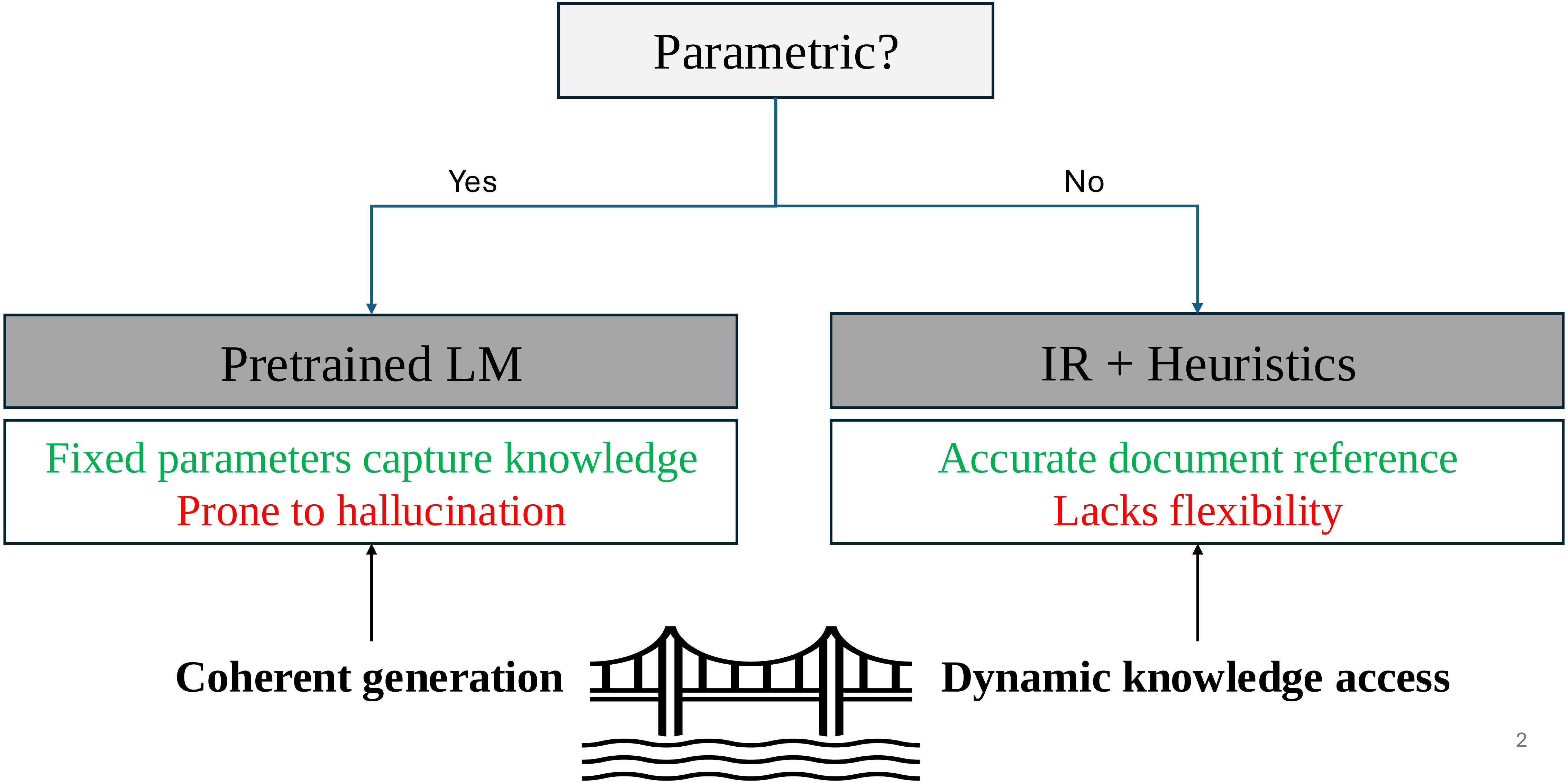
Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

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

Knowledge-Intensive NLP Tasks



Parametric?

