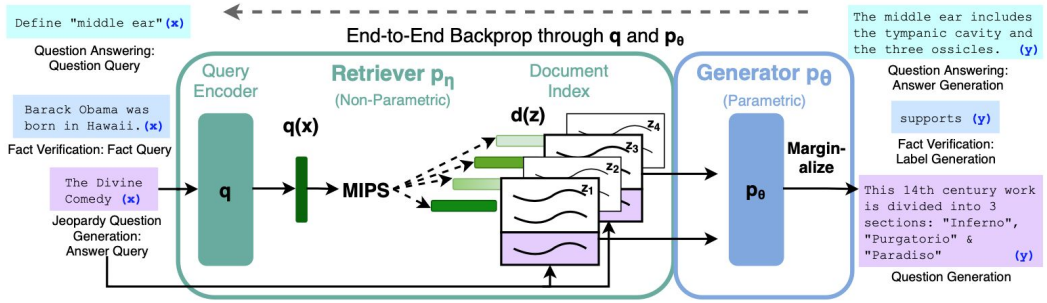
Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis[†‡], Ethan Perez[⋆],

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel[†‡], Sebastian Riedel[†‡], Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; [⋆]New York University;
plewis@fb.com



RETRIEVAL-AUGMENTED GENERATION FOR KNOWLEDGE-INTENSIVE NLP TASKS

Paper highlights by **Andrei Hera**

https://arxiv.org/pdf/2005.11401

WE'LL LOOK INTO...

How can RAG help us

Some limitations of LARGE language models

How RAG works:

Dense Passage Retreiver (Siamese BERT)

Token and Sequence Generators (with BART)

Uses for RAG

Key takeaways

ABOUT ME

- 2013 – First encounter with AI trying to apply it in FPGAs
- 2015 – Crash course in Machine Learning by Google
- 2018 – Joined the Timisoara Deep Learning Meetup
  – @work switched from SW engineering to AI Research: Computer Vision in automotive
- 2019 – First encounter with NLP
  – Changed jobs: Automotive to Financial / CV to NLP
- 2021 – I still ask myself why do people believe what I say when I talk about AI ☺️

Andrei Hera
Timisoara - RO

andrei.hera@gmail.com

BACK TO **RAG**…

$y = f(x)$ where x & y are text sequences

autoregresive
- text generation

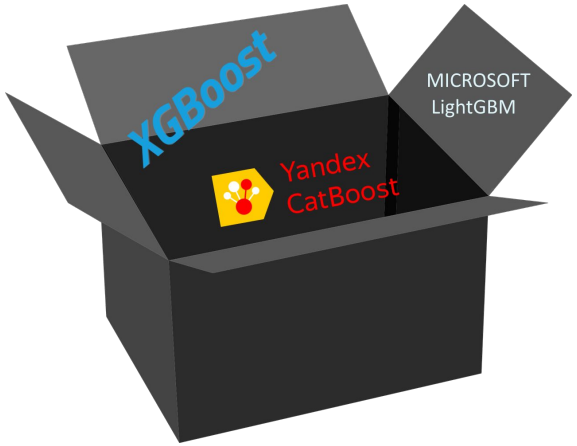
seq2seq
- Q&A
- summarization
- language translation

Why is generating text so hard?

