

DMQA Open Seminar

Retriever for Language Models

2024.06.07

고려대학교 산업경영공학과

Data Mining & Quality Analytics Lab.

이정민

발표자 소개



❖ 이정민(JungMin Lee)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab.(김성범 교수님)
- 석박 통합 과정(2022.03~Present)

❖ Research Interest

- Uncertainty Quantification
- Large Language Models

❖ Contact

- jungmin9195@korea.ac.kr

Contents

❖ Introduction

- What is LLM?
- Development of LLM

❖ Retrieval with Language Models

- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2020, NeurIPS)
- Improving Language Models by Retrieving from Trillions of Tokens (2022, PMLR)
- Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study (2023, EMNLP)

❖ Conclusion

❖ References

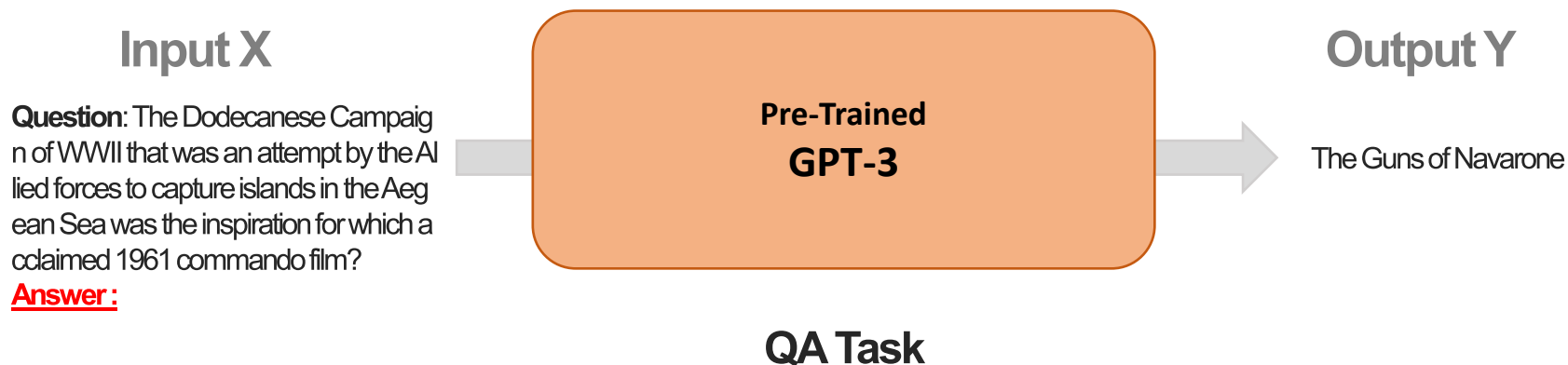
Introduction

Introduction

What is LLM?

❖ Large Language Models(LLM)

- 방대한 양의 텍스트로 사전 학습 된 언어 모델
- 대용량의 언어 모델을 통해 다양한 task를 수행할 수 있음



Introduction

What is LLM?

❖ Large Language Models(LLM)

- 방대한 양의 텍스트로 사전 학습 된 언어 모델
- 대용량의 언어 모델을 통해 다양한 task를 수행할 수 있음



[ChatGPT]