Yes

No

Pretrained LM

Fixed parameters capture knowledge

Prone to hallucination

IR + Heuristics

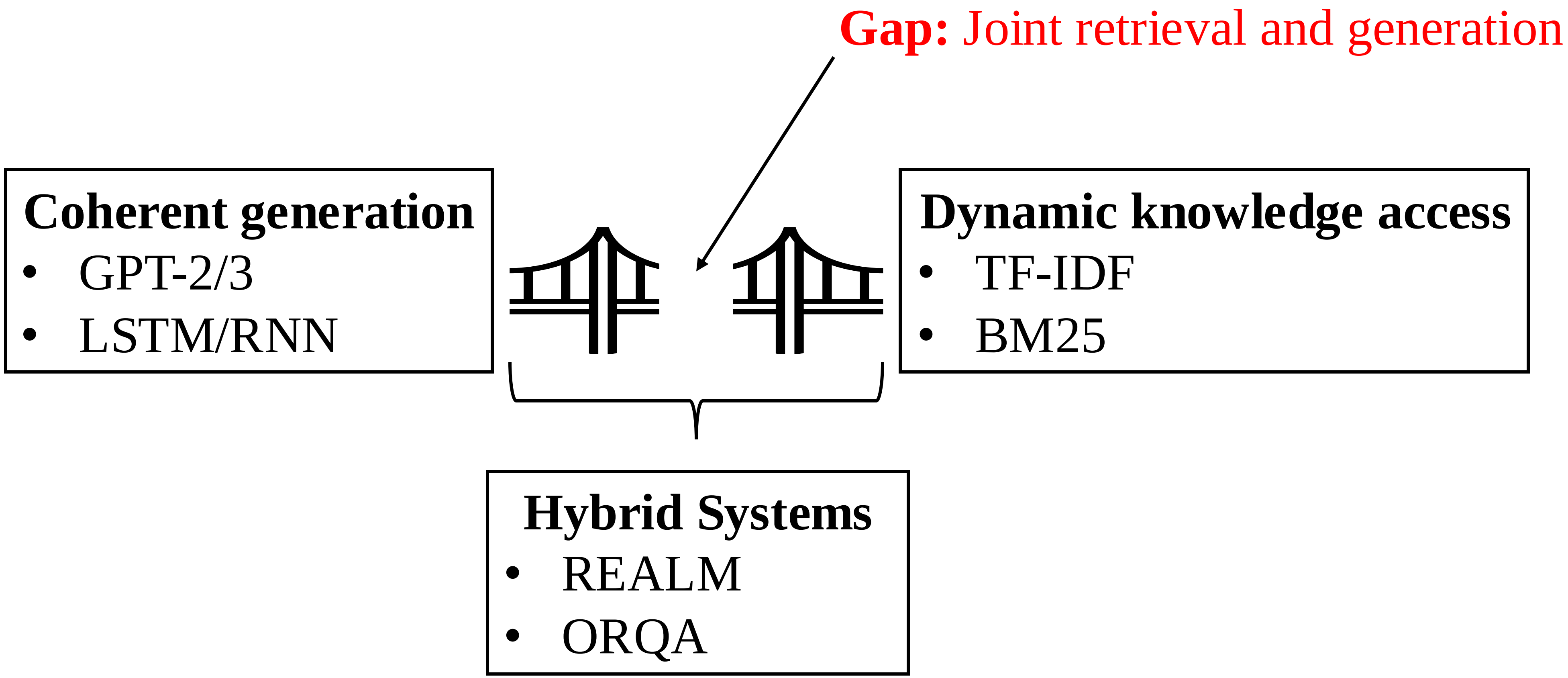
Accurate document reference

Lacks flexibility

Coherent generation

Dynamic knowledge access

Previous Approaches



Gap: Joint retrieval and generation

Coherent generation

- GPT-2/3
- LSTM/RNN

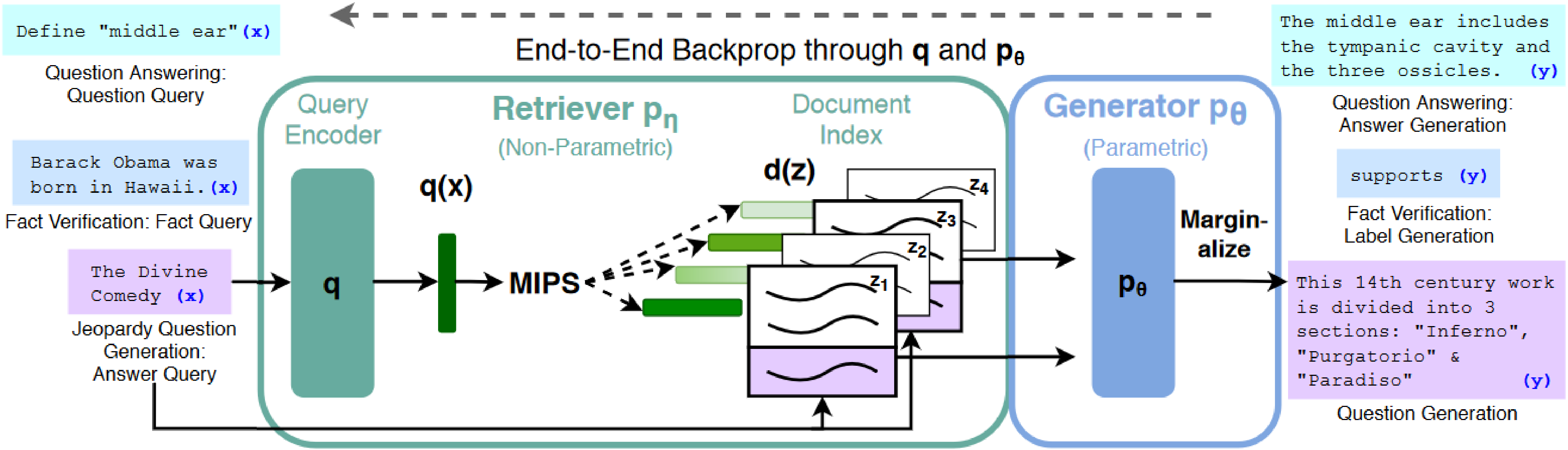
Dynamic knowledge access

- TF-IDF
- BM25

Hybrid Systems

- REALM
- ORQA

Retrieval-Augmented Generation (RAG)


Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and p_θ
Query Encoder
q
q(x)
Retriever p_η
(Non-Parametric)
MIPS
Document Index
d(z)
z_4
z_3
z_2
z_1
Generator p_θ
(Parametric)
p_θ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