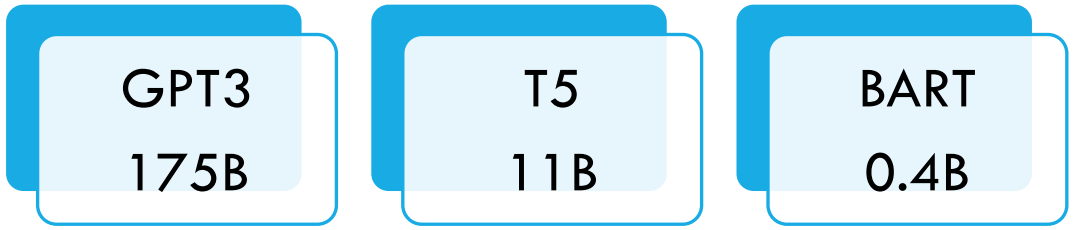
hallucinations

Update memory and facts

Insight into predictions



CLOSED BOOK OR PARAMETRIC MEMORY



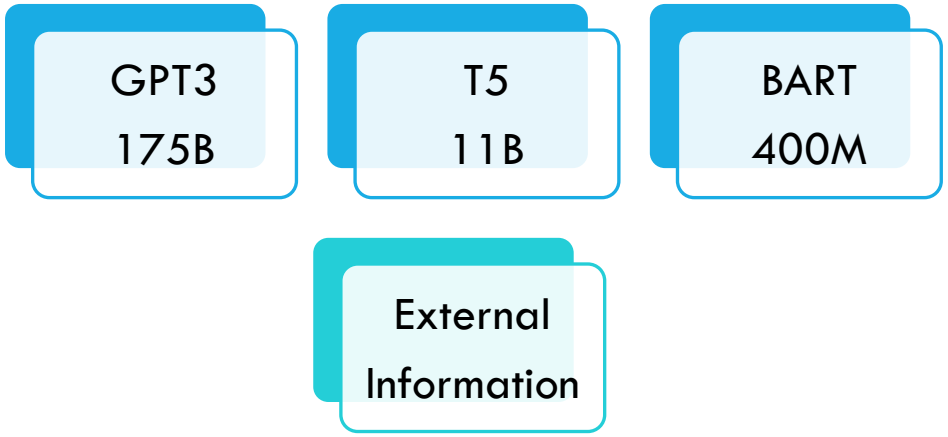
They are like Romanians: They know everything about anything

Abstractive Open Domain QA

Input: how many calories in average apple

BART: The average apple contains 1,000 calories in an average apple and 1,200 calories in a medium apple