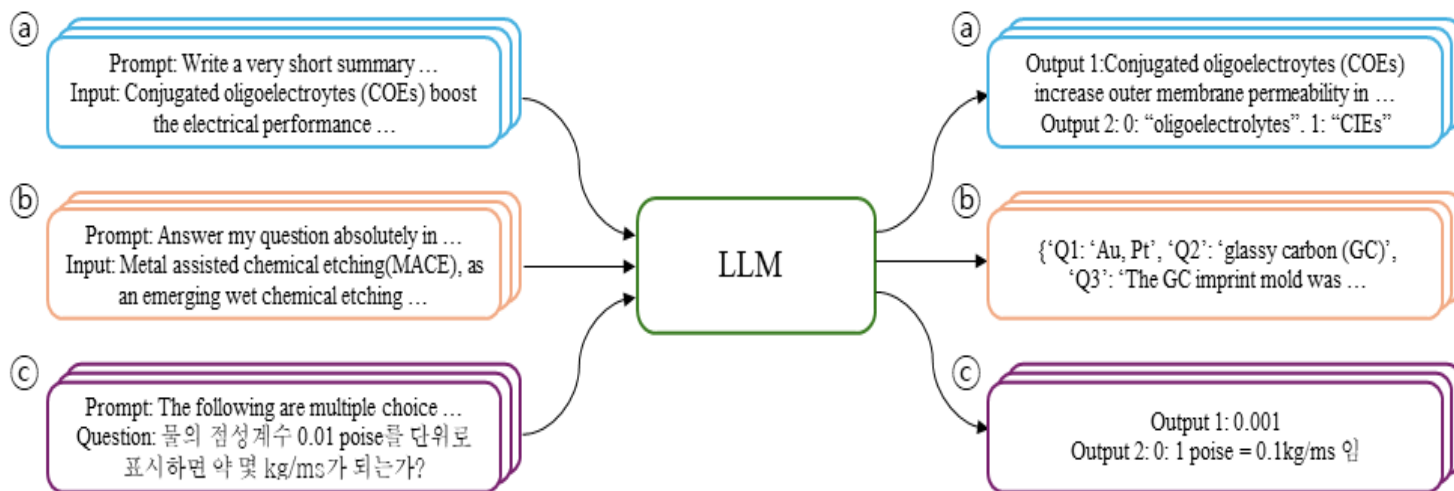
[LLaMA]

Introduction

What is LLM?

❖ Instruction-tuning

- Instruction(지시문) / input text / output text 데이터셋을 통해 지도 학습 형식으로 학습
- 사용자의 의도 및 task에 맞게 LLM을 잘 활용하기 위해서는 instruction-tuning이 필수적



[Instruction-tuning]


Introduction

Related Seminar


❖ Language Model들의 발전과 LLM에 대해 소개한 세미나




종료

What is LLM and ChatGPT?


2023. 07. 28
Data Mining & Security Analytics Lab.

What is LLM and ChatGPT?


발표자:  **채고은**

 2023년 7월 28일
 오후 12시 ~
 온라인 비디오 시청 (YouTube)


세미나 정보 보기 →




종료

Training Techniques and Research Trends of LLM


2023. 08. 04
Data Mining & Security Analytics Lab.
김현지

Training Techniques and Research Trends

발표자:  **김현지**

 2023년 8월 4일
 오전 12시 ~
 온라인 비디오 시청 (YouTube)

