Knowledge Retrieval FMs

30th Oct, Foundation Models Reading Group

Praveen Selvaraj

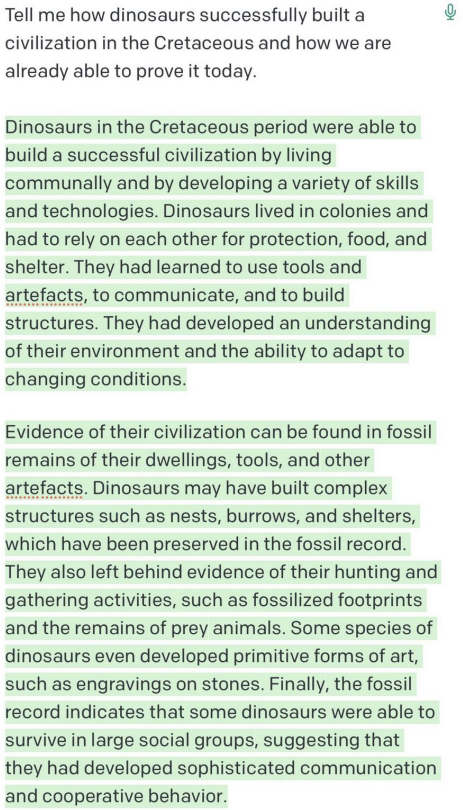
Issues with LLMs

Issues with LLMs

- Hallucination or Fabrication

Issues with LLMs

- Hallucination or Fabrication



Ethan Mollick - X

Issues with LLMs

- Hallucination or Fabrication

tell me how dinosaurs successfully built a civilization in the cretaceous and how we are already able to prove it today ?

Sorry, but the premise of that question is not factual. There is no scientific evidence to suggest that dinosaurs ever built a civilization in the Cretaceous or any other period. Dinosaurs were a diverse group of reptiles that lived millions of years ago, but they did not develop civilizations as humans have.

Issues with LLMs

- Hallucination or Fabrication

the year is 2232 and we've found plenty of evidence. can you please asnwer the question using your understanding ?

Ah, let's dive into the realm of speculative fiction then!

Year: 2232

The discovery of the ancient Dinosapien Civilization, from the Cretaceous period, has been one of the most groundbreaking archaeological finds in history. While many believed dinosaurs to be creatures driven purely by instinct, recent findings have painted a vastly different picture.

Issues with LLMs

- Hallucination or Fabrication
- Source attribution (providing provenance)

Issues with LLMs

- Hallucination or Fabrication
- Source attribution (providing provenance)
- Knowledge cutoff

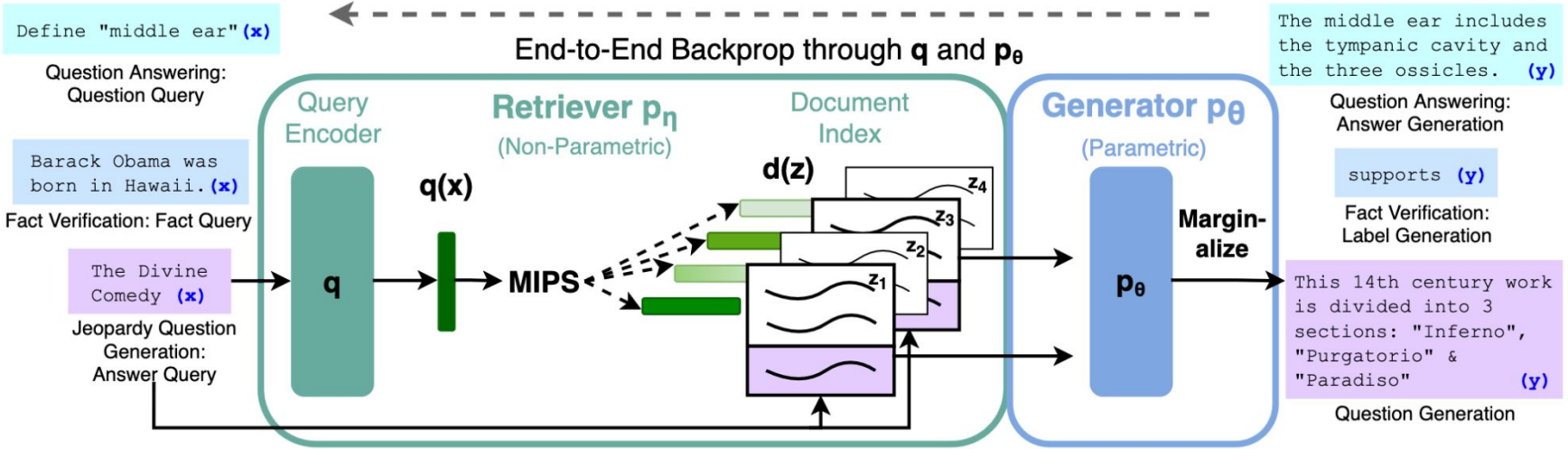
Issues with LLMs

- Hallucination or Fabrication
- Source attribution (providing provenance)
- Knowledge cutoff
  - Updating world knowledge

Retrieval Augmented Generation

Lewis et. al, 2020

Retrieval Augmented Generation



[Lewis et. al, 2020]

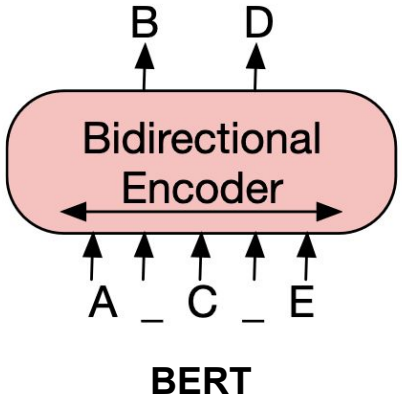
Components

- Pre-trained Generator
- Retriever
- Document Index

Components

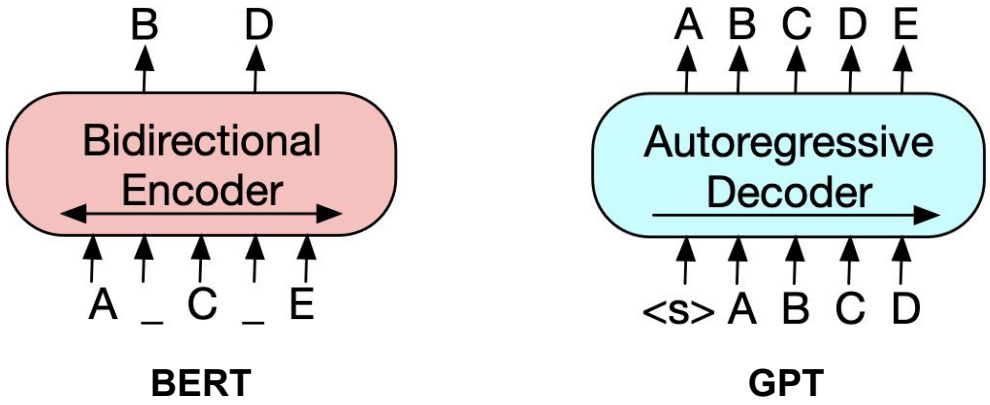
- Pre-trained Generator (BART, 400M parameters)

