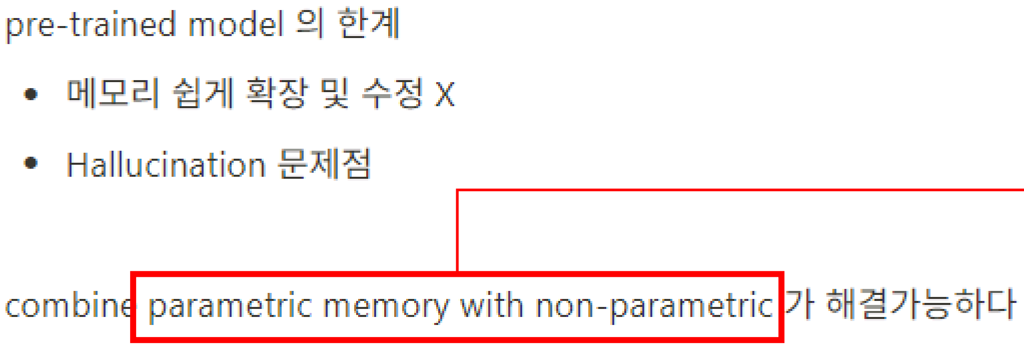
간단 논세

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

24.11.14
유하영

Introduction



pre-trained model 의 한계

- 메모리 쉽게 확장 및 수정 X
- Hallucination 문제점

combine parametric memory with non-parametric가 해결가능하다

parametric vs Non-parametric

parametric model

: 학습하면서 결정해야 하는 모델의 파라미터 수가 명확하게 결정되어 있음

non-parametric model

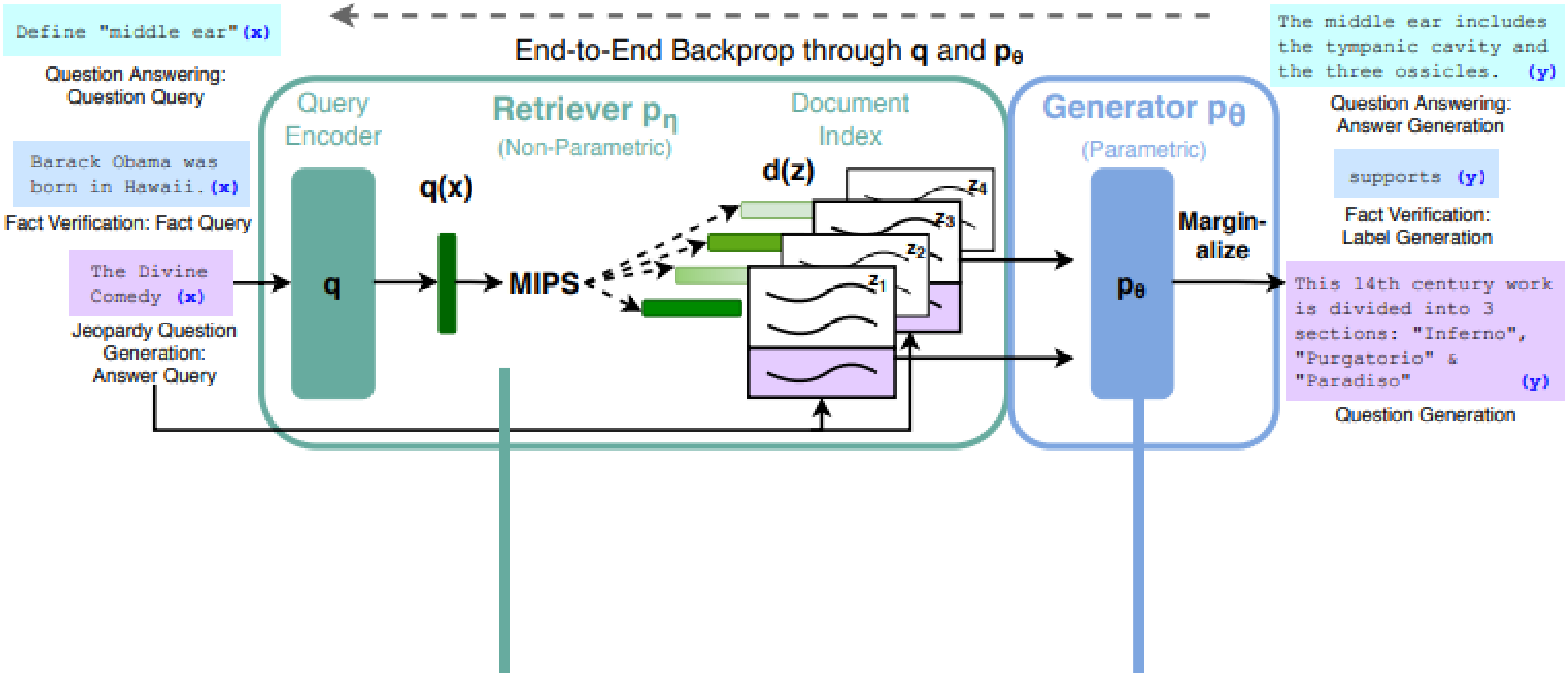
: 파라미터가 명확하게 정해져 있지 X.