OPEN BOOK AND NON-PARAMETRIC MEMORY



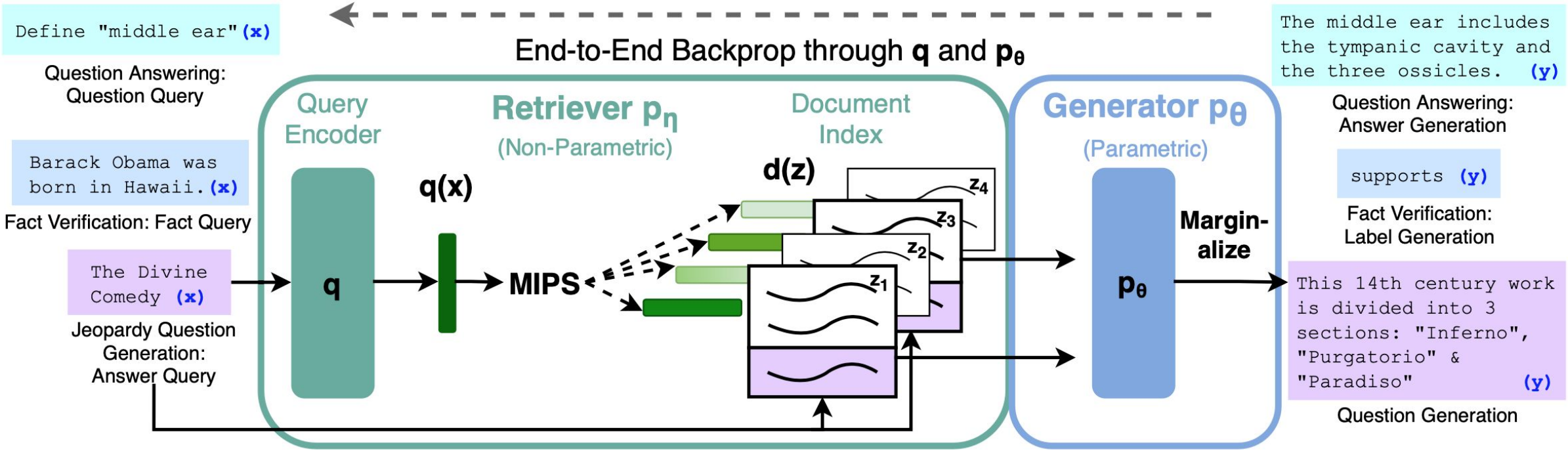
Abstractive Open Domain QA

Input: how many calories in average apple

BART: The average apple contains 1,000 calories in an average apple and 1,200 calories in a medium apple

RAG: There are 126 calories in an average apple, while an extra large size apple has 172 calories.

GOING IN DEPTH

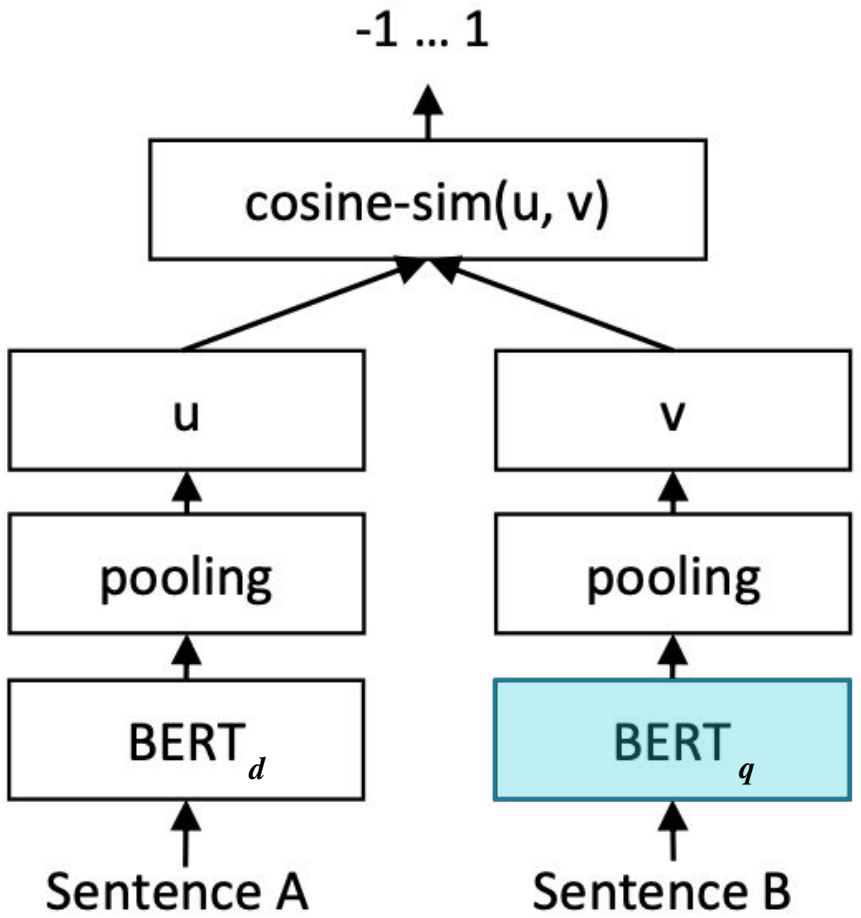


RETRIEVER

- Model: Dense Passage Retriever
- Siemese network / Two tower approach
- Querry encoder weights are **updated**
- Vector similarity search
- Centroid search
- Uses FAISS

https://arxiv.org/abs/1908.10084v1

$$p_\eta(z|x) \propto \exp\left(\mathbf{d}(z)^\top \mathbf{q}(x)\right)$$



(z - document_n)

(x - querry)

$$\mathbf{d}(z) = \mathrm{BERT}_d(z), \quad \mathbf{q}(x) = \mathrm{BERT}_q(x)$$

HOW DOES THE RAG MODEL WORK?