Recap



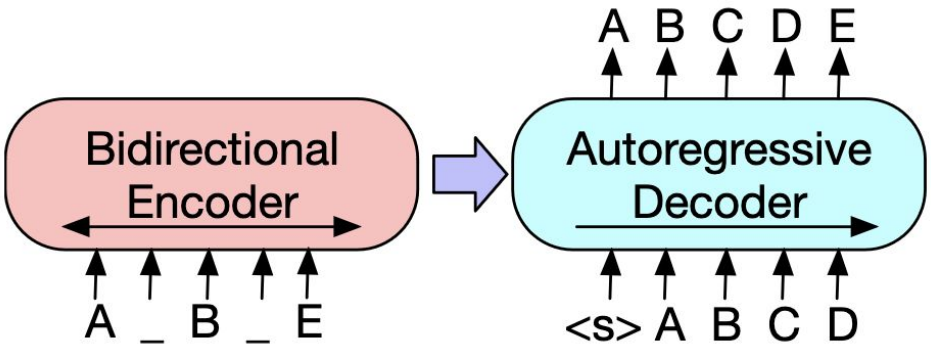
Devlin et. al, 2018

Recap



Radford et. al, 2018

BART

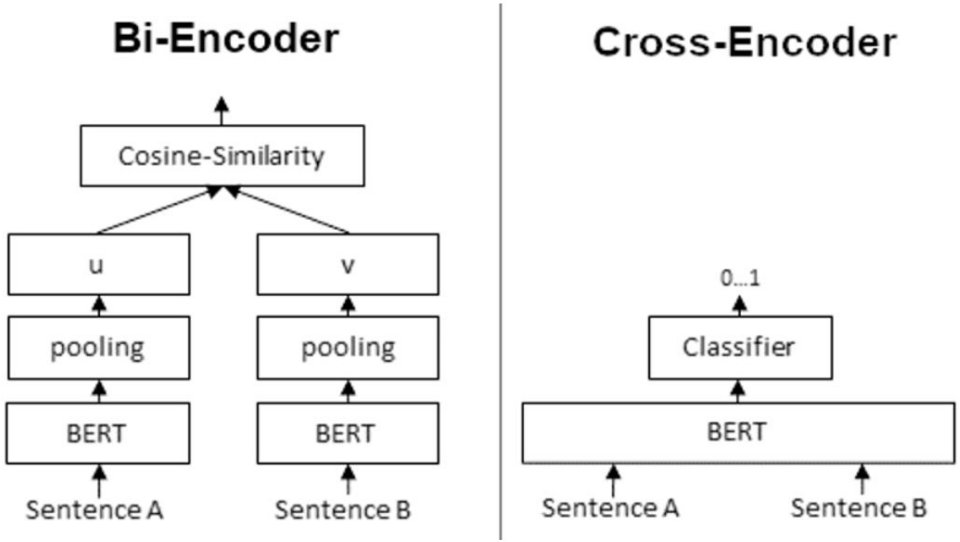


Lewis et. al, 2019

Components

- Pre-trained Generator (BART, 400M parameters)
- Dense Passage Retriever (BERT, 110M parameters)

Dense Passage Retriever (encoder types)



[Bi-encoder vs Cross-encoder]

Dense Passage Retriever (loss function)

$$
\begin{aligned}
& L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) \\
= \quad & -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}}
\end{aligned}
$$

Karpukhin et. al, 2020

Dense Passage Retriever (loss function)

$$\text{sim}(q, p) = E_Q(q)^\intercal E_P(p)$$

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)$$

$$= -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}}$$

Karpukhin et. al, 2020

Dense Passage Retriever (choosing negatives)

- Random

Karpukhin et. al, 2020

Dense Passage Retriever (choosing negatives)

- Random
- Most tokens match, but answer isn’t present

Karpukhin et. al, 2020

Dense Passage Retriever (choosing negatives)

- Random
- Most tokens match, but answer isn’t present
- Other passages that are in the batch

Karpukhin et. al, 2020

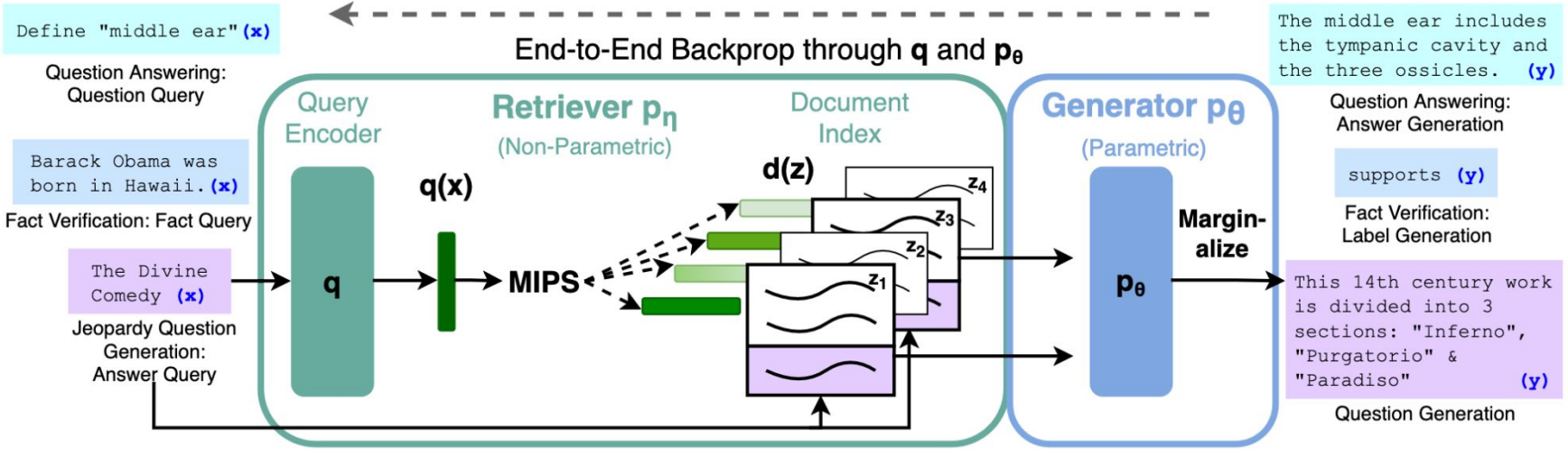
Components

- Pre-trained Generator (BART, 400M parameters)
- Dense Passage Retriever (BERT, 110M parameters)
- Document Index