RAG Variants

x ← Input sequence

y ← Output sequence

z ← Latent document

Sequence-level marginalization

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

Token-level marginalization

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

$p_\eta(\cdot)$ ← Retriever

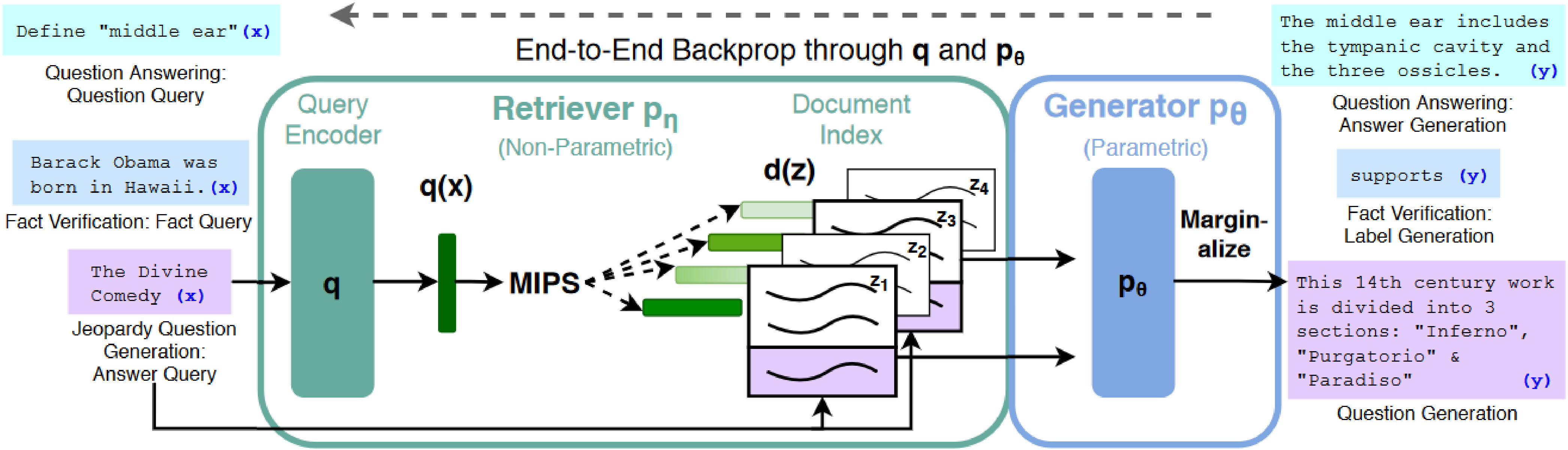
$p_\theta(\cdot)$ ← Generator

Training: minimization of negative marginal log likelihood

$$(x_j, y_j) \to \textstyle\sum_j -\log p(y_j|x_j)$$

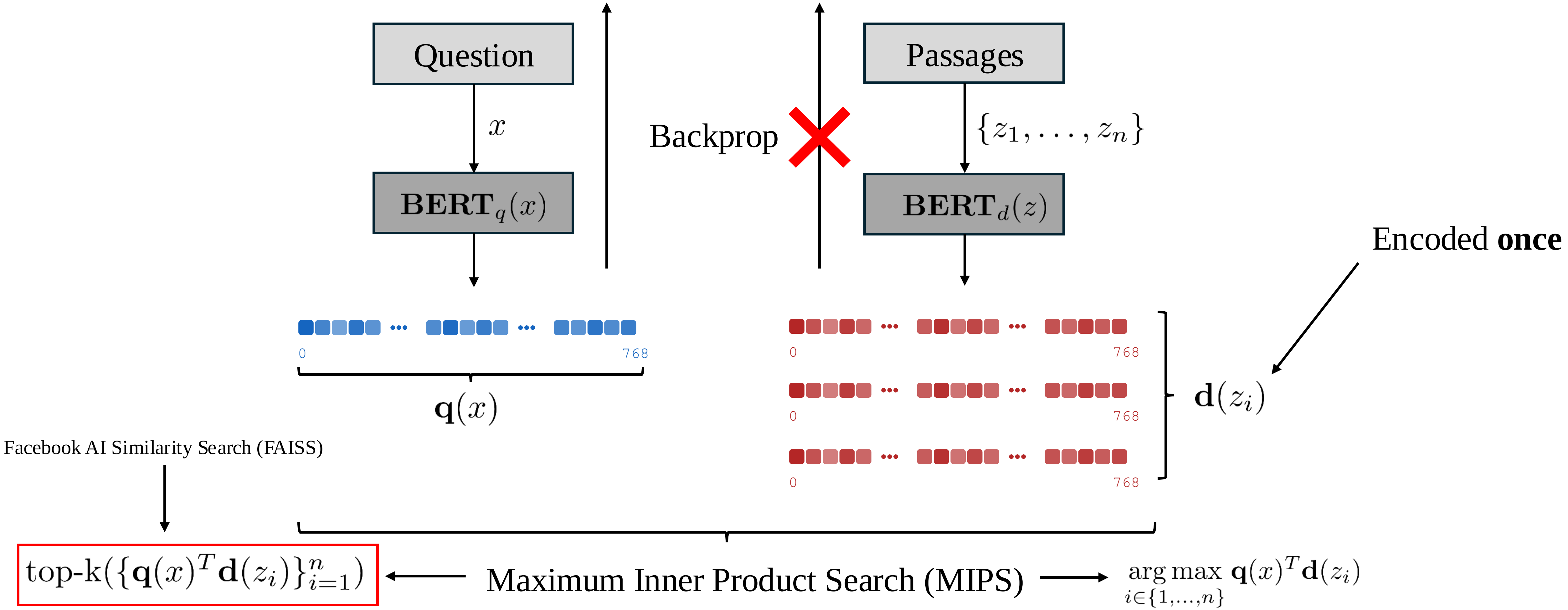
Karpukhin, V. et al. (2020). *Dense Passage Retrieval for Open-Domain Question Answering.* EMNLP 2020, Online. ACL.