Soruce: Facebook AI Research

θ - generator
η - retriever

GENERATOR

- Model: BART
- Token Generator
- Sequence Generator

RAG - Token

Standard Beam Search with transition probability

$$p_{\text{RAG-Token}}(y|x) \approx \prod_{i}^{N} \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})$$

RAG - Sequence

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x, z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_{i}^{N} p_\theta(y_i|x, z, y_{1:i-1})$$

https://arxiv.org/abs/1910.13461

GENERATOR

- Model: BART
- Token Generator
- Sequence Generator

Θ - generator
η - retriever

RAG - Token

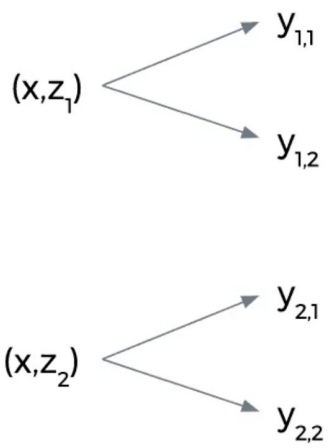
Standard Beam Search

$$p_{\text{RAG-Token}}(y|x) \approx \prod_{i}^{N} \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})$$

RAG - Sequence

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x, z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_{i}^{N} p_\theta(y_i|x, z, y_{1:i-1})$$

Beam search



https://arxiv.org/abs/1910.13461

GENERATOR

- Model: BART
- Token Generator
- Sequence Generator

Θ - generator
η - retriever

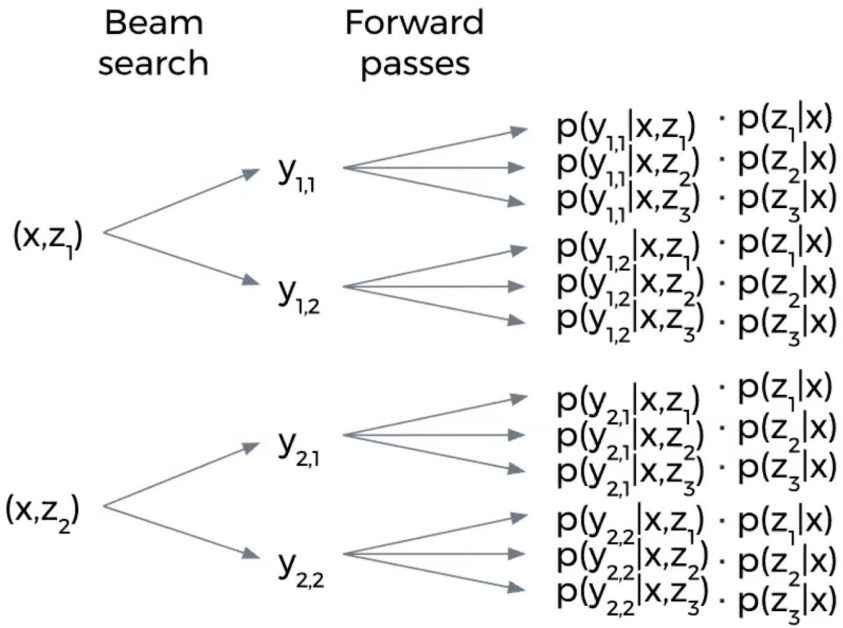
RAG - Token

Standard Beam Search

$$p_{\text{RAG-Token}}(y|x) \approx \prod_{i}^{N} \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})$$