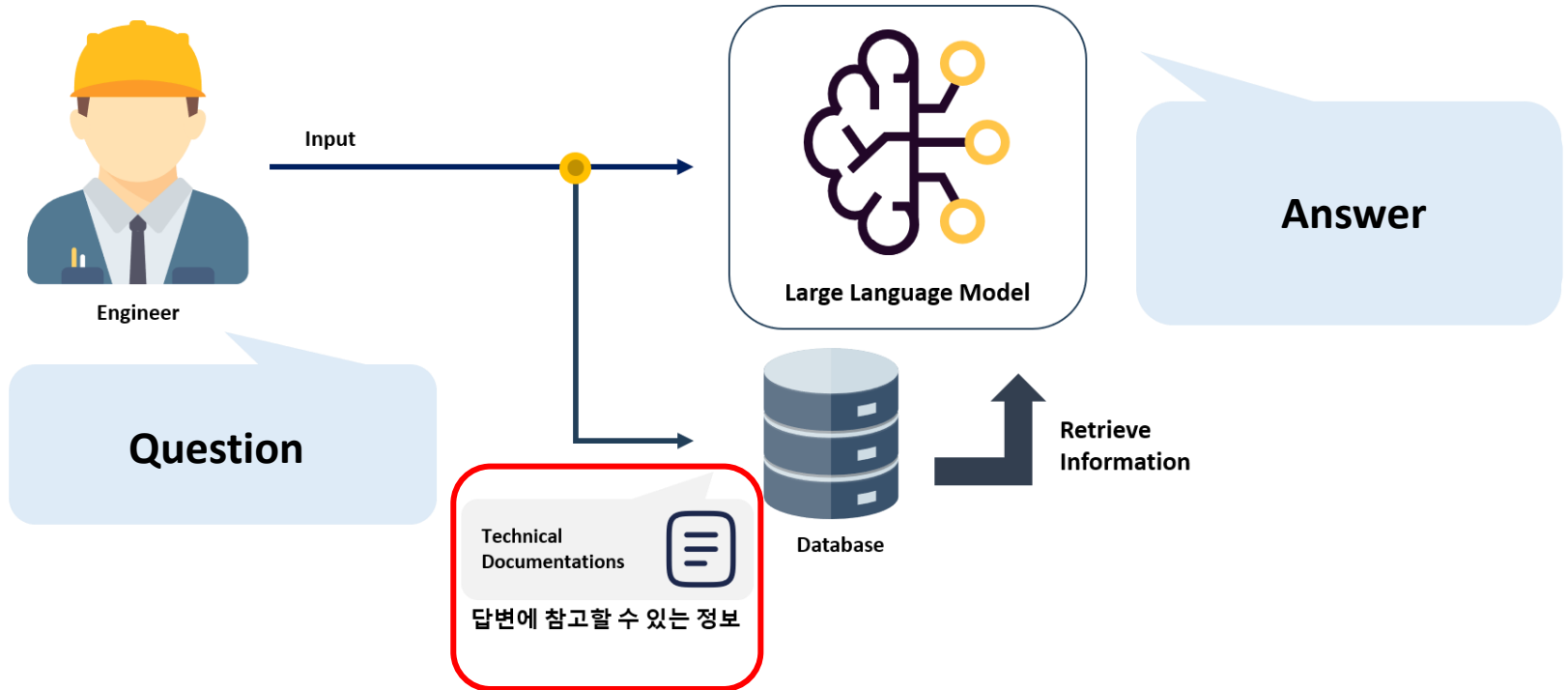
세미나 정보 보기 →

Introduction

Development of LLM

❖ What is Retrieval?

- 기존 LLM의 한계: 특정 도메인 텍스트에 대해서는 부적절한 답변 출력
- 회수 모델(Retriever)이 질문과 관련된 정보를 데이터베이스에서 탐색하여 모델에 함께 입력

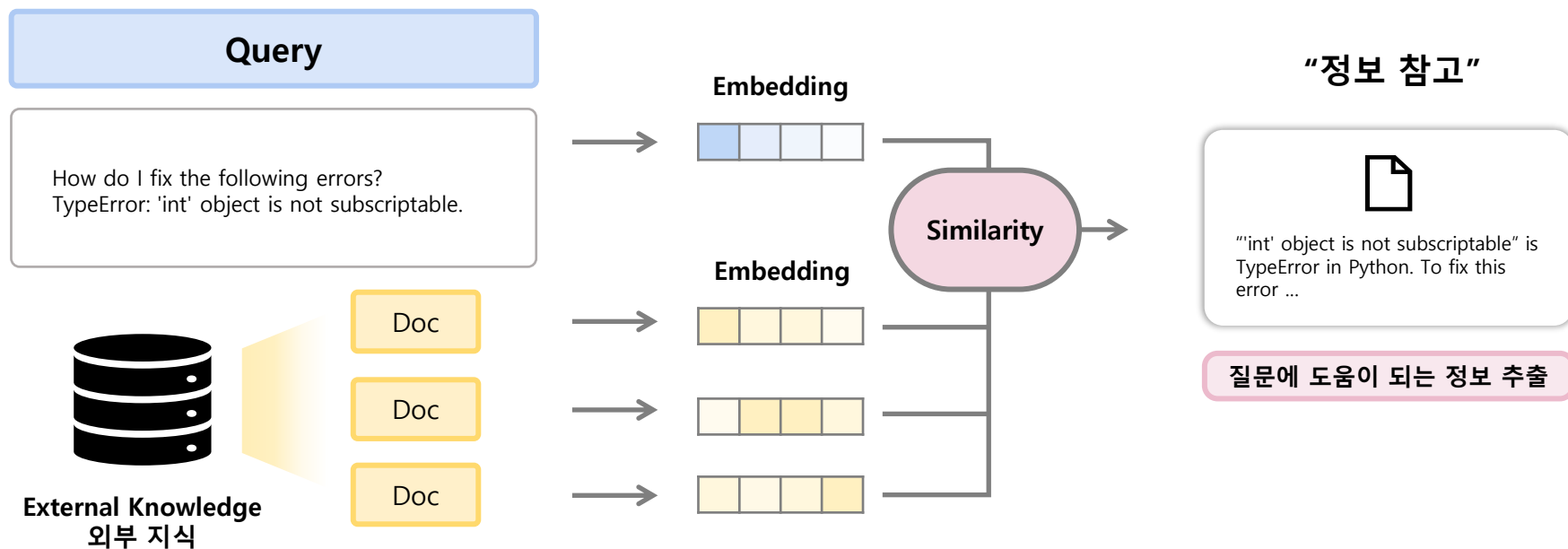


Introduction

Development of LLM

❖ What is Retrieval?

