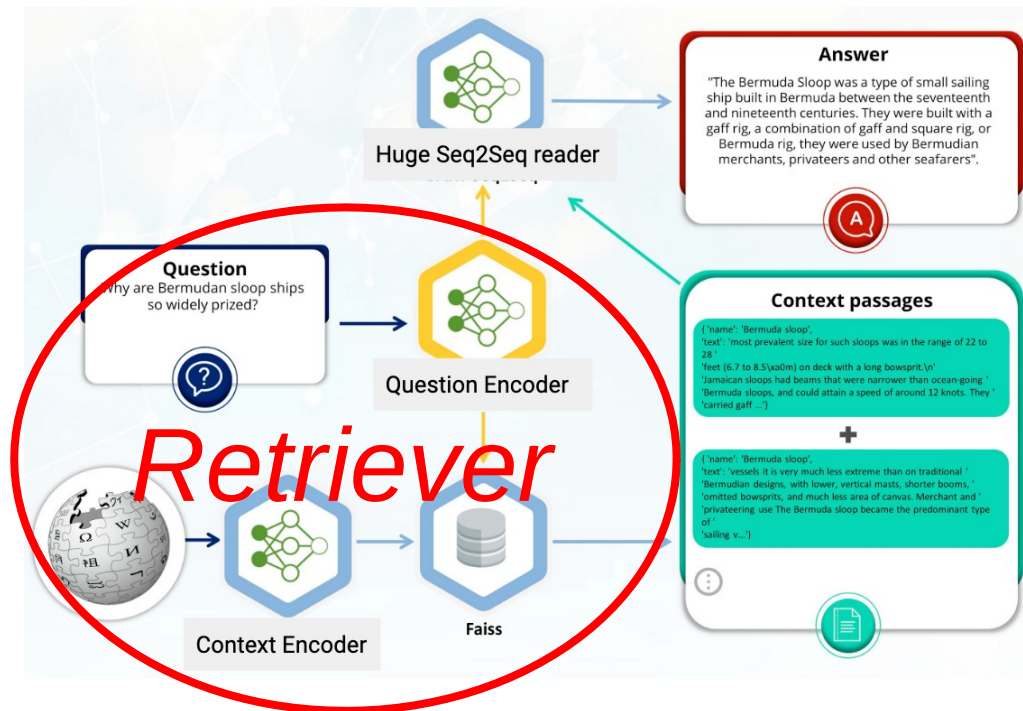


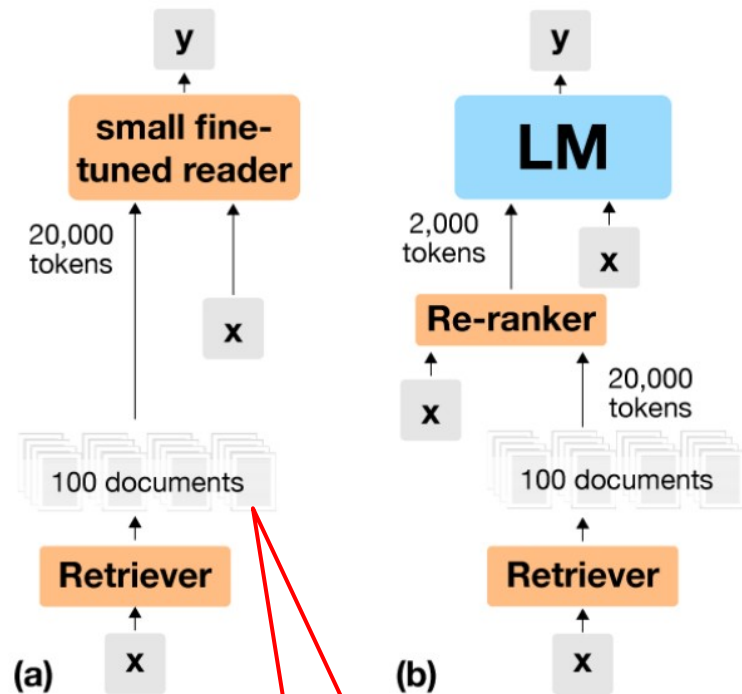
(b)