Retrieval-Augmented Generation (RAG)



Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and p_θ
Query Encoder
Retriever p_η
(Non-Parametric)
Document Index
q
q(x)
MIPS
d(z)
z_4
z_3
z_2
z_1
Generator p_θ
(Parametric)
p_θ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

Dense Passage Retrieval (DPR)



Question

x

$\mathbf{BERT}_q(x)$

$\mathbf{q}(x)$

Backprop

Passages

$\{z_1, \ldots, z_n\}$

$\mathbf{BERT}_d(z)$

Encoded **once**

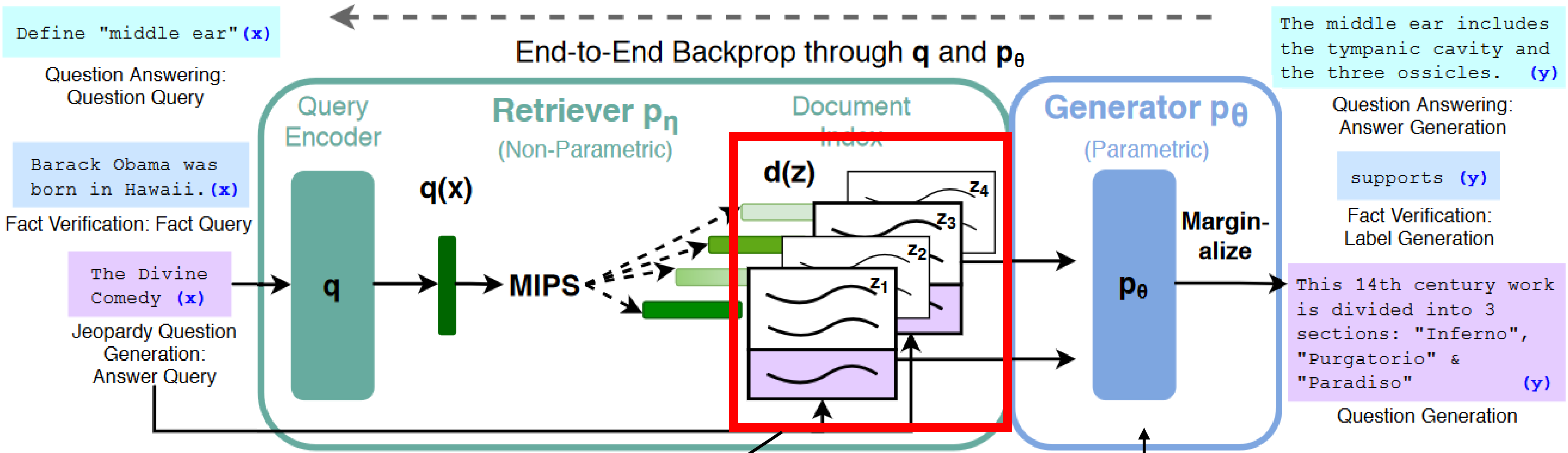
$\mathbf{d}(z_i)$

Facebook AI Similarity Search (FAISS)

$\boxed{\text{top-k}(\{\mathbf{q}(x)^T\mathbf{d}(z_i)\}_{i=1}^{n})}$ ← Maximum Inner Product Search (MIPS) → $\arg\max_{i\in\{1,\ldots,n\}} \mathbf{q}(x)^T\mathbf{d}(z_i)$

Karpukhin, V. et al. (2020). *Dense Passage Retrieval for Open-Domain Question Answering*. EMNLP 2020, Online. ACL.

Retrieval-Augmented Generation (RAG)



Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and pθ
Query Encoder
q
q(x)
Retriever pη
(Non-Parametric)
MIPS
Document Index
d(z)
z4
z3
z2
z1
Generator pθ
(Parametric)
pθ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

Query *concatenated* to **documents**

Pretrained Seq2seq (BART-Large)

x ← Input sequence

y ← Output sequence

z ← Latent document

RAG Variants

Sequence-level marginalization

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

Token-level marginalization

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

$p_\eta(\cdot)$ ← Retriever

$p_\theta(\cdot)$ ← Generator

Training: minimization of negative marginal log likelihood

$$(x_j, y_j) \rightarrow \textstyle\sum_j -\log p(y_j|x_j)$$

Karpukhin, V. et al. (2020). *Dense Passage Retrieval for Open-Domain Question Answering*. EMNLP 2020, Online. ACL.