- 임베딩 벡터 간 유사도 기반으로 쿼리와 관련된 정보 추출

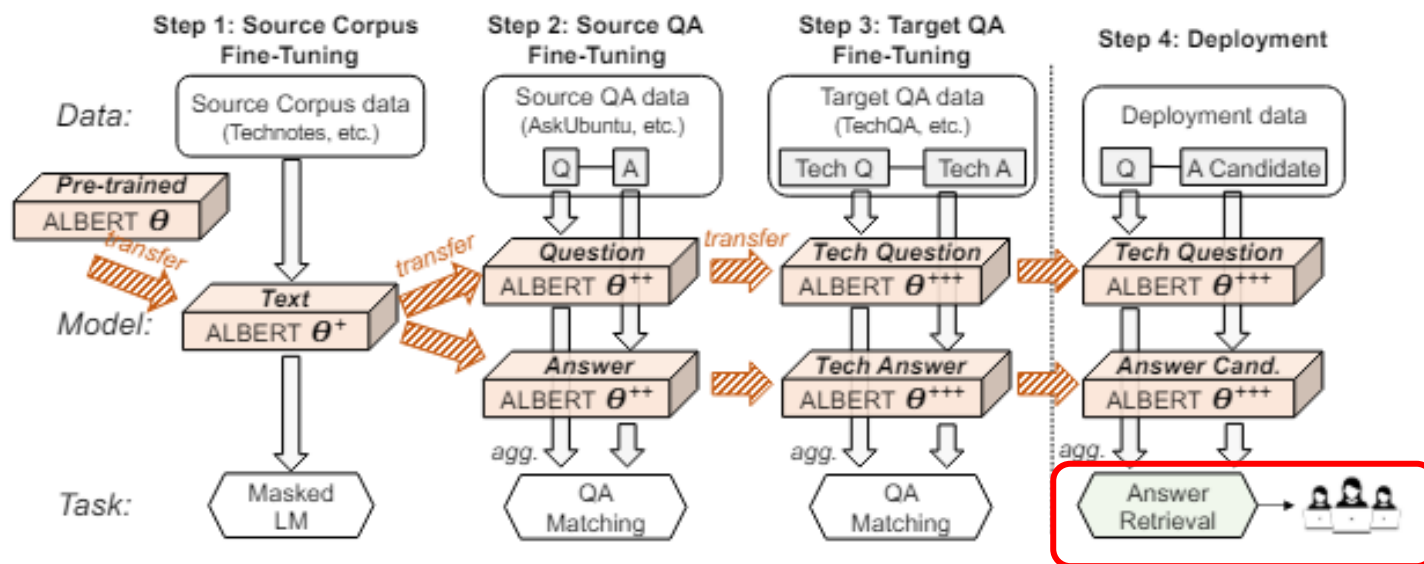


Introduction

Development of LLM

❖ What is Retrieval?

- 질문과 관련된 정보를 잘 회수하도록 모델을 학습하는 방법론들도 지속적으로 연구되고 있음
- ALBERT: BERT를 경량화 시킨 모델



Yu, W., Wu, L., Deng, Y., Mahindru, R., Zeng, Q., Guven, S., & Jiang, M. (2020, October). A technical question answering system with transfer learning. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 92-99).

Introduction

Development of LLM

❖ What is Retrieval?