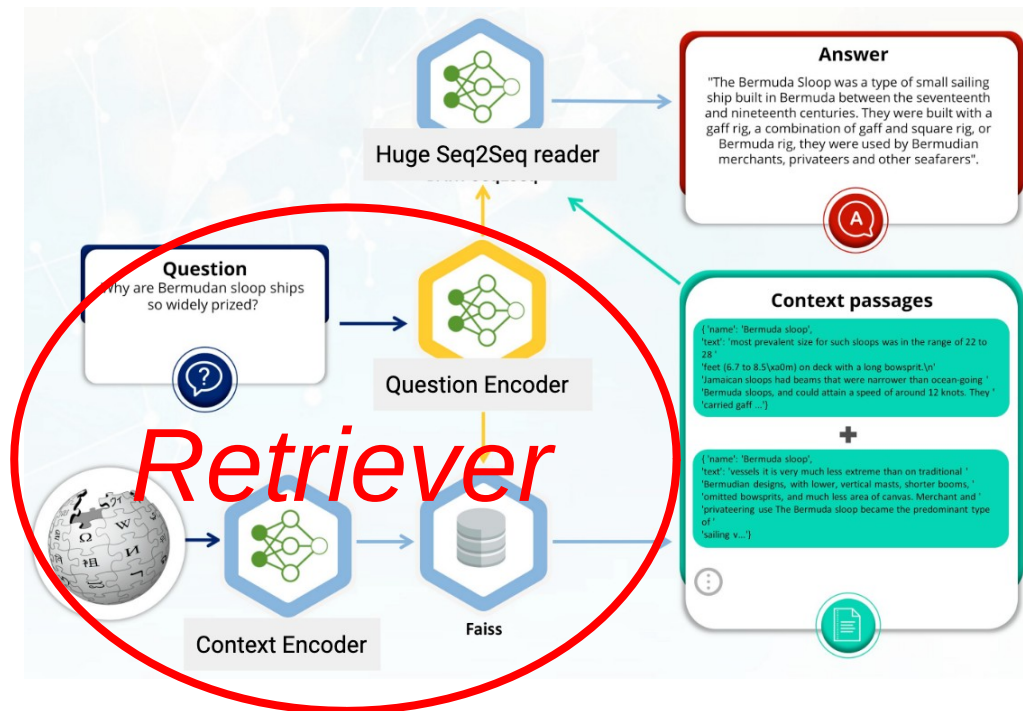


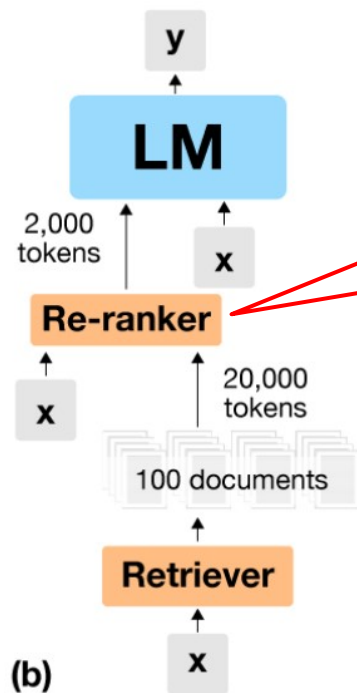
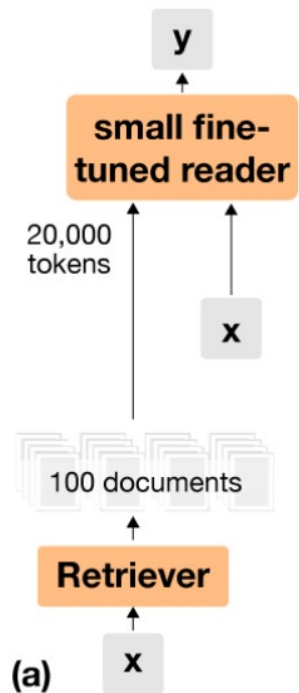
Used a successful base retriever: Spider-NQ + BM25



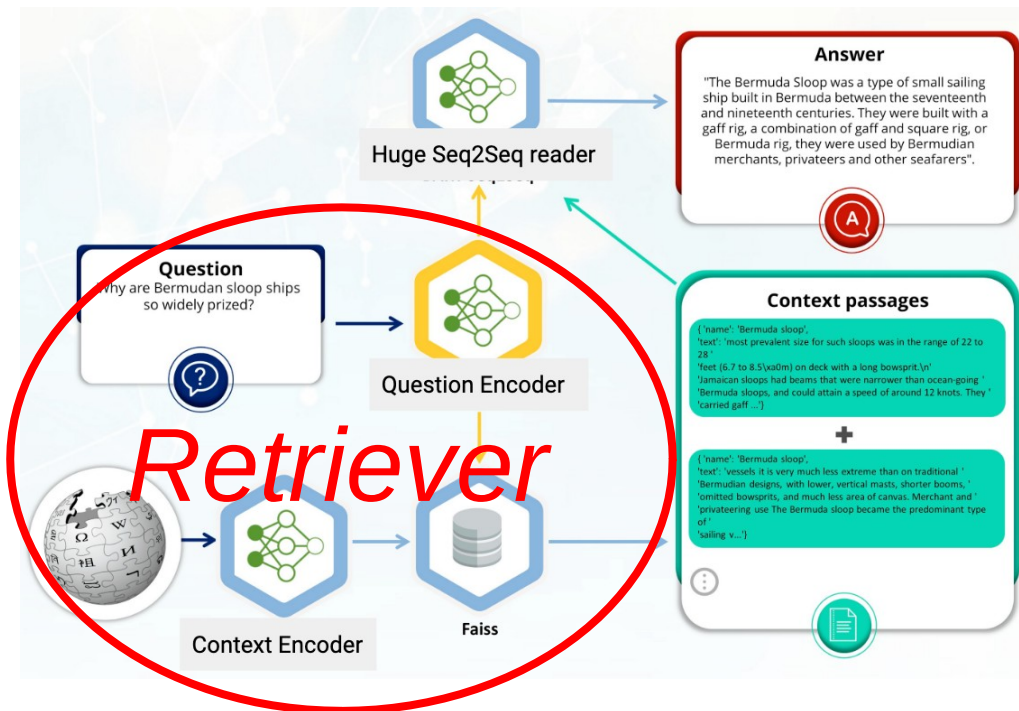


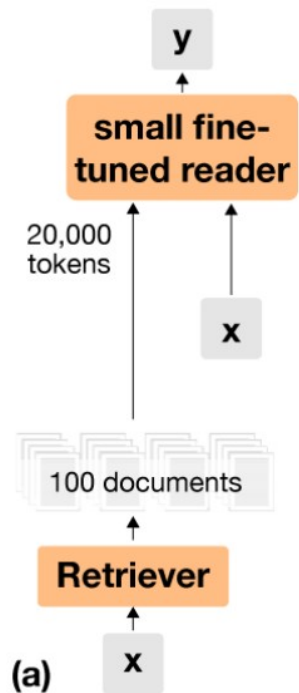
100 docs are too long for an off-the-shelf huge LM



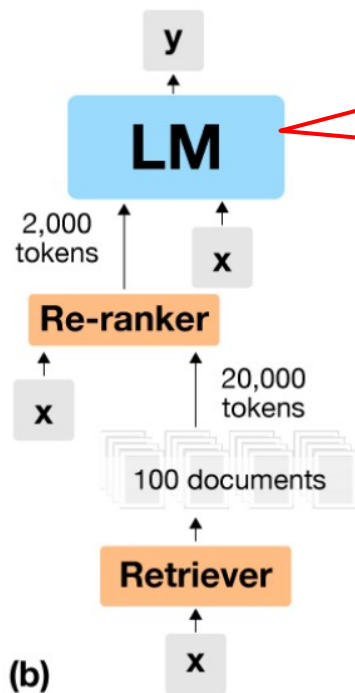


Train a *small* re-ranker (BERT-base, 110M params) to fit the best docs into LM's input length