Experiments

- Can be applied to any task with input/output sequence

- Focus is on tasks with need for precise knowledge access

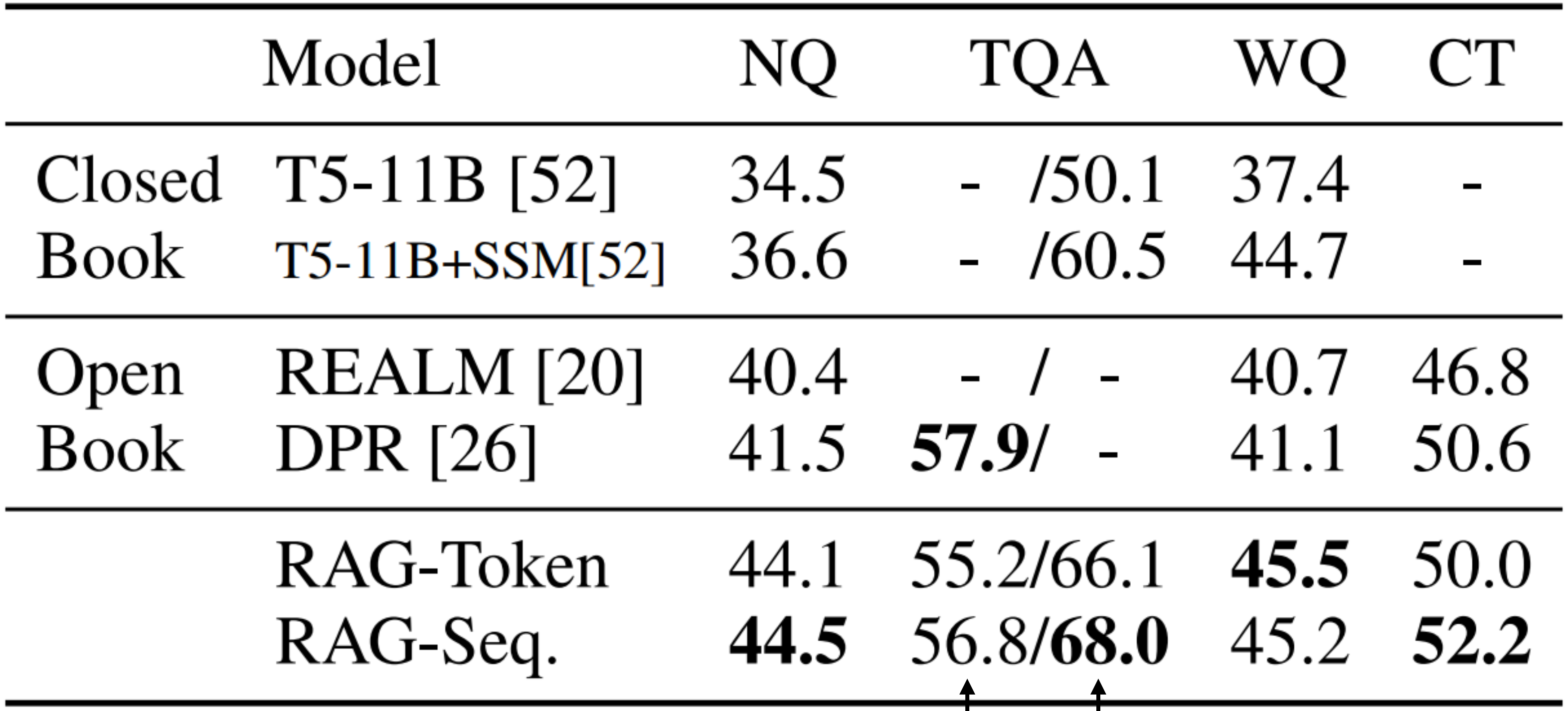
Open-domain QA

Question Generation

Abstractive open-domain QA

Fact Verification

Results – Open-Domain QA



	Model	NQ	TQA	WQ	CT
Closed	T5-11B [52]	34.5	- /50.1	37.4	-
Book	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open	REALM [20]	40.4	- / -	40.7	46.8
Book	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**

Standard test set

TQA-Wiki test set

Results – Abstractive QA

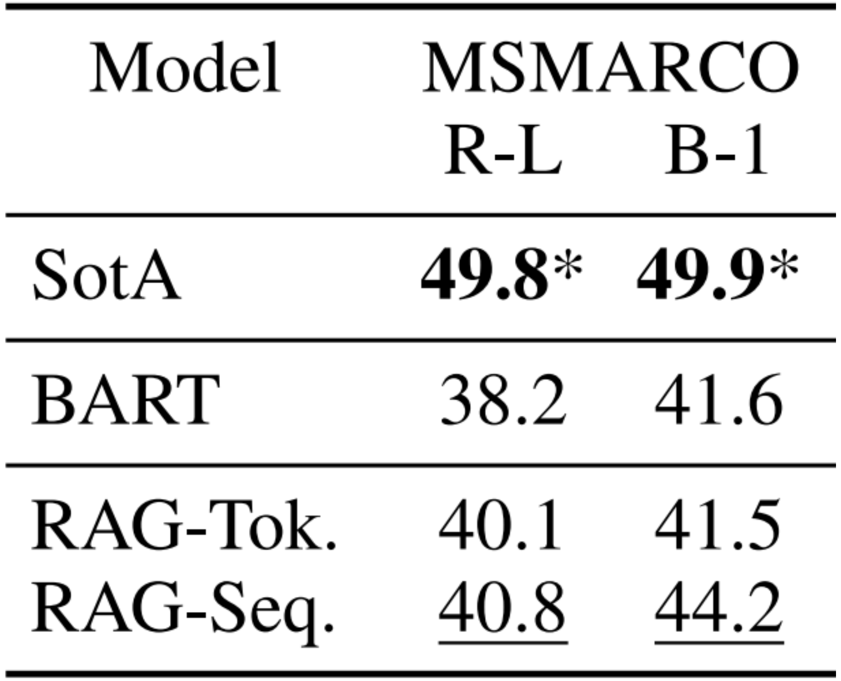
Input: how many calories in average apple