seq2seq모델에 도입.

이전 연구에서는 주로 **텍스트 추출** 방식으로 질문에 답변

→ **질문에 대한 답변을 직접 generation 하는 방법을 본 논문에서 제안**

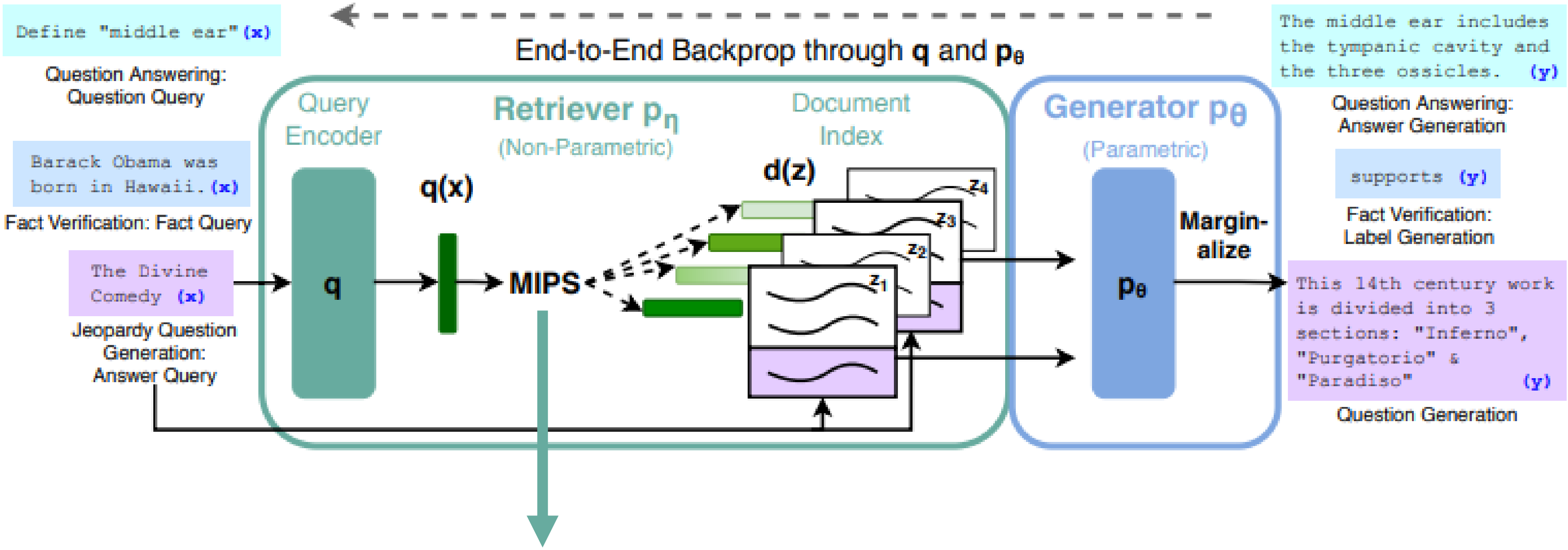
Method


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
Retriever pη
(Non-Parametric)
Document Index
q
q(x)
MIPS
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