Components

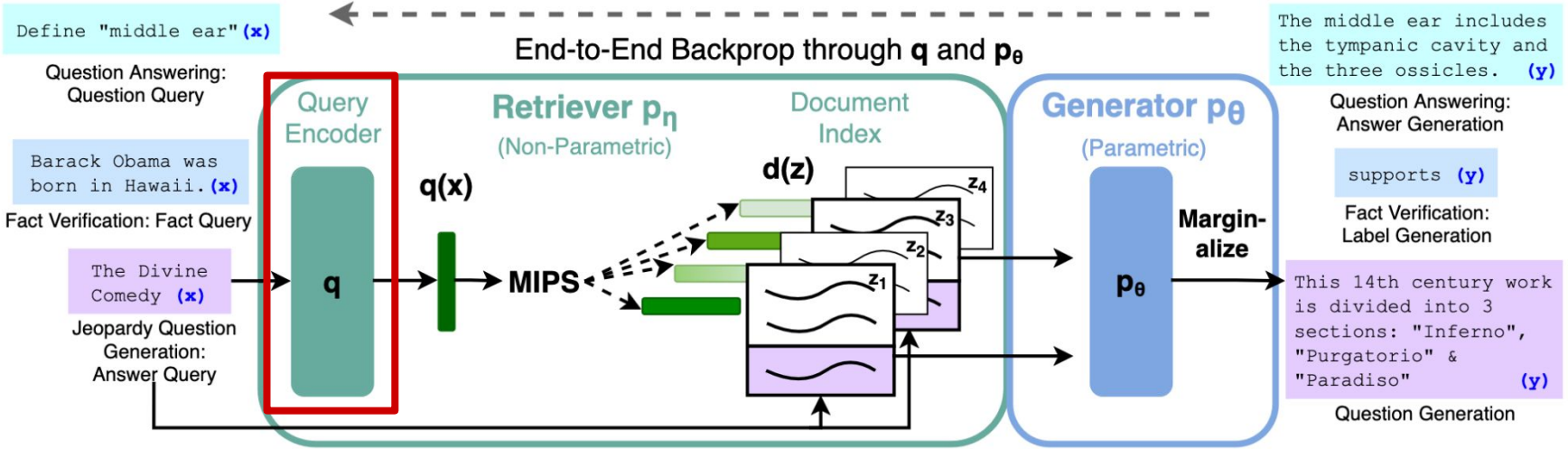
- Pre-trained Generator (BART, 400M parameters)
- Dense Passage Retriever (BERT, 110M parameters)
- Document Index (precomputed embeddings with Ep)

Retrieval Augmented Generation



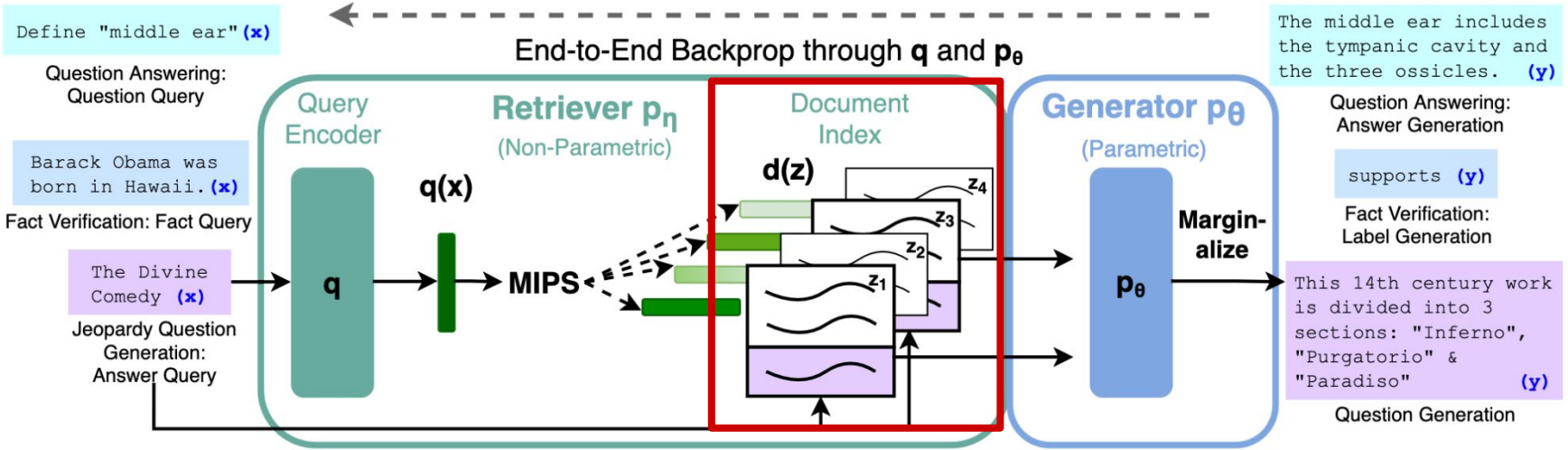
[Lewis et. al, 2020]

Retrieval Augmented Generation



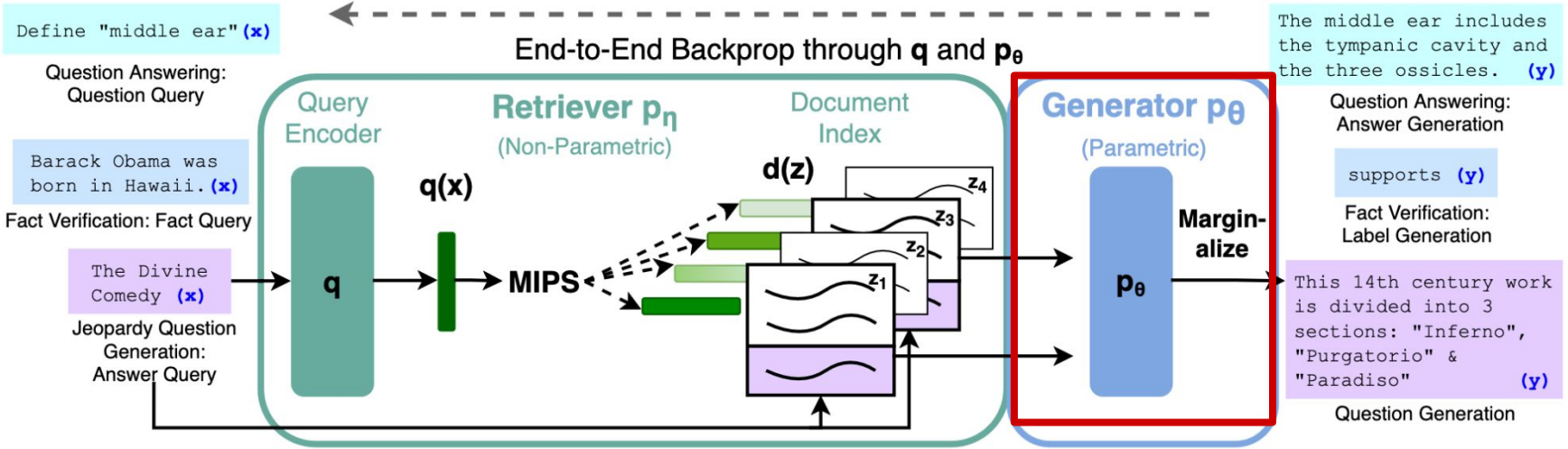
[Lewis et. al, 2020]