GOLD: an average apple has 80 calories

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

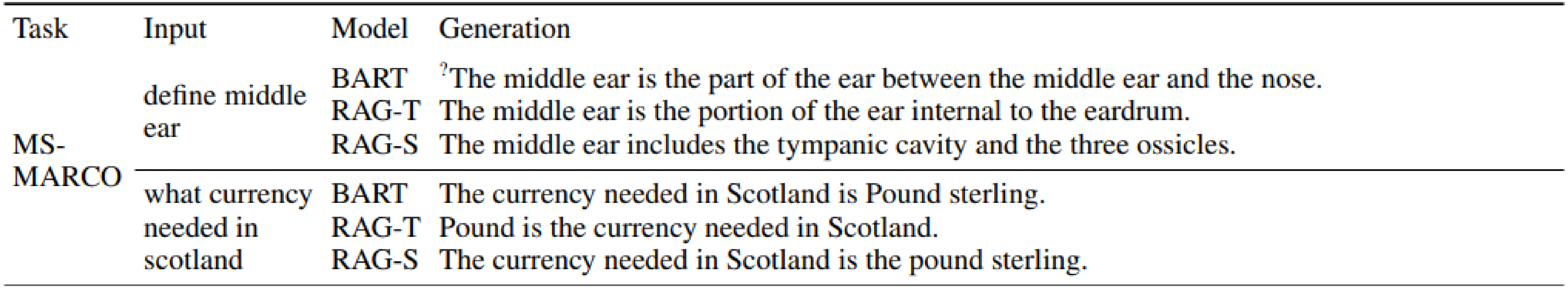
Top Retrieved doc: A typical apple serving weighs 242 grams and provides 126 calories with a moderate content of dietary fiber (table). Otherwise, there is ... is usually not eaten and is discarded.



Model	MSMARCO R-L	 B-1
SotA	**49.8***	**49.9***
BART	38.2	41.6
RAG-Tok.	40.1	41.5
RAG-Seq.	40.8	44.2

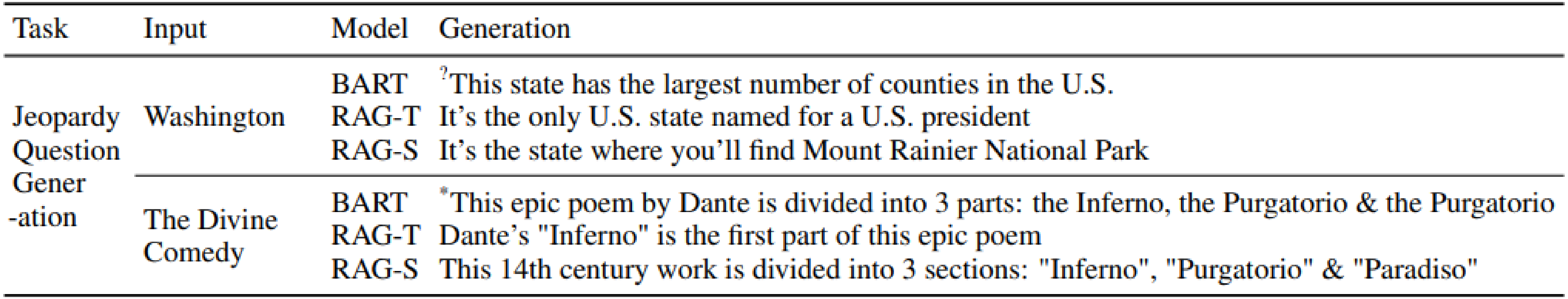
*Uses gold context/evidence. Best model without gold access underlined.

Results – Abstractive QA

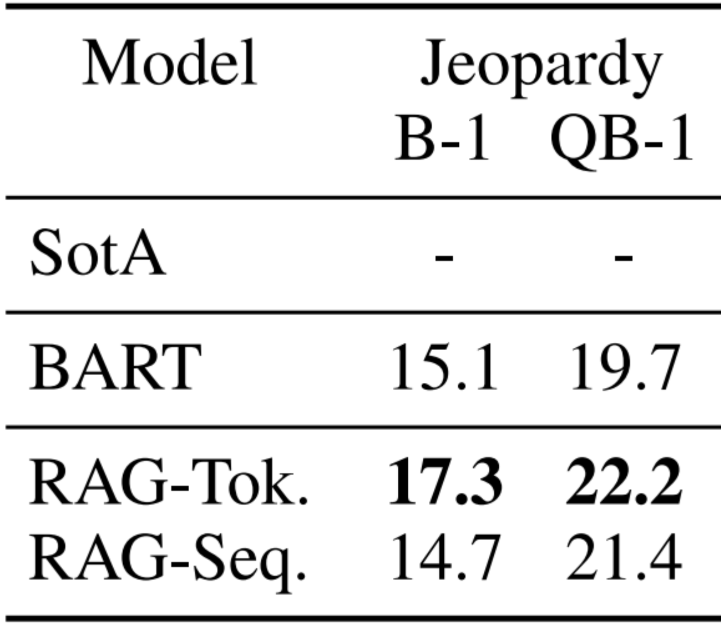


Task	Input	Model	Generation
MS-MARCO	define middle ear	BART	[?]The middle ear is the part of the ear between the middle ear and the nose.
		RAG-T	The middle ear is the portion of the ear internal to the eardrum.
		RAG-S	The middle ear includes the tympanic cavity and the three ossicles.
	what currency 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.