(i) a retriever $p_\eta(z|x)$

(ii) a generator $p_\theta(y_i|x, z, y_{1:i-1})$

Method


End-to-End Backprop through q and pθ
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
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

$$p_\eta(z|x) \propto \exp\left(\mathbf{d}(z)^\top \mathbf{q}(x)\right) \qquad \mathbf{d}(z) = \text{BERT}_d(z), \;\; \mathbf{q}(x) = \text{BERT}_q(x)$$

$$\downarrow$$

$$p(z|x) = \frac{\exp f(x,z)}{\sum_{z'} \exp f(x,z')}$$

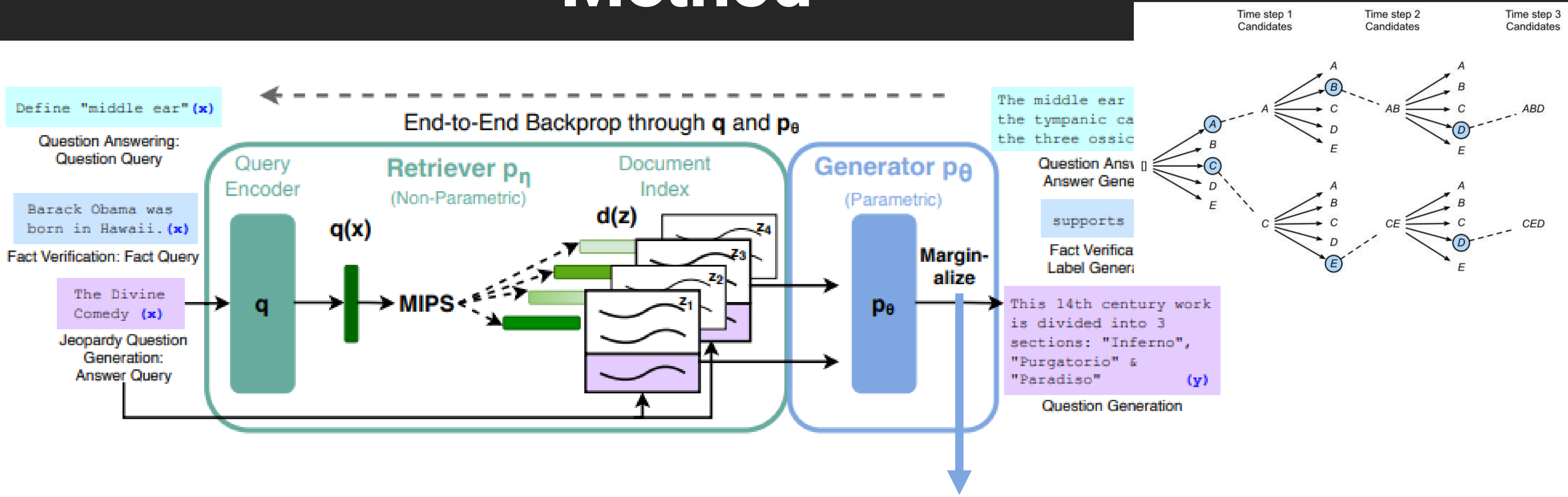
Dense Passage Retriever

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

Maximum Inner Product Search

$$\underset{d \in D}{\text{argmax}}\; q^T d$$