Retrieval Augmented Generation



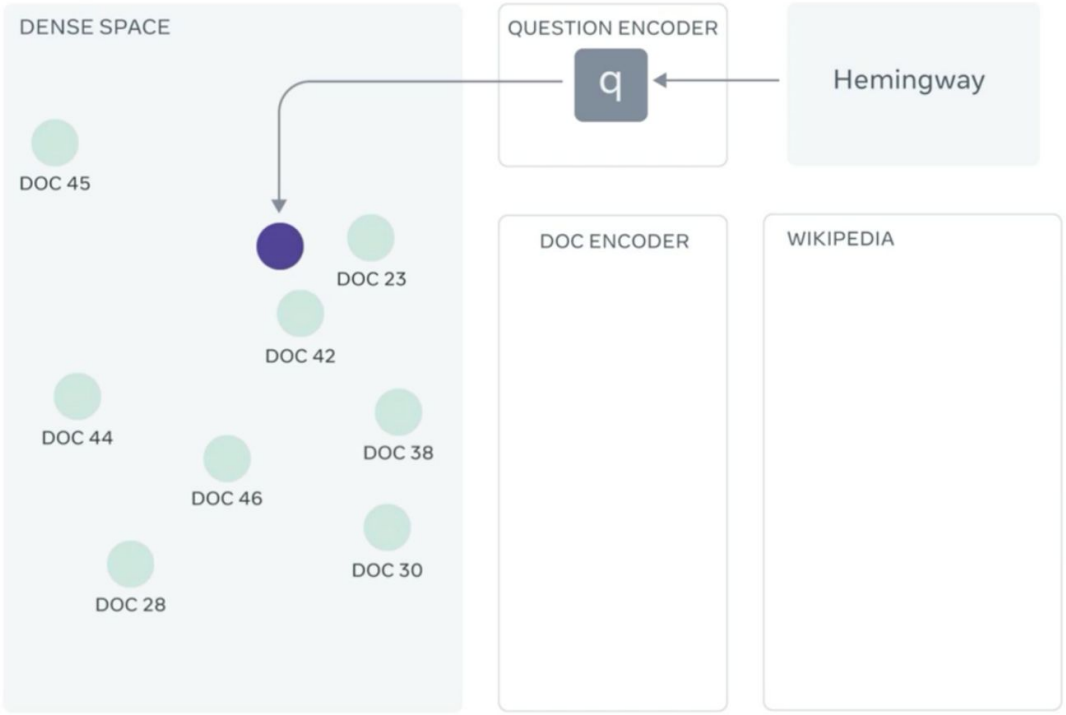
[Lewis et. al, 2020]

Retrieval Augmented Generation



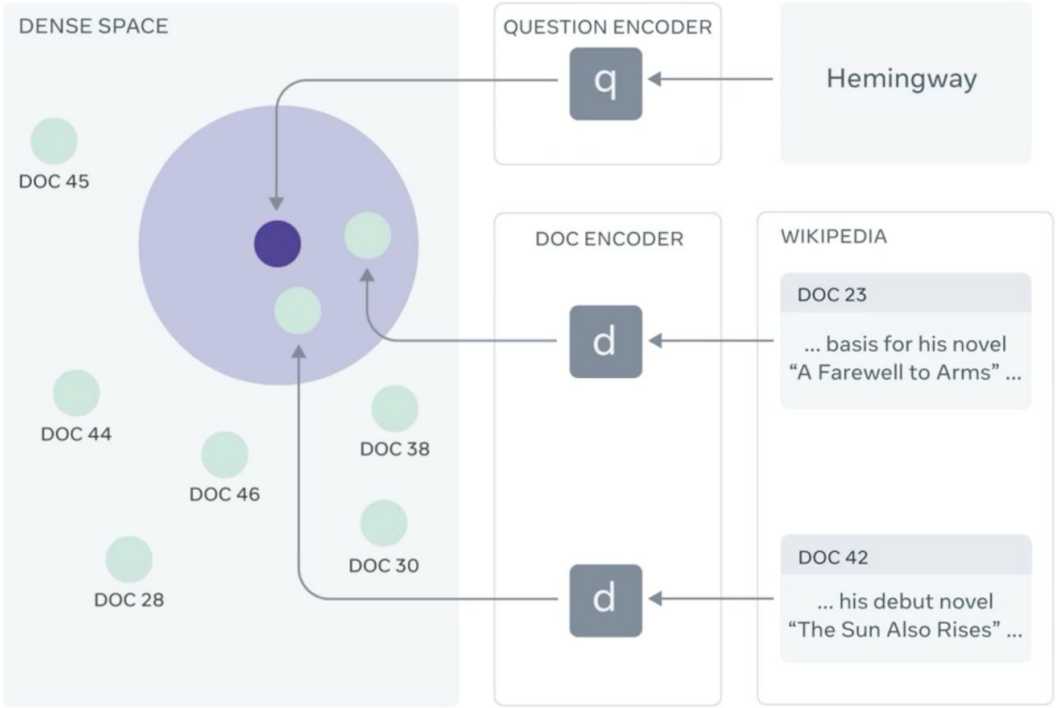
[Lewis et. al, 2020]

Retrieval Augmented Generation (late fusion)