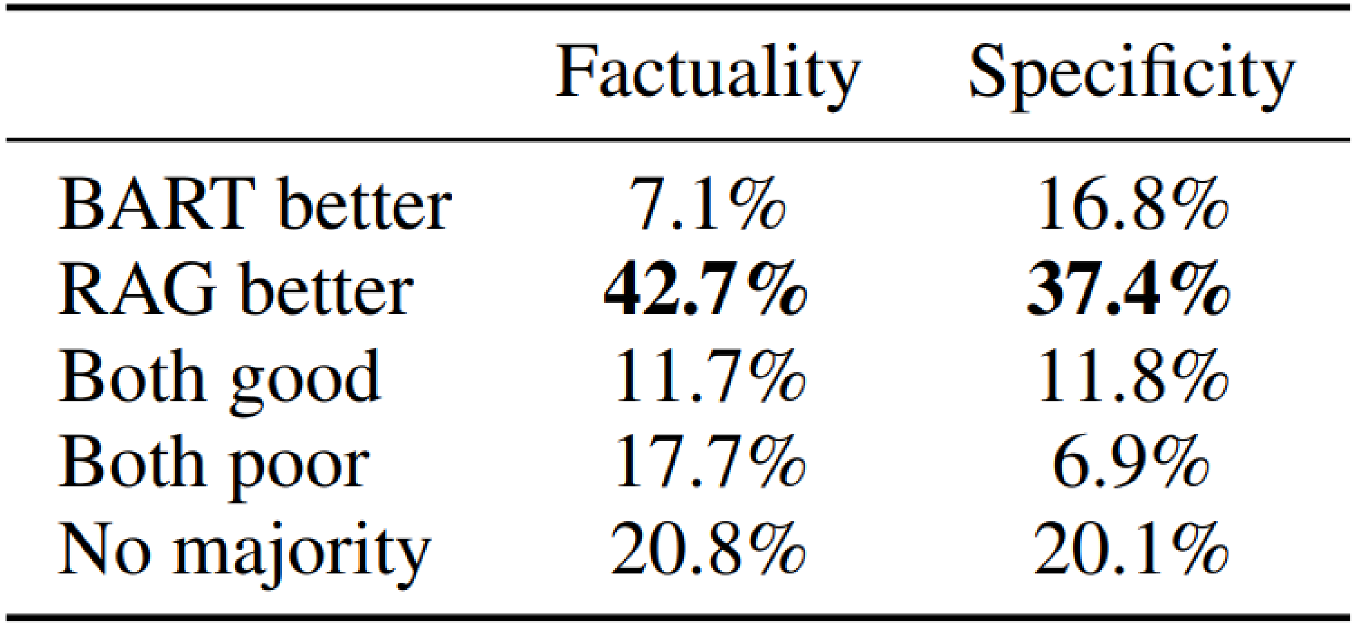
Results – Jeopardy Question Generation



Task	Input	Model	Generation
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"

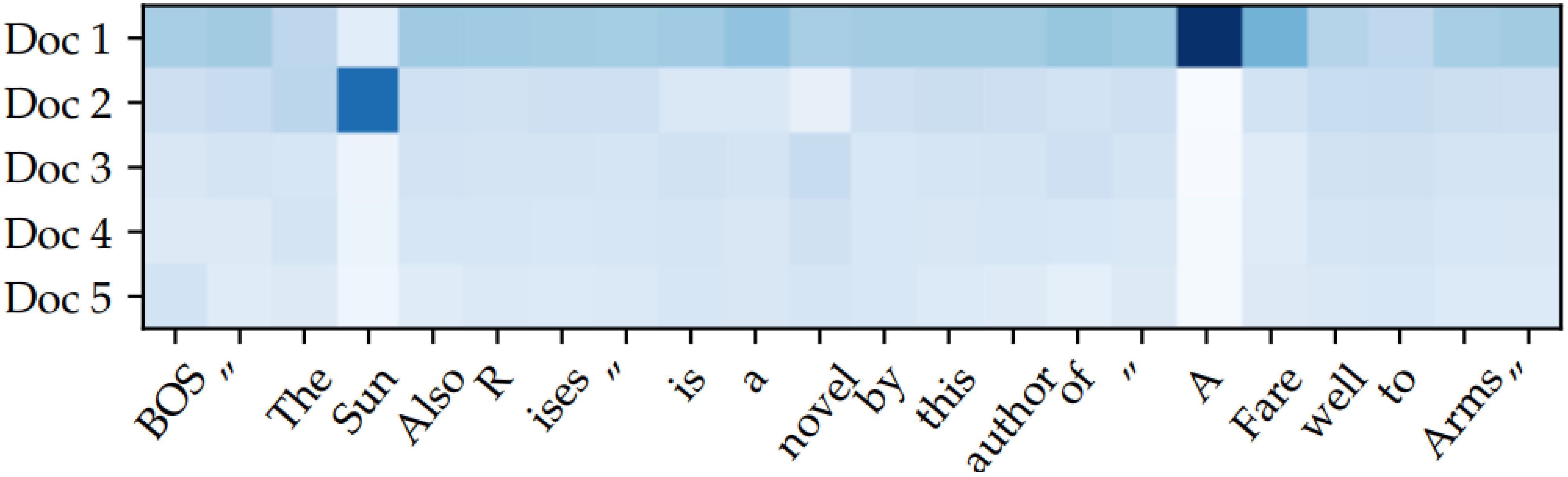


Model	Jeopardy	
	B-1	QB-1
SotA	-	-
BART	15.1	19.7
RAG-Tok.	**17.3**	**22.2**
RAG-Seq.	14.7	21.4



	Factuality	Specificity
BART better	7.1%	16.8%
RAG better	**42.7%**	**37.4%**
Both good	11.7%	11.8%
Both poor	17.7%	6.9%
No majority	20.8%	20.1%