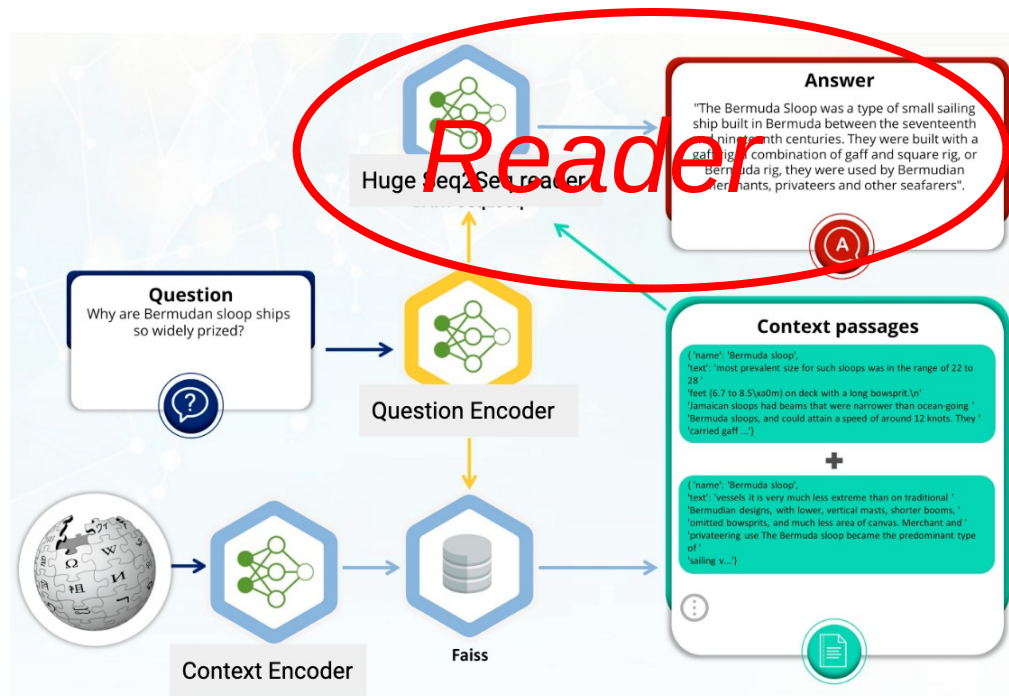


(a)

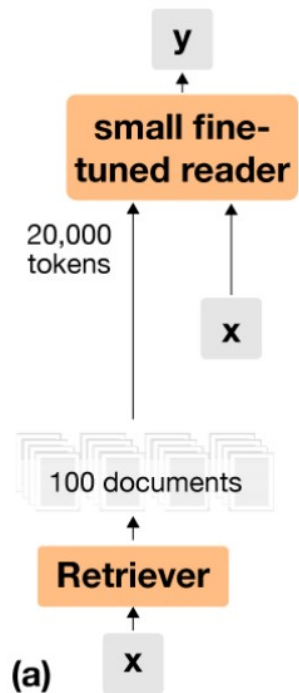


(b)

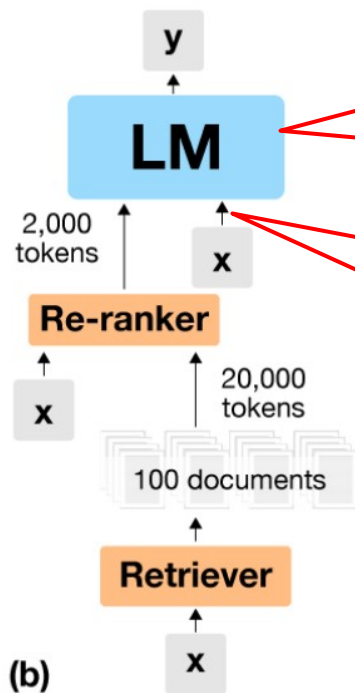
Use a huge pretrained Seq2Seq LM, Jurassic-1 (7B/17B params)







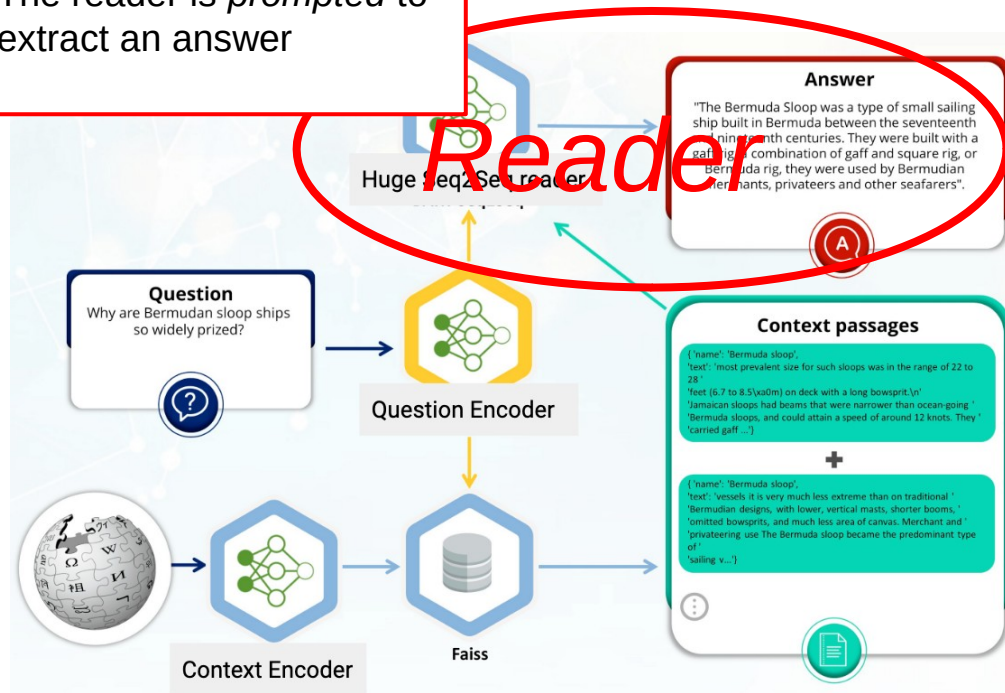
(a)



(b)