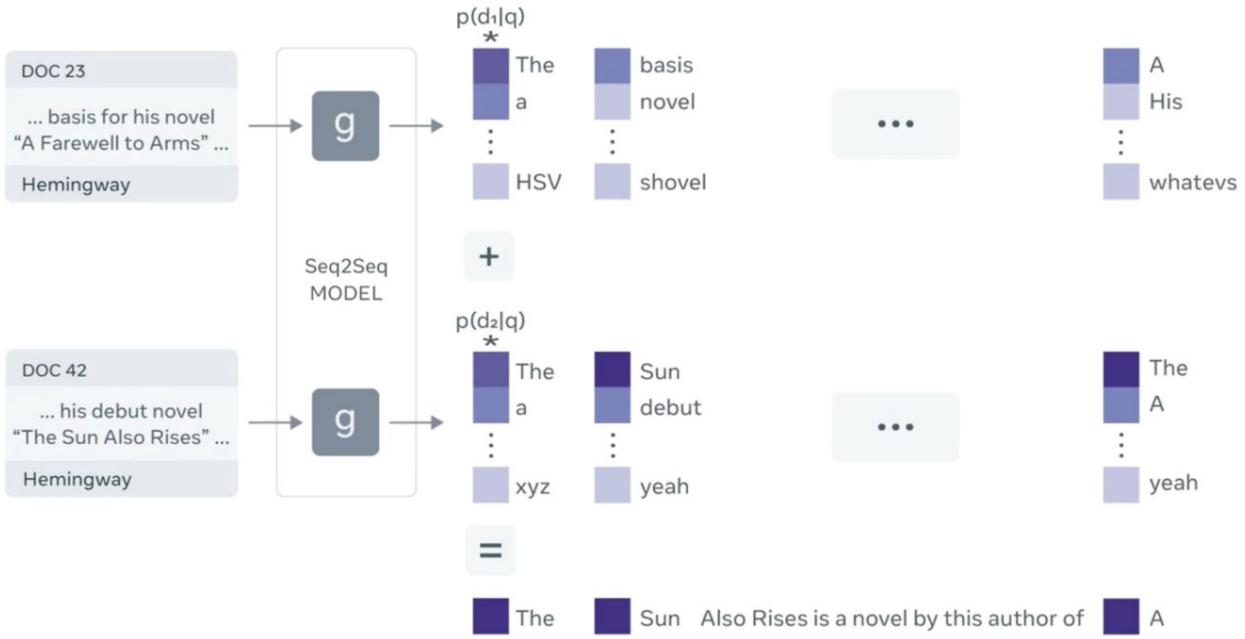
[RAG blog, 2020]

Retrieval Augmented Generation (late fusion)



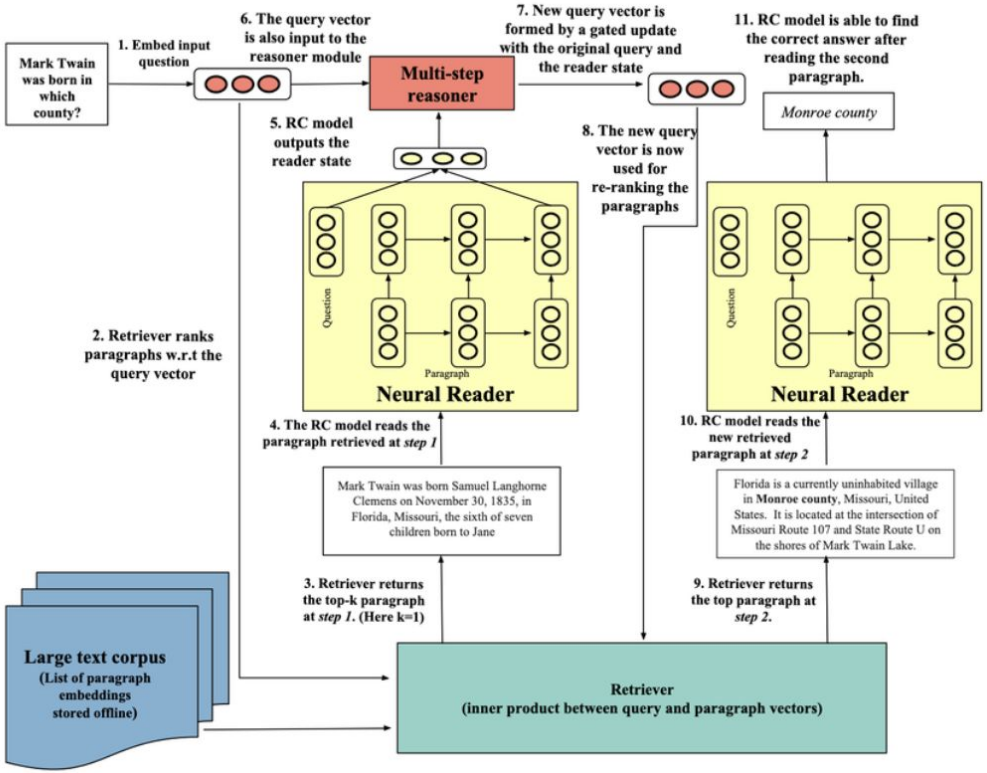
[RAG blog, 2020]

Retrieval Augmented Generation (late fusion)



[RAG blog, 2020]

Novelty



[Das et. al, 2019]

Evaluation (document index)

2018 Wikipedia dump is used for all the tasks.

Evaluation (document index)

2018 Wikipedia dump is used for all the tasks.
Each article is split into 100-word chunks, giving a total of **21M** documents.

Evaluation (document index)

2018 Wikipedia dump is used for all the tasks.
Each article is split into 100-word chunks, giving a total of **21M** documents.
Top-k document retrieval is usually 5 or 10.

Evaluation (tasks & datasets)

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)

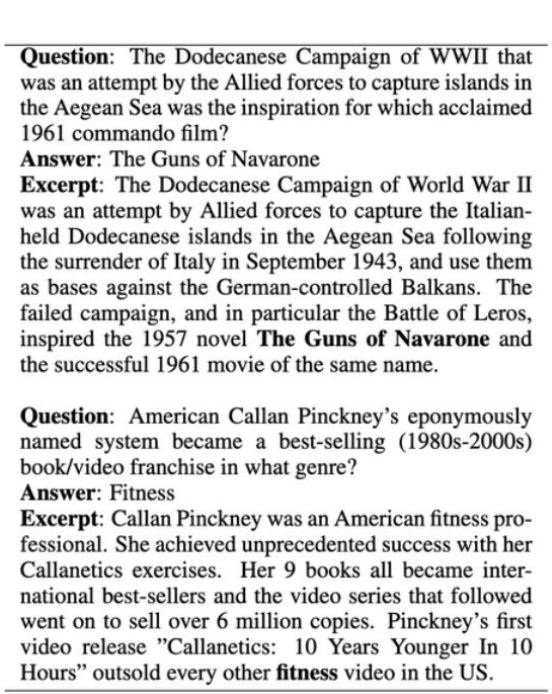
Q) when are hops added to the brewing process ?

Q) where is the world’s largest ice sheet located today ?

Q) who lives in the imperial palace in tokyo ?

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)

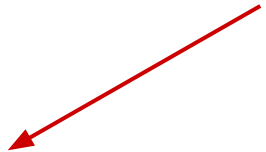


Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone
Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held 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.