Method



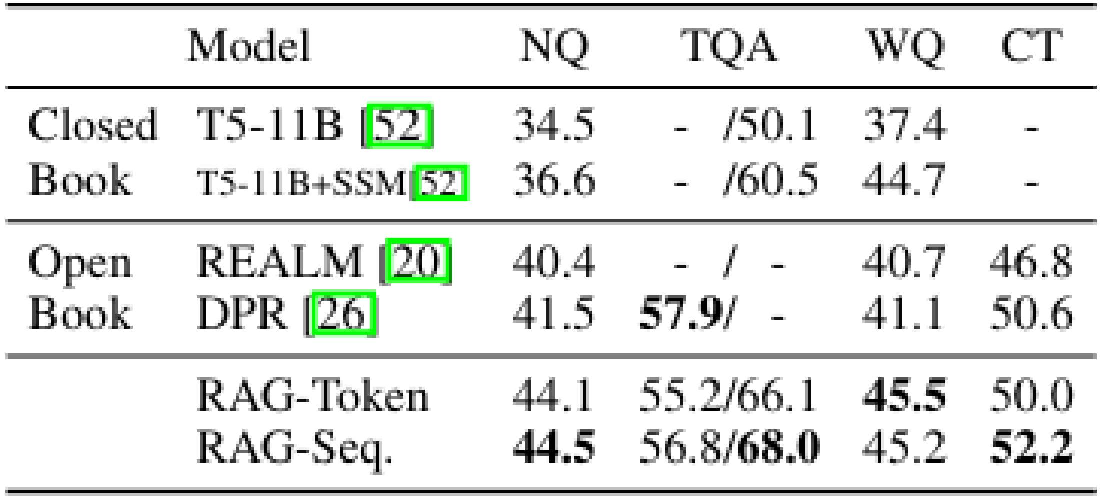
RAG-Sequence Model

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

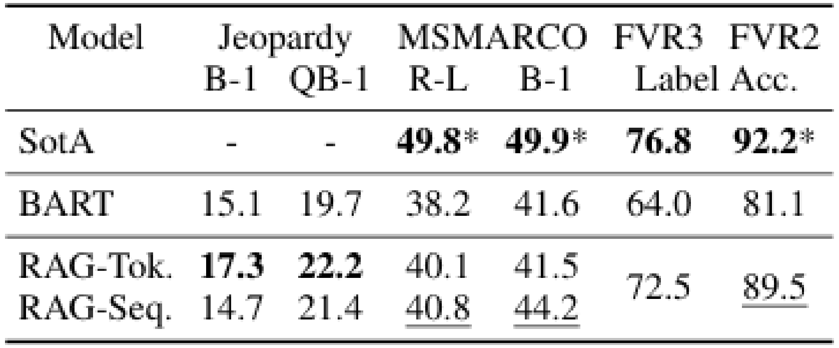
RAG-Token Model

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

Experiments

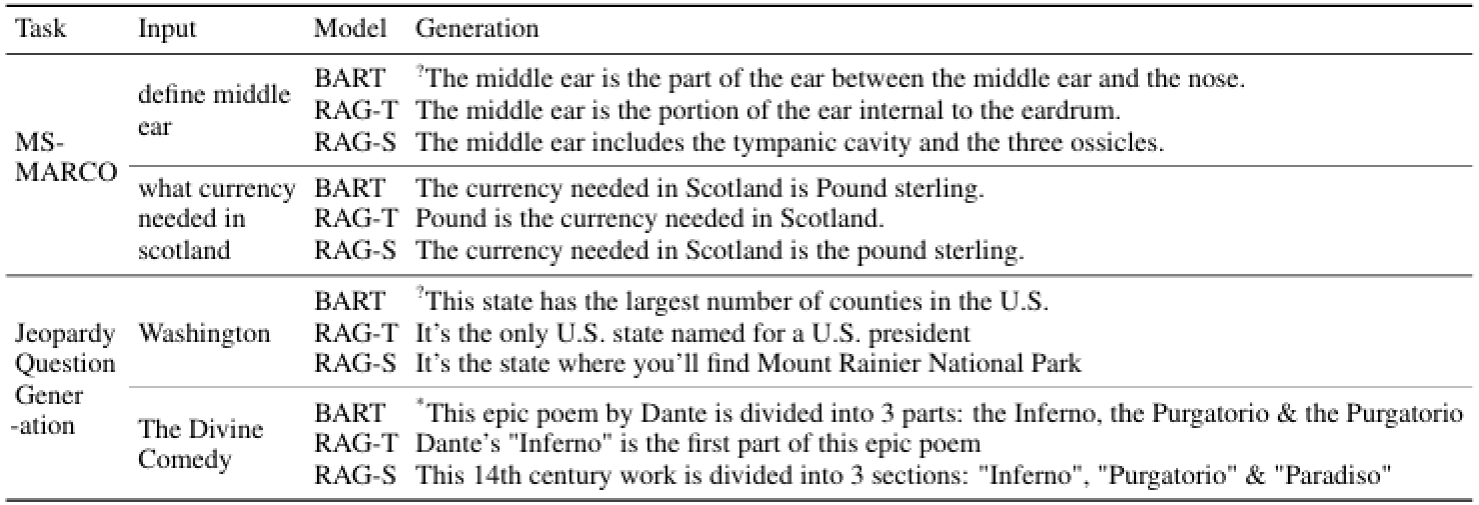


	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**



Model	Jeopardy		MSMARCO		FVR3	FVR2
	B-1	QB-1	R-L	B-1	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		

Examples from generation tasks.



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"

Discussion

Retriever와 Generator가 함께 학습될 수 있는지?

Parametric Memory와 Non-parametric Memory의 상호작용