Use a huge pretrained Seq2Seq LM, Jurassic-1 (7B/17B params)

The reader is *prompted* to extract an answer





# Does the re-ranker help?

<b>Passage score</b>	<b>Reader</b>	<b>Retriever</b>	<b>Recall @ J1 input</b>	<b>Avg. #docs</b>	<b>Dev EM</b>
Retriever	J1-Large-7B	Spider+BM25	81.4	17	49.5
Re-ranker	J1-Large-7B	Spider+BM25	83.2	17	50.8

# Exact Match Performance

Model	Reader	Retriever	Test EM
EMDR <sup>2</sup> (Singh et al., 2021)	Fine-tuned T5-Base	EMDR <sup>2</sup>	52.5
FiD-Distill (Izacard & Grave, 2020b)	Fine-tuned T5-Large	Distilled DPR	<b>53.7</b>
Frozen LM reader (Ours)	J1-Large-7B	Spider+BM25	51.9
Frozen LM reader (Ours)	J1-Grande-17B	Spider+BM25	<b>53.7</b>

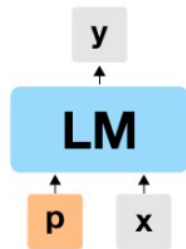
# AI21's Jurassic-X

LFQA function

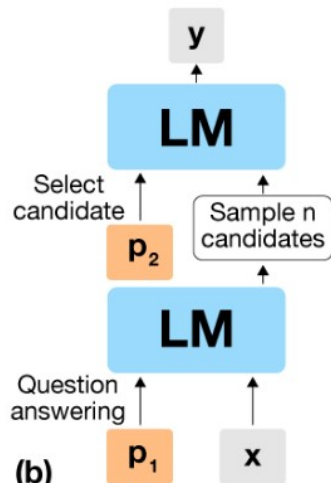
RECURSIVELY APPLYING A FROZEN LM

In *closed book* LFQA

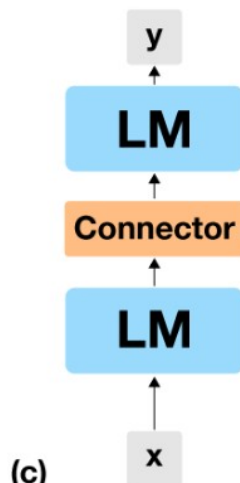




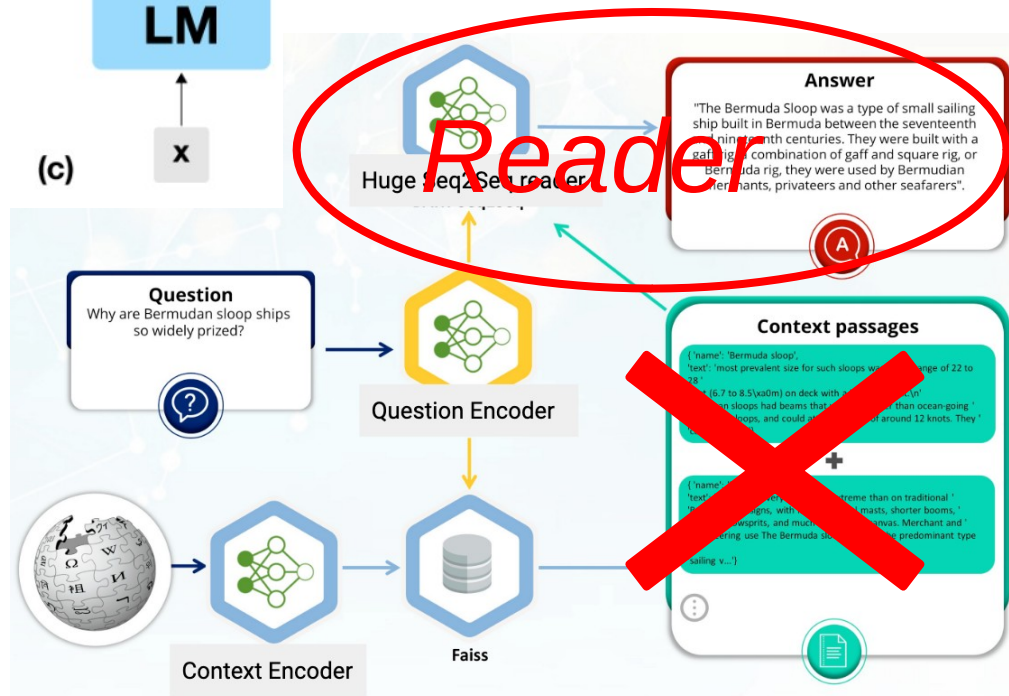
(a)

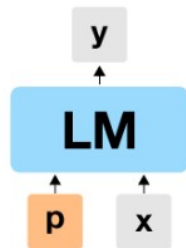


(b)

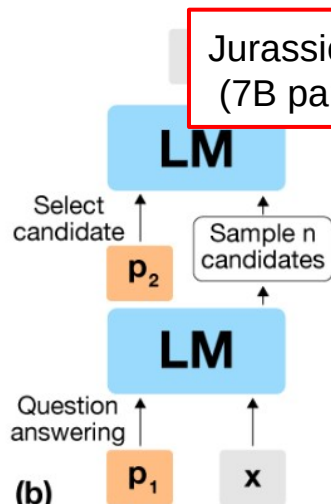


(c)



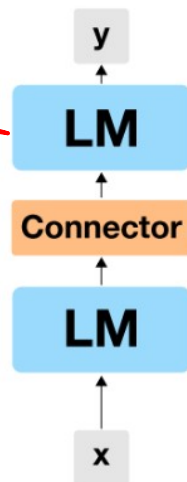


(a)