RAG - Sequence

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x, z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_{i}^{N} p_\theta(y_i|x, z, y_{1:i-1})$$



https://arxiv.org/abs/1910.13461

GENERATOR

- Model: BART
- Token Generator
- Sequence Generator

Θ - generator
η - retriever

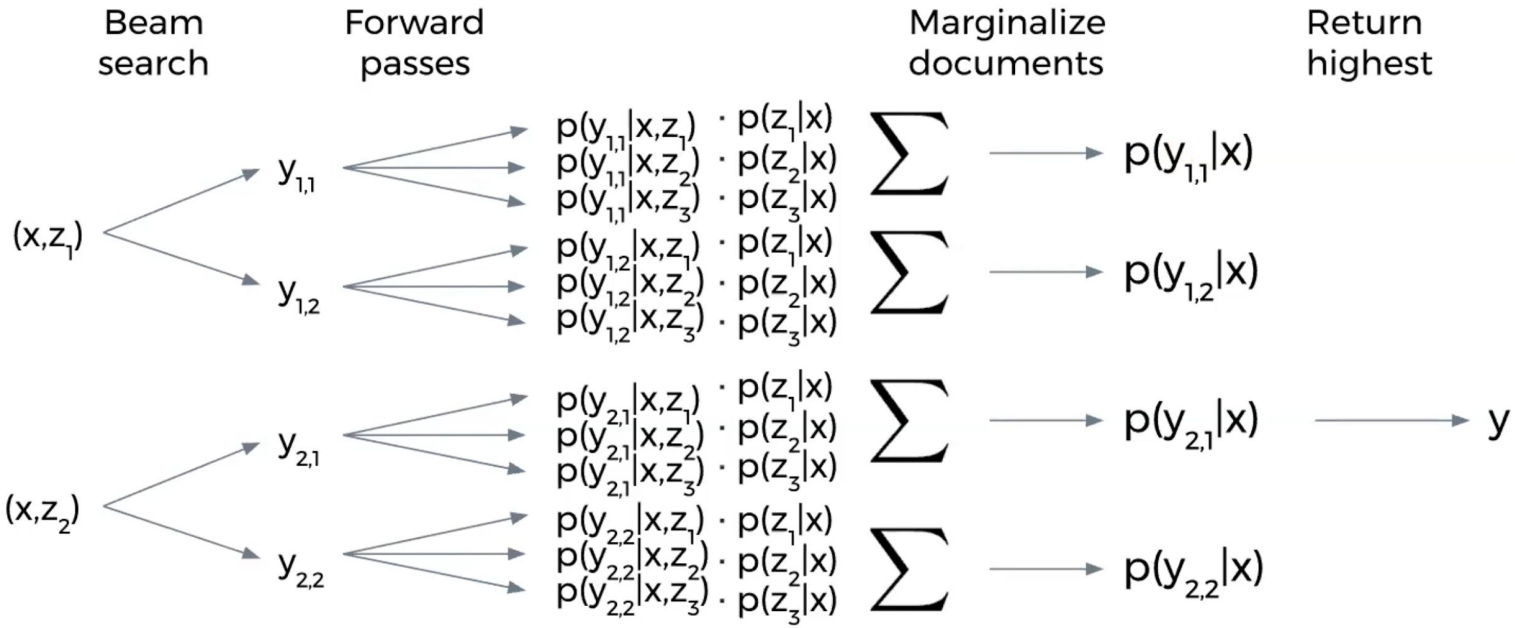
RAG - Token

Standard Beam Search

$$p_{\text{RAG-Token}}(y|x) \approx \prod_{i}^{N} \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})$$

RAG - Sequence

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x, z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_{i}^{N} p_\theta(y_i|x, z, y_{1:i-1})$$



https://arxiv.org/abs/1910.13461

EXPERIMENTS – OPEN BOOK MEETS CLOSED BOOK IN **RAG**

Document 1: his works are considered classics of American literature ... His wartime experiences formed the basis for his novel **"A Farewell to Arms"** (1929) ...

Document 2: ... artists of the 1920s "Lost Generation" expatriate community. His debut novel, **"The Sun Also Rises"**, was published in 1926.