- 질문과 관련된 정보를 잘 회수하도록 모델을 학습하는 방법론들도 지속적으로 연구되고 있음

C-Pack: Packed Resources For General Chinese Embeddings

Shitao Xiao[†]
stxiao@baai.ac.cn
Beijing Academy of AI
Beijing, China

Niklas Muennighoff
n.muennighoff@gmail.com
HuggingFace
Beijing, China

Zheng Liu^{†*}
zhengliu1026@gmail.com
Beijing Academy of AI
Beijing, China

Defu Lian
liandefu@ustc.edu.cn
USTC
Hefei, China

Peitian Zhang
namespace.pt@gmail.com
Renmin University of China
Beijing, China

Jian-Yun Nie
nie@iro.umontreal.ca
University of Montreal
Montreal, Canada

ABSTRACT

We introduce **C-Pack**, a package of resources that significantly advances the field of general text embeddings for Chinese. **C-Pack** includes three critical resources. 1) **C-MTP** is a massive training dataset for text embedding, which is based on the curation of vast unlabeled corpora and the integration of high-quality labeled corpora. 2) **C-MTEB** is a comprehensive benchmark for Chinese text embeddings covering 6 tasks and 35 datasets. 3) **BGE** is a family of embedding models covering multiple sizes. Our models outperform all prior Chinese text embeddings on **C-MTEB** by more than +10% upon the time of the release. We also integrate and optimize the entire suite of training methods for **BGE**. Along with our resources on general Chinese embedding, we release our data and models for English text embeddings. The English models also achieve state-of-the-art performance on the MTEB benchmark; meanwhile, our released English data is 2 times larger than the Chinese data. Both Chinese and English datasets are the largest public release of training data for text embeddings. All these resources are made publicly available at <https://github.com/FlagOpen/FlagEmbedding>.

CCS CONCEPTS

• Information systems → Retrieval models and ranking.

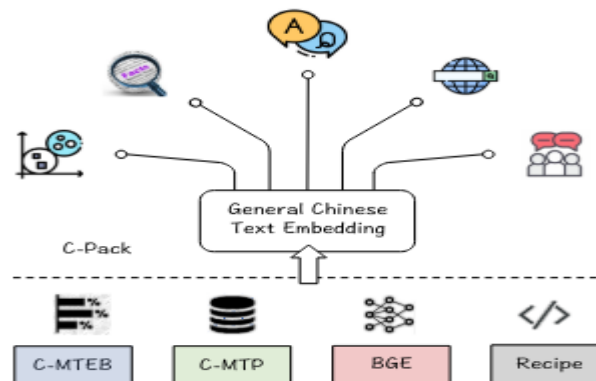


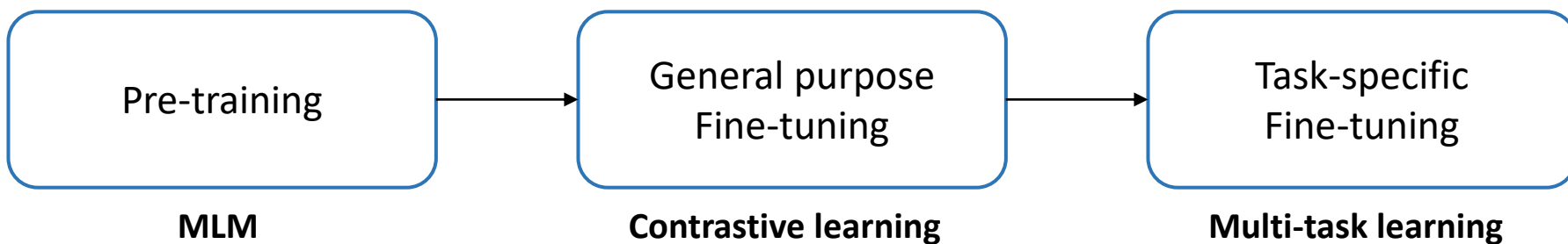
Figure 1: C-Pack presents 4 critical resources to support general Chinese embedding: C-MTEB (comprehensive evaluation benchmark), C-MTP (massive training data), BGE (powerful pre-trained models), the entire-suite of training recipe.