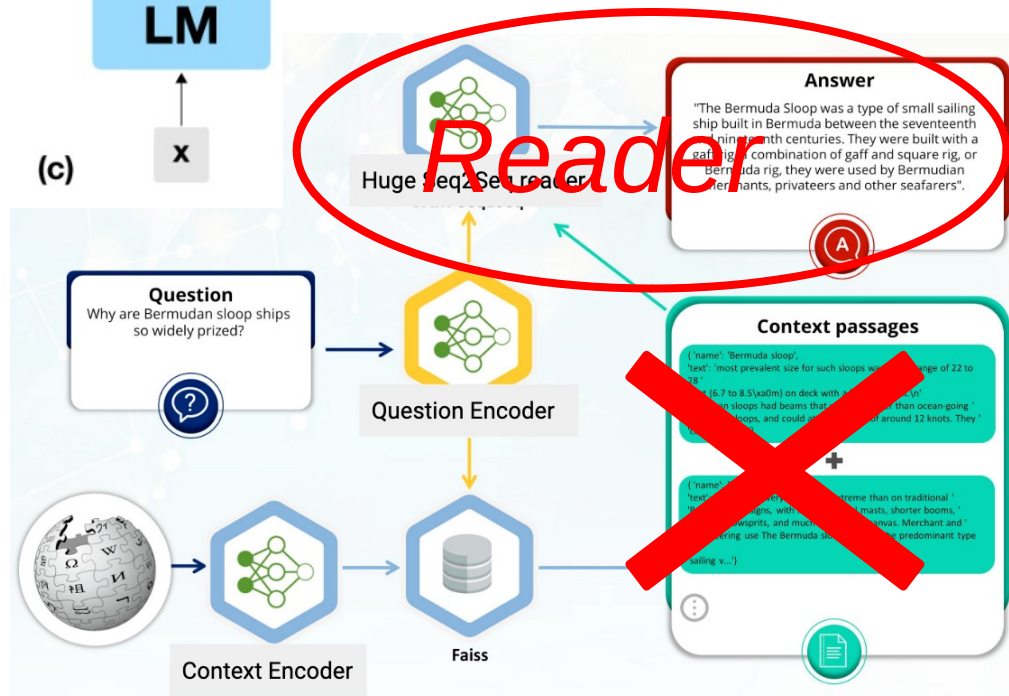


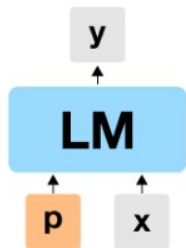
(b)

Jurassic-1 Large  
(7B params)

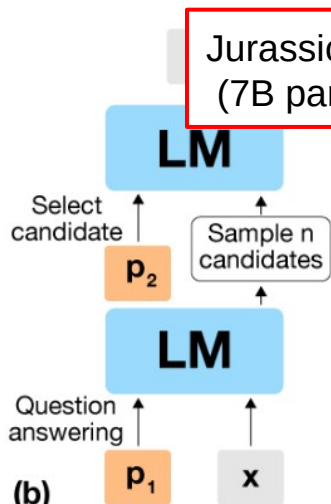


(c)

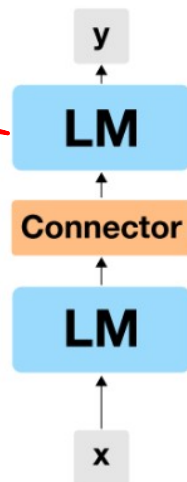




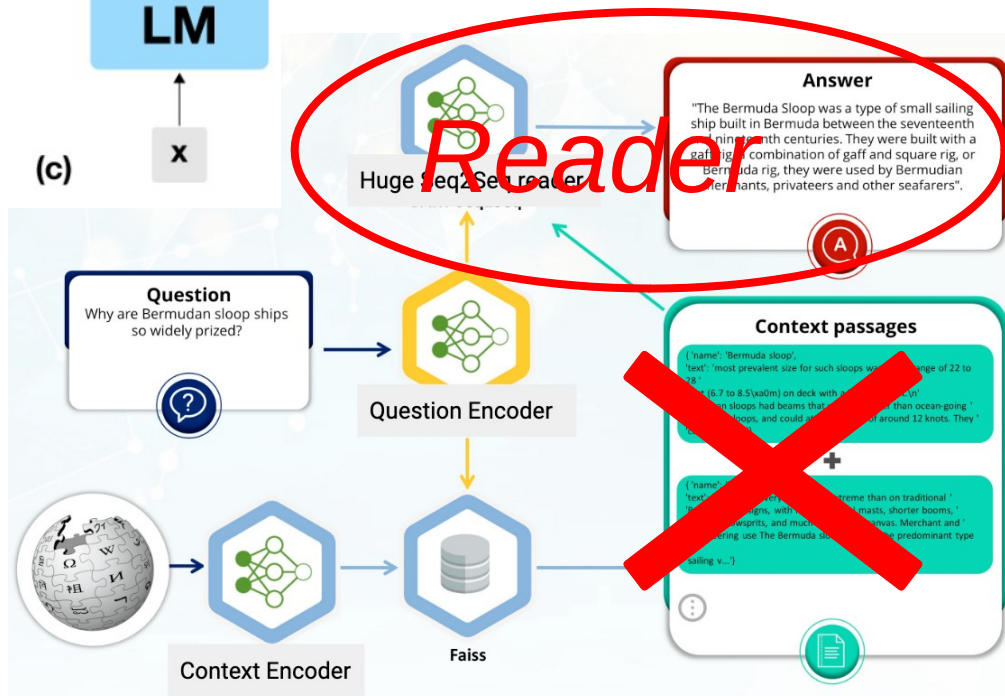
(a)

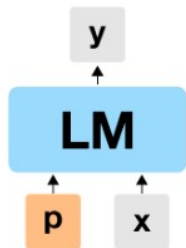


(b)

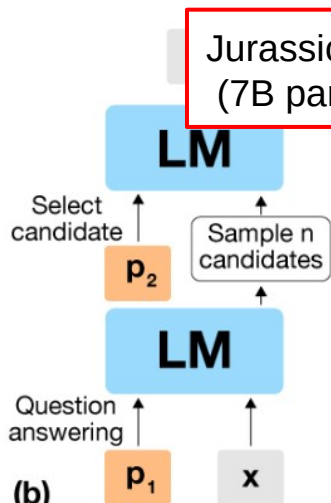


(c)



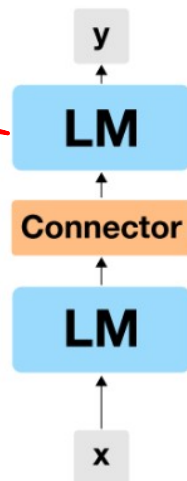


(a)