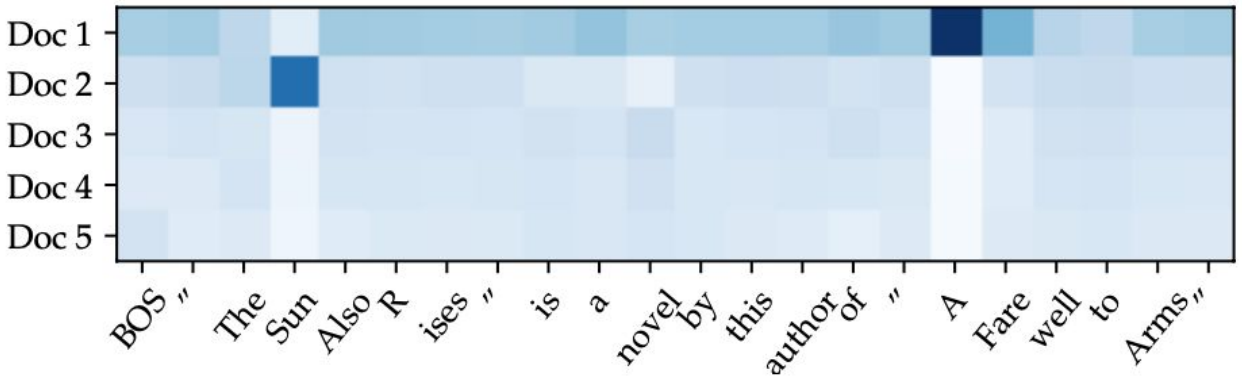


Figure 2: RAG-Token document posterior $p(z_i|x, y_i, y_{-i})$ for each generated token for input "Hemingway" for Jeopardy generation with 5 retrieved documents. The posterior for document 1 is high when generating "A Farewell to Arms" and for document 2 when generating "The Sun Also Rises".

EXPERIMENTS

Table 3: Examples from generation tasks. RAG models generate more specific and factually accurate responses. '?' indicates factually incorrect responses, * indicates partially correct responses.

Task	Input	Model	Generation
MS-MARCO	define middle ear	BART	?The middle ear is the part of the ear between the middle ear and the nose.
		RAG-T	The middle ear is the portion of the ear internal to the eardrum.
		RAG-S	The middle ear includes the tympanic cavity and the three ossicles.
	what currency needed in scotland	BART	The currency needed in Scotland is Pound sterling.
		RAG-T	Pound is the currency needed in Scotland.
		RAG-S	The currency needed in Scotland is the pound sterling.
Jeopardy Question Gener -ation	Washington	BART	?This state has the largest number of counties in the U.S.
		RAG-T	It's the only U.S. state named for a U.S. president
		RAG-S	It's the state where you'll find Mount Rainier National Park
	The Divine Comedy	BART	*This epic poem by Dante is divided into 3 parts: the Inferno, the Purgatorio & the Purgatorio
		RAG-T	Dante's "Inferno" is the first part of this epic poem
		RAG-S	This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso"

RESULTS

Table 1: Open-Domain QA Test Scores. For TQA, left column uses the standard test set for Open-Domain QA, right column uses the TQA-Wiki test set. See Appendix D for further details.

	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**

Table 2: Generation and classification Test Scores. MS-MARCO SotA is [4], FEVER-3 is [68] and FEVER-2 is [57] *Uses gold context/evidence. Best model without gold access underlined.

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		

- Other aspect such as improving quality of generated texts is also improved according to human evaluators

KEY TAKEAWAYS

A FLEXIBLE WAY TO COMBINE LMS WITH EXTRINSIC DATA

REDUCES HALLUCINATIONS

IMPROVES OVERALL QUALITY OF GENERATED TEXTS

We’re not there yet… so keep on learning ☺️

THANK YOU!

REFERENCES AND CREDITS

Great videos that helped me:

Patrick Lewis*:
https://www.youtube.com/watch?v=JGpmQvlYRdU

Henry AI Labs:
https://www.youtube.com/watch?v=dzChvuZI6D4

Demo:

https://huggingface.co/rag

Other references:

https://ai.facebook.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/

(*) He is one of the authors. Some diagrams are screenshots from this video.