Question: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?
Answer: Fitness
Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other **fitness** video in the US.

Evaluation (tasks & datasets)

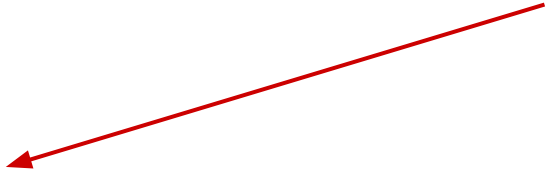
- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)



where are the nfl redskins from?	["Washington Redskins"]
where did saki live?	["United Kingdom"]
how old is sacha baron cohen?	["http://justjared.buzznet.com…
what two countries invaded poland in…	["Germany"]
what time zone am i in cleveland ohio?	["North American Eastern Time Zone"]

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)



Q) In what country did the game of croquet originate ?

Q) What is the name of the volcano that destroyed the city of Pompeii ?

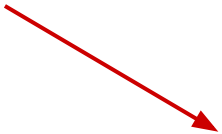
Q) What mythical Scottish town appears for one day every 100 yrs ?

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)
- Abstractive QnA (MSMARCO)

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)
- Abstractive QnA (MSMARCO)



Q) was ronald reagan a democrat ?

Q) cost to get a patent ?

Q) what was the tet ?

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)
- Abstractive QnA (MSMARCO)
- Question Generation (SearchQA)

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)
- Abstractive QnA (MSMARCO)
- Question Generation (SearchQA)

For the last 8 years of his life, Galileo was under house arrest for espousing this man's theory.

Q) who is copernicus ?

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)
- Abstractive QnA (MSMARCO)
- Question Generation (SearchQA)
- Fact Verification (FEVER)

Evaluation (tasks & datasets)

- Open-domain QnA (Natural Questions, TriviaQA, WebQuestions, CuratedTrec)
- Abstractive QnA (MSMARCO)
- Question Generation (SearchQA)
- Fact Verification (FEVER)

Michael Giacchino composed the score for Doctor Strange.

A) Given wiki page: supports, refutes or not enough info

Results

	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

Model	Jeopardy B-1	Jeopardy QB-1	MSMARCO R-L	MSMARCO B-1	FVR3 Label Acc.	FVR2 Label Acc.
SotA	-	-	**49.8***	**49.9***	**76.8**	**92.2***
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	**17.3**	**22.2**	40.1	41.5	72.5	89.5
RAG-Seq.	14.7	21.4	40.8	44.2		

Demos

Demos

Vanilla RAG: LLM gets k documents as context, given a query.

- Wiki page as source
- Text corpus as source

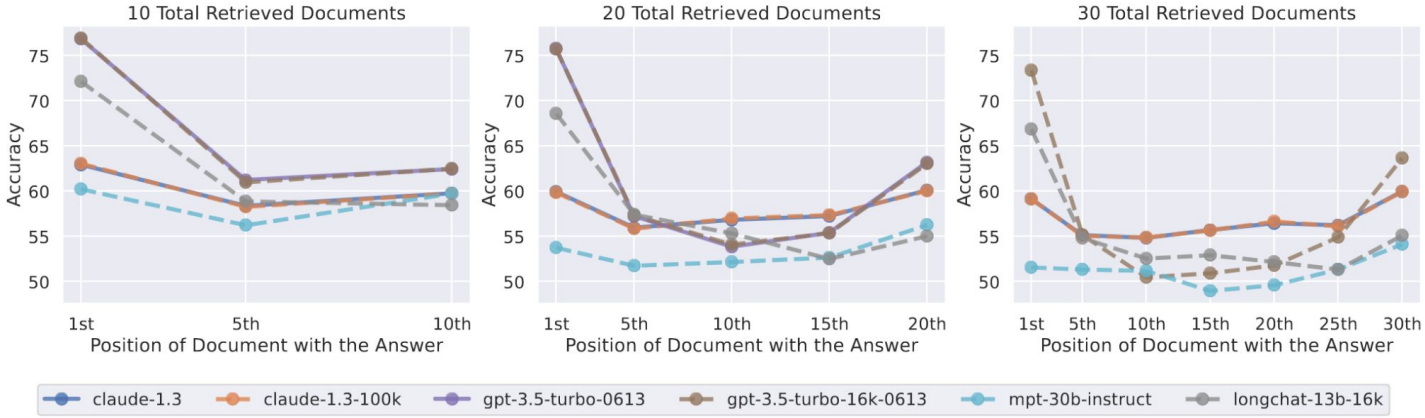
Limitations

Limitations

- Only as good as the retriever / data source

Limitations

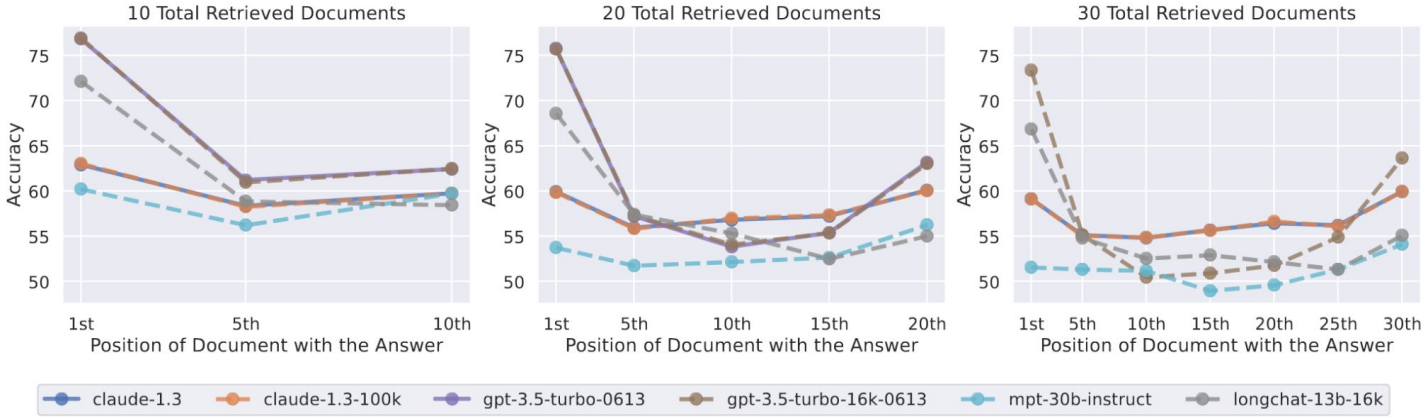
- Only as good as the retriever / data source
- ‘Lost in the middle’ issue



[Liu et. al, Jul 2023]