(b)

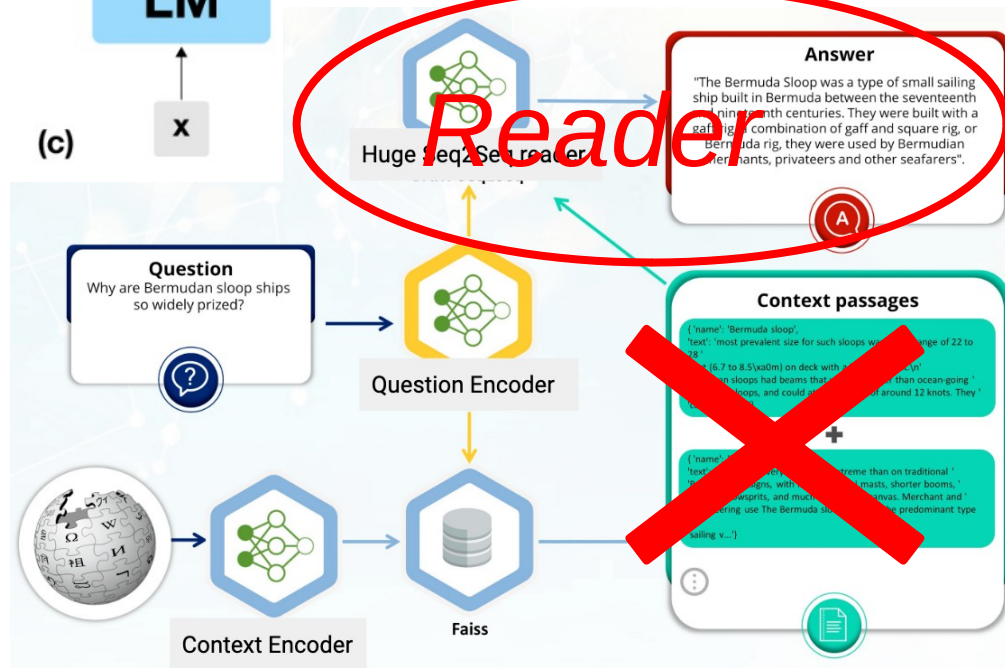
Jurassic-1 Large  
(7B params)

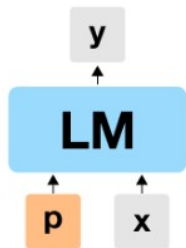


(c)

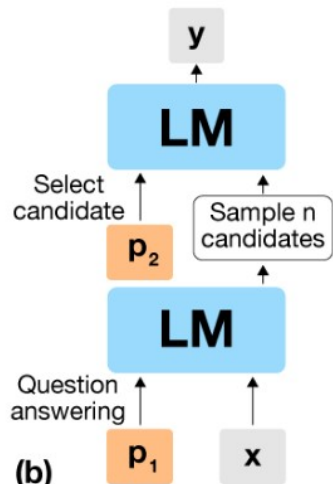
A Seq2Seq LM both reads and writes text!

It's already pre-trained for multiple tasks!

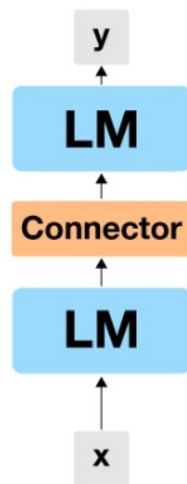




(a)

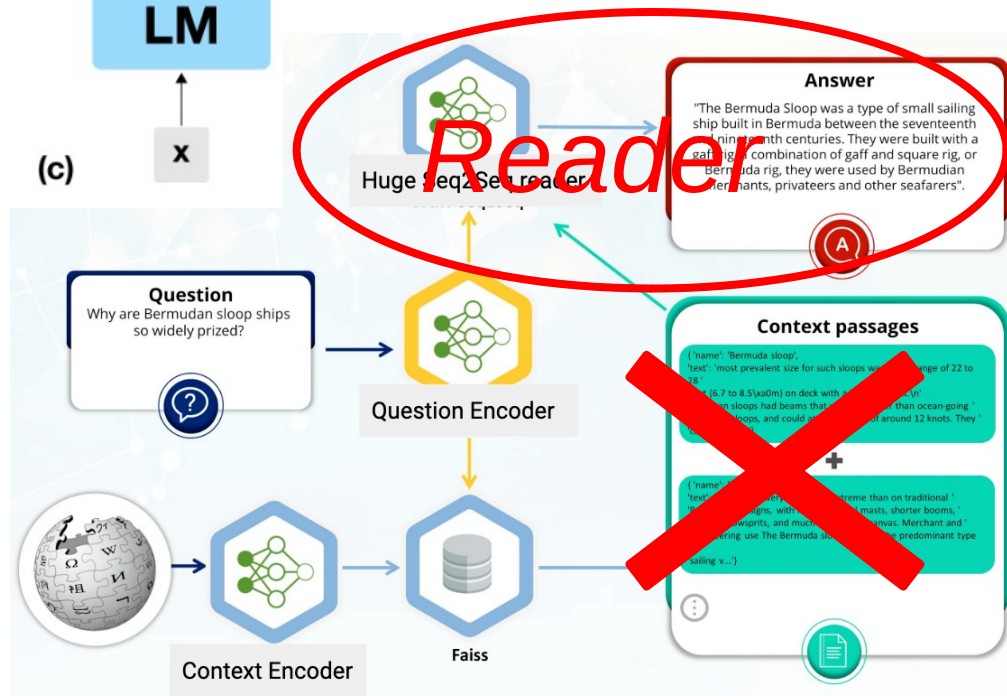


(b)



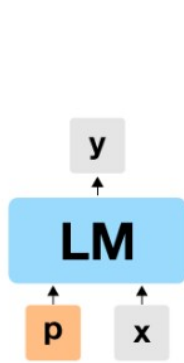
(c)