Results – Jeopardy Question Generation

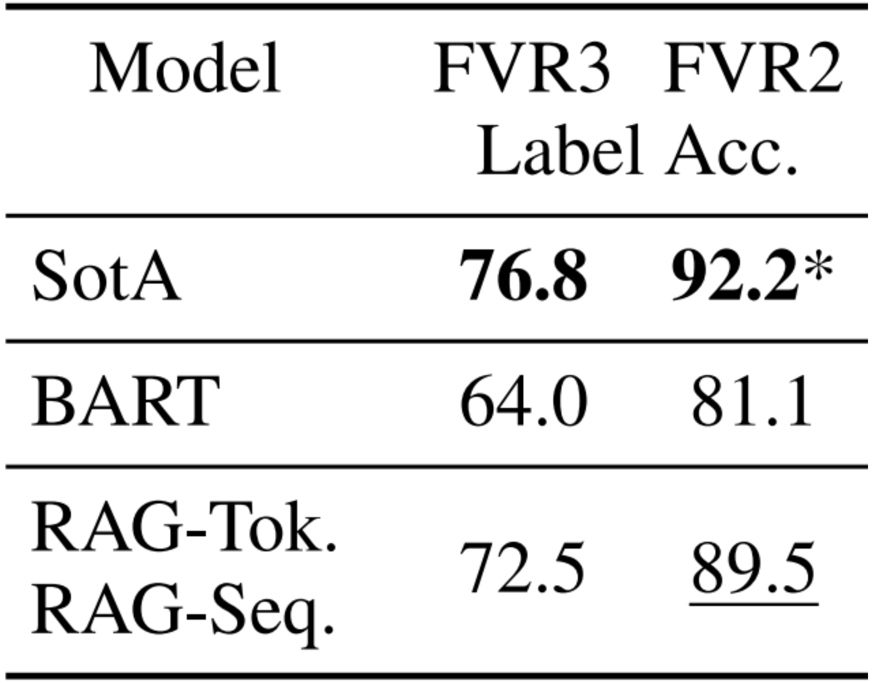


Document 1: his works are considered classics of American literature ... His wartime experiences formed the basis for his novel **"A Farewell to Arms"** (1929) ...

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

Results – Fact checking

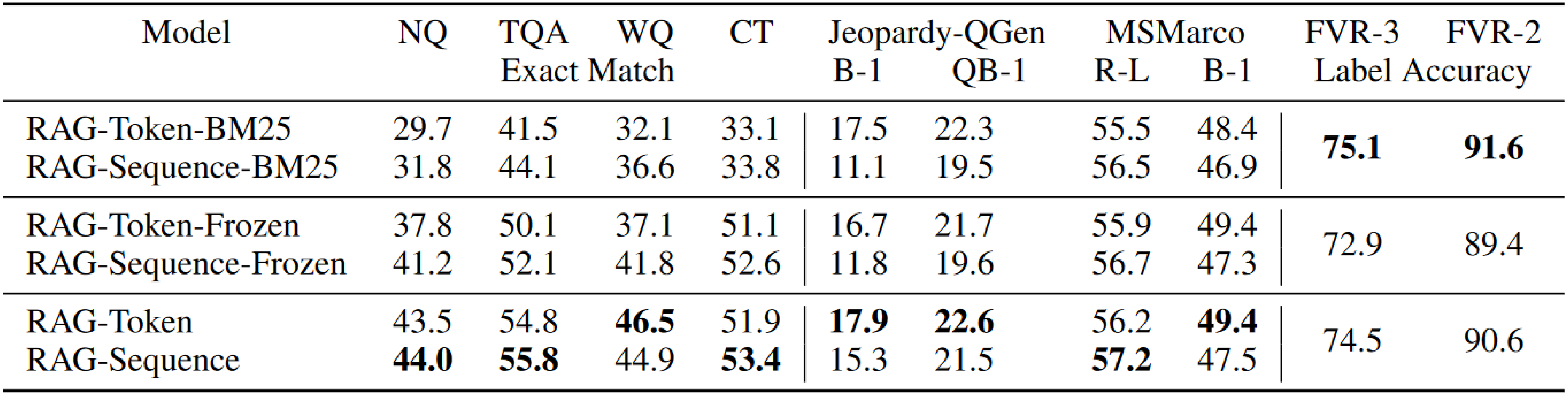
- **FVR3**: supports/refutes/not enough info
- **FVR2**: supports/refutes
- **SotA:** complex pipeline, retrieval supervision
- **RAG:** No supervision on retrieved evidence



Model	FVR3	FVR2
	Label Acc.	
SotA	**76.8**	**92.2***
BART	64.0	81.1
RAG-Tok. RAG-Seq.	72.5	<u>89.5</u>

*Uses gold context/evidence. Best model without gold access underlined.

Results – Ablation



Model	NQ	TQA	WQ	CT	Jeopardy-QGen		MSMarco		FVR-3	FVR-2
		Exact Match			B-1	QB-1	R-L	B-1	Label Accuracy	
RAG-Token-BM25	29.7	41.5	32.1	33.1	17.5	22.3	55.5	48.4	**75.1**	**91.6**
RAG-Sequence-BM25	31.8	44.1	36.6	33.8	11.1	19.5	56.5	46.9		
RAG-Token-Frozen	37.8	50.1	37.1	51.1	16.7	21.7	55.9	49.4	72.9	89.4
RAG-Sequence-Frozen	41.2	52.1	41.8	52.6	11.8	19.6	56.7	47.3		
RAG-Token	43.5	54.8	**46.5**	51.9	**17.9**	**22.6**	56.2	**49.4**	74.5	90.6
RAG-Sequence	**44.0**	**55.8**	44.9	**53.4**	15.3	21.5	**57.2**	47.5		

Conclusion

- **Hybrid generation**: Access to parametric/non-parametric memory
- **Learned retrieval**: Ablations support trainable retriever
- **Index hot-Swapping**: Update model memory on the fly
- **Fewer hallucinations**
- **Knowledge source bias**

Questions?