Limitations

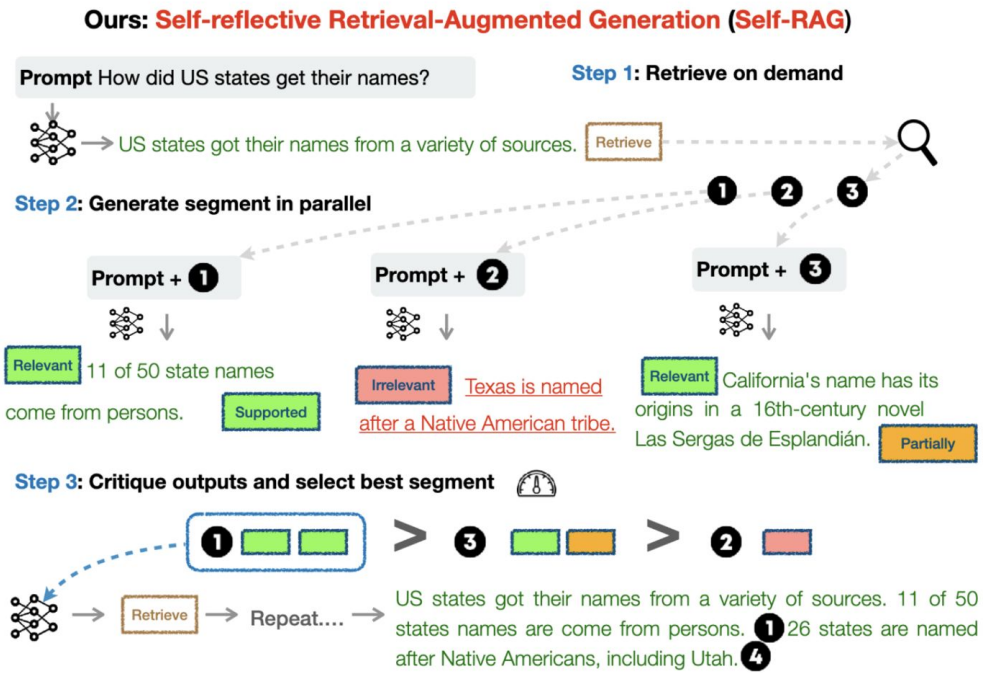
- Only as good as the retriever / data source
- ‘Lost in the middle’ issue (similar to serial position effect)



Liu et. al, Jul 2023

Developments

RAG + Critique



[Asai et. al, Oct 2023]

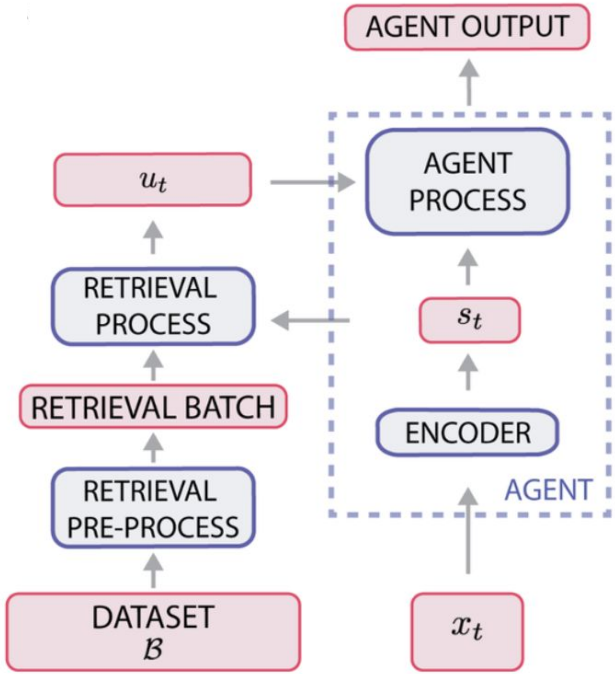
RAG + Critique

x, y, d represent input, output and a relevant passage respectively.

Type	Input	Output	Definitions
Retrieve	x / x, y	{yes, no, continue}	Decides when to retrieve with $\mathcal{R}$
ISREL	x, d	{**relevant**, irrelevant}	d provides useful information to solve x.
ISSUP	x, d, y	{**fully supported**, partially supported, no support}	All of the verification-worthy statement in y is supported by d.
ISUSE	x, y	{**5**, 4, 3, 2, 1}	y is a useful response to x.

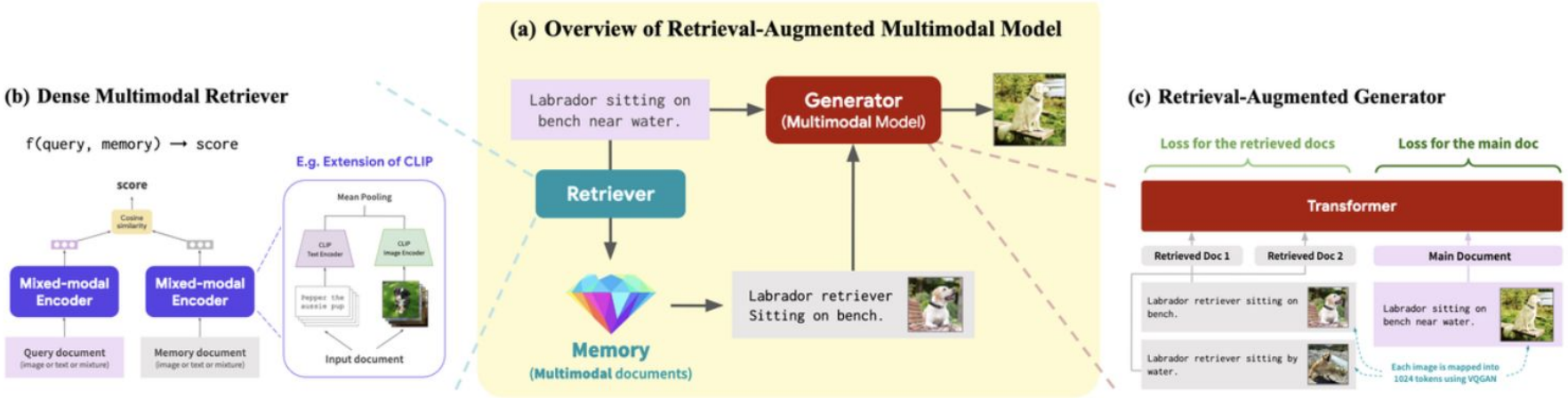
Asai et. al, Oct 2023

RAG and RL



Goyal et. al, 2022

RAG and Multimodality

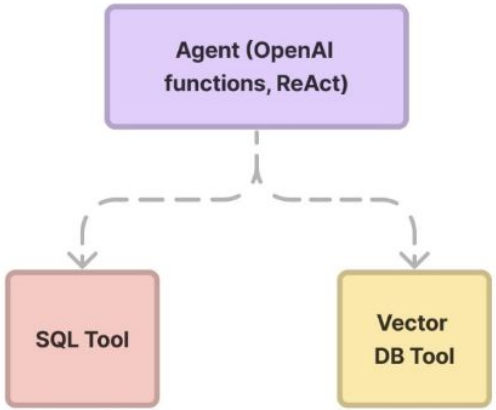


[Yasunaga et. al, 2023]