Textual approach,  
text->LM->text->LM->text

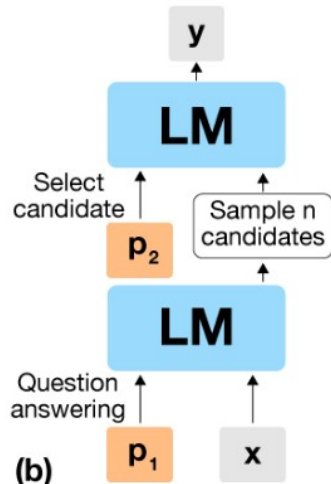








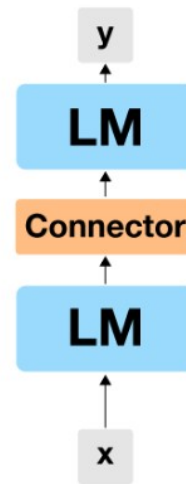
(a)



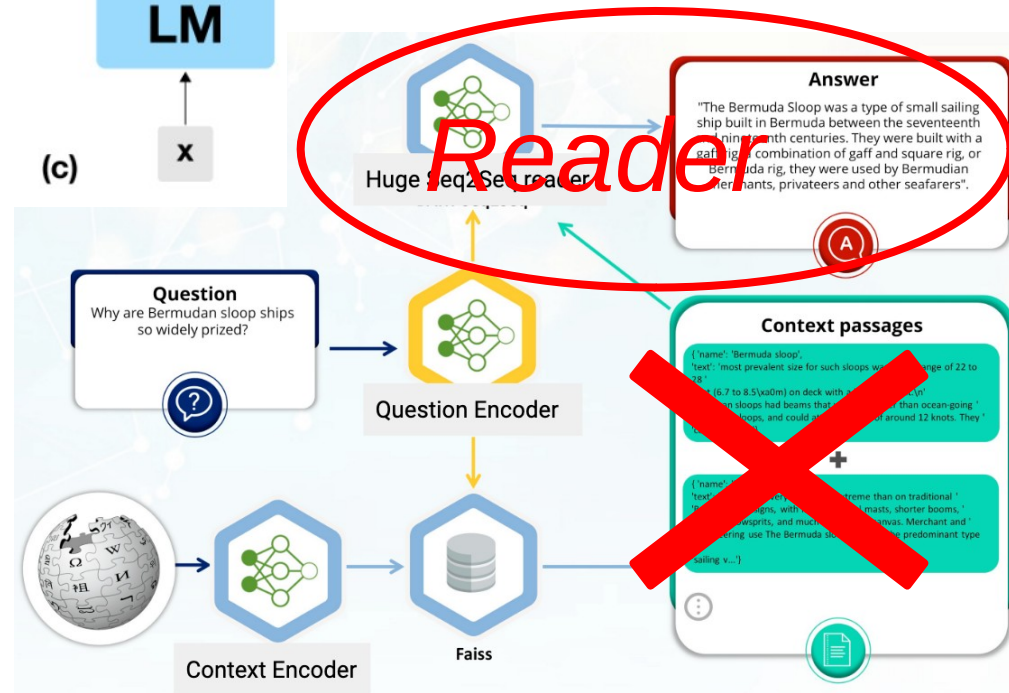
(b)

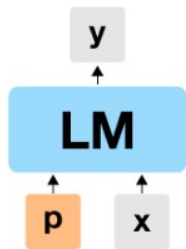
### Disadvantages of textual LM recursion

1. Slow
2. Text representation loses some info
3. p<sub>1</sub> could be trained for a more general task
4. Too weak (few params)

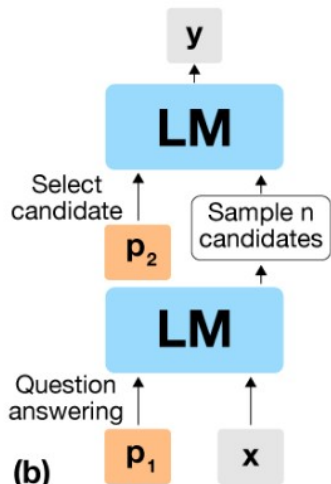


(c)

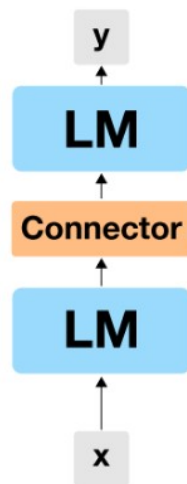




(a)

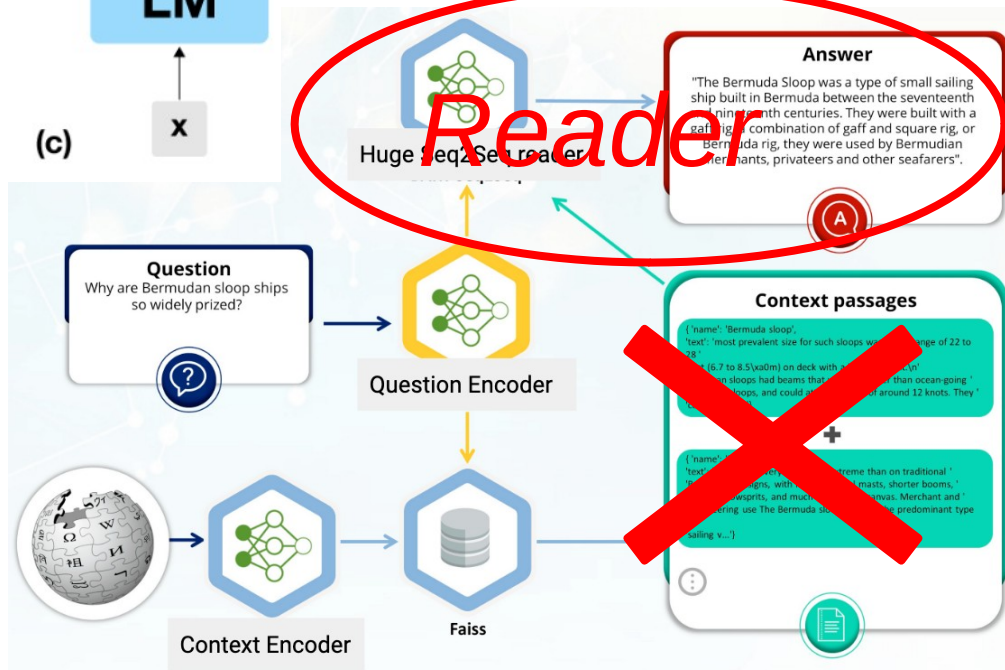


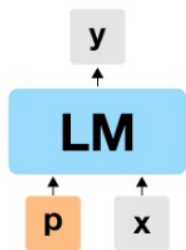
(b)



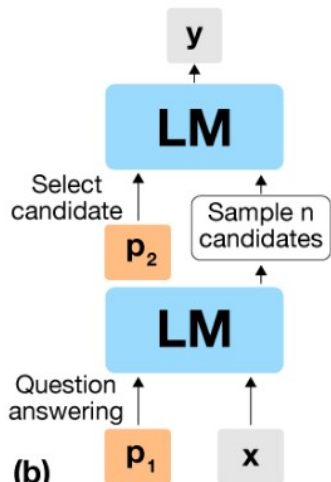
(c)

Neural approach,  
text->LM->logits->LM->text

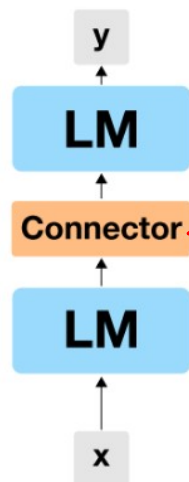




(a)



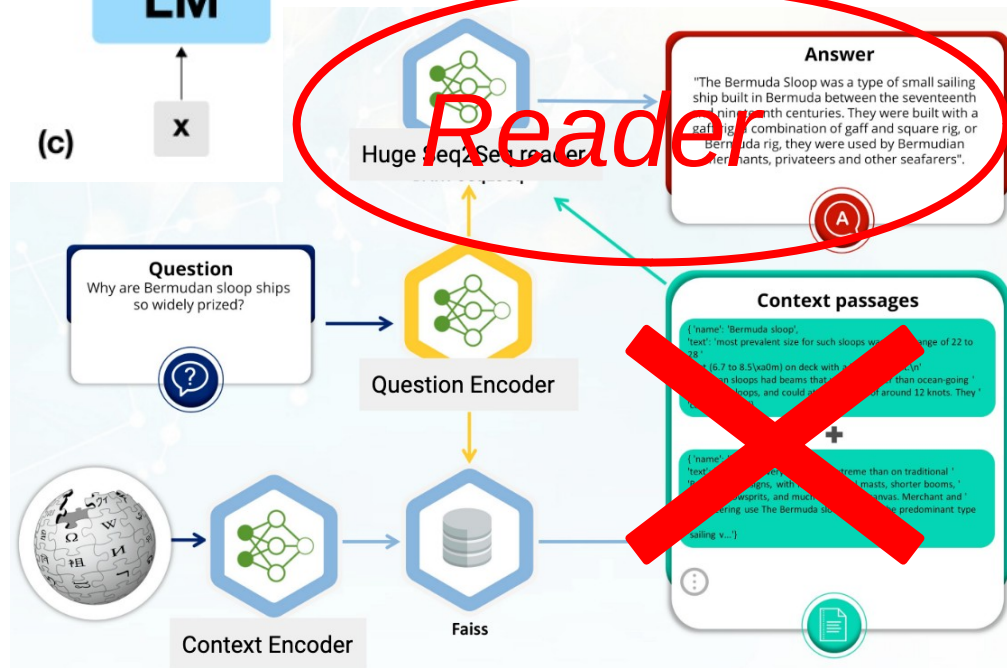
(b)



(c)

Neural approach,  
text  $\rightarrow$  LM  $\rightarrow$  logits  $\rightarrow$  LM  $\rightarrow$  text

1 or 2 transformer layers, plugged  
into the 1st transformer layer of the  
second LM (w/ 32 layers)



# Does *neural* LM recursion help?

LM Passes	LM size	Model	Connector Layers	Pretrain Init?	EM
1	7B	Prompt-tuned	–	No	17.1
1	7B	Prompt-tuned	–	Yes	17.0
1	7B	LM–Connector	2	Yes	17.3
1	7B	Connector–LM	2	Yes	18.6
2	7B	LM–Connector–LM	2	No	18.7
2	7B	LM–Connector–LM	1	Yes	19.6
2	7B	LM–Connector–LM	2	Yes	<b>20.8</b>
1	17B	Prompt-tuned	–	No	22.1

Can multiple passes through a small LM could serve as an alternative to a larger model?

# Does *any* LM recursion help?

LM Passes	LM Size	Method	EM (dev)	EM (test)
1	7B	Prompt-tuned	19.8	21.6
2	7B	Textual-recursive (n=16)	22.2	23.4
2	7B	Neural-recursive (2 pretrained Connector layers)	<b>25.6</b>	<b>26.0</b>

Now instead of pre-training, fine tuning and deploying more and bigger LMs, we can re-run 1 LM and get more performance :)

# Recap

## HF ELI5 open QA

Train ranker(s) with *contrastive learning* to embed questions and docs into FAISS

Each question is paired with 10 most similar docs

Fine tune a large reader, to generate answers

## Jurassic-X open QA

Train a small reranker

Huge frozen reader

Prompt tuning

## Jurassic-X closed

No document store

Huge frozen reader

LM Recursion, textual or neural

WHO WON? WHO'S NEXT?

Demo

S!

LET THE PEOPLE  
OF THE WORLD  
YOU DECIDE!



# Links

Explain Anything Like I'm Five:

A Model for Open Domain Long Form Question Answering ([notebook](#))

[ELI5: Long Form Question Answering \(Fan et al., 2019\)](#)

[Long-Form QA beyond ELI5: an updated dataset and approach](#)

[LFOA demo, by HuggingFace](#)

[Long-Form QA beyond ELI5: an updated dataset and approach \(Blagojevic, towardsdatascience\)](#)

[Jurassic-X: Crossing the Neuro-Symbolic Chasm with the MRKL System](#)

# Human LFQA

tnx!