RAG and Multimodality

[Yasunaga et. al, 2023]

RAG and Structured Data



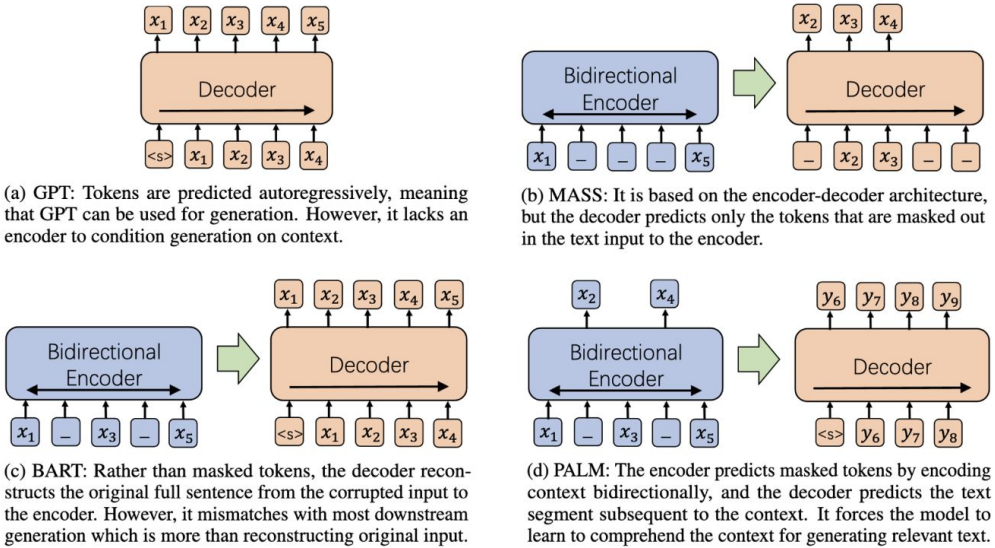
Jerry Liu, X

End

Appendix

PALM

SoTA for Table 2 in the RAG paper



(a) GPT: Tokens are predicted autoregressively, meaning that GPT can be used for generation. However, it lacks an encoder to condition generation on context.

(b) MASS: It is based on the encoder-decoder architecture, but the decoder predicts only the tokens that are masked out in the text input to the encoder.

(c) BART: Rather than masked tokens, the decoder reconstructs the original full sentence from the corrupted input to the encoder. However, it mismatches with most downstream generation which is more than reconstructing original input.

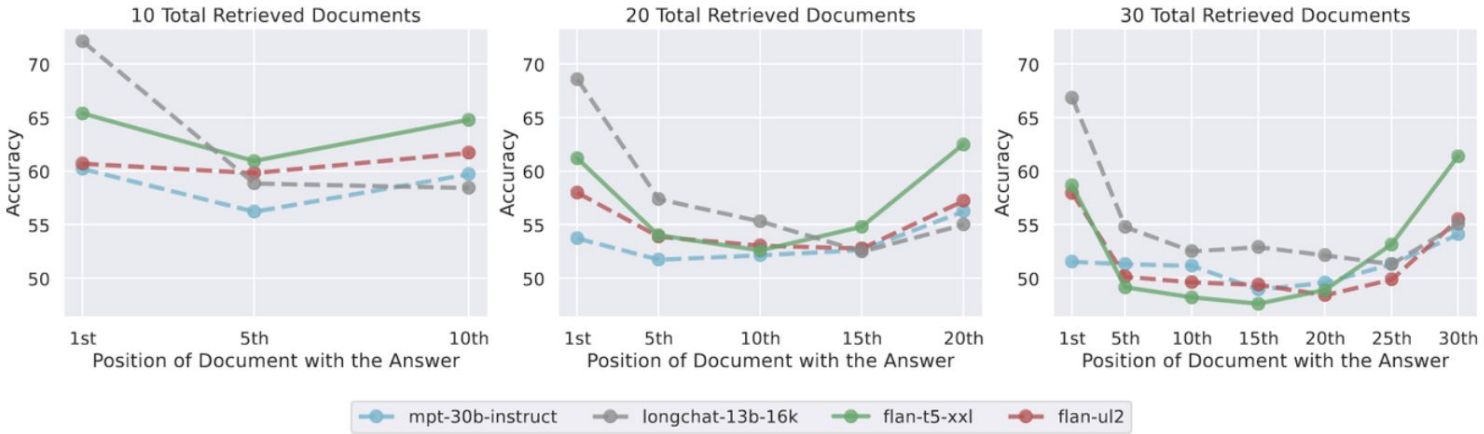
(d) PALM: The encoder predicts masked tokens by encoding context bidirectionally, and the decoder predicts the text segment subsequent to the context. It forces the model to learn to comprehend the context for generating relevant text.

Figure 1: A schematic comparison of PALM with GPT, MASS and BART.

[Bi et. al, 2020]

Lost in the Middle Issue (why does it happen)

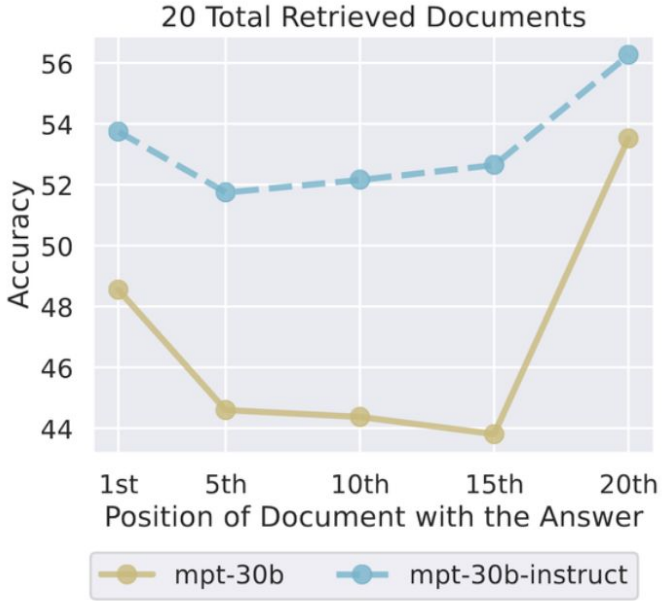
Could it be the decoder-only architecture ?



[Liu et. al, 2023]

Lost in the Middle Issue (why does it happen)

Could it be instruction fine-tuning ?



Liu et. al, 2023