Introduction

Development of LLM

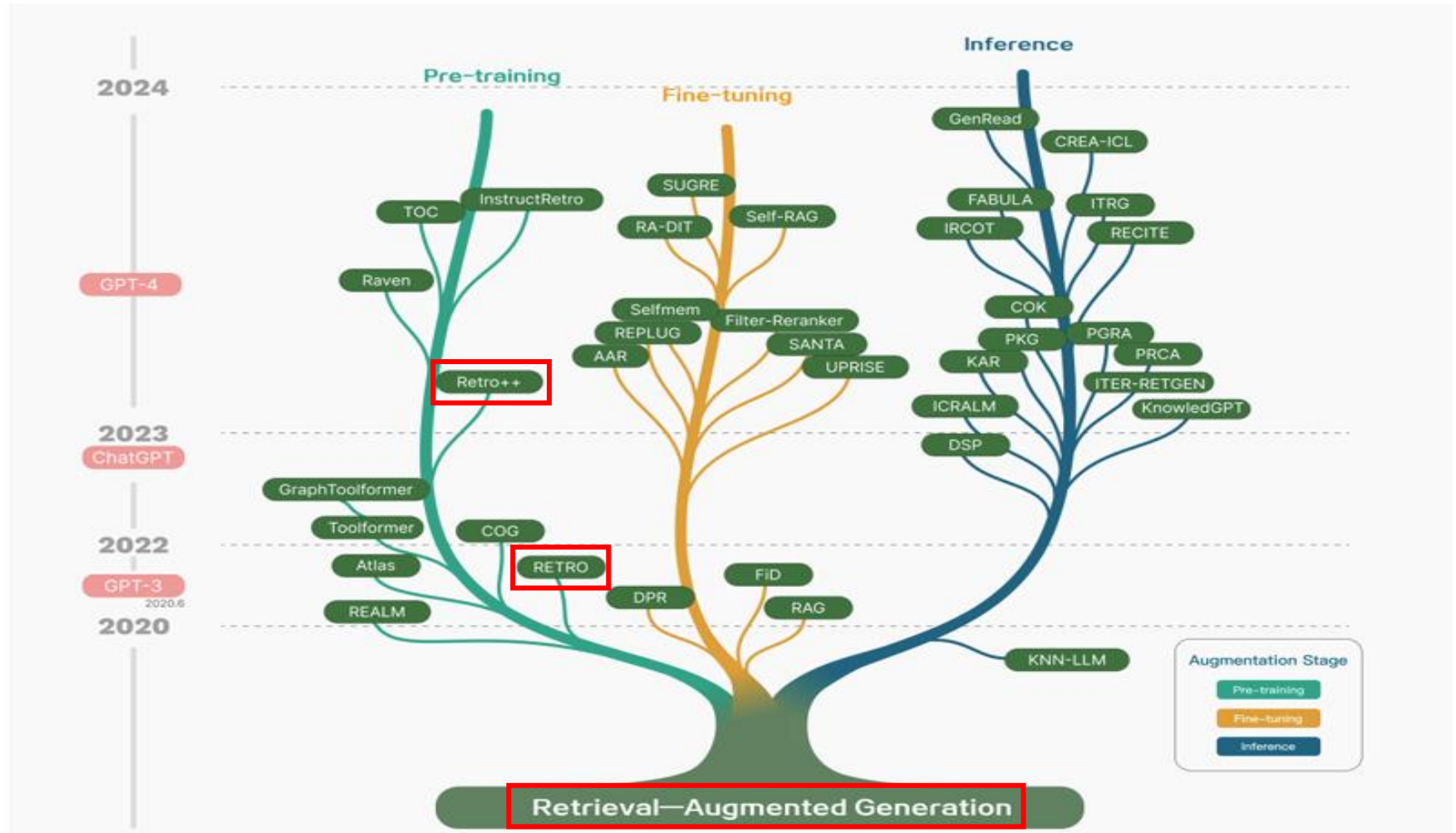
❖ What is Retrieval?

- 질문과 관련된 정보를 잘 회수하도록 모델을 학습하는 방법론들도 지속적으로 연구되고 있음
- BGE: 대용량의 임베딩 모델로 많은 연구에서 retriever로 사용 됨
 - BERT 기반으로 다양한 데이터셋 및 tasks(retrieval 포함)로 학습된 모델



Retrieval with Language Models

Retrieval with Language Models



Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Retrieval with Language Models

Paper

❖ Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2020, NeurIPS)

- RAG 파이프라인을 처음으로 제시한 연구

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

Abstract

Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory can overcome this issue, but have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) — models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, and another which can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state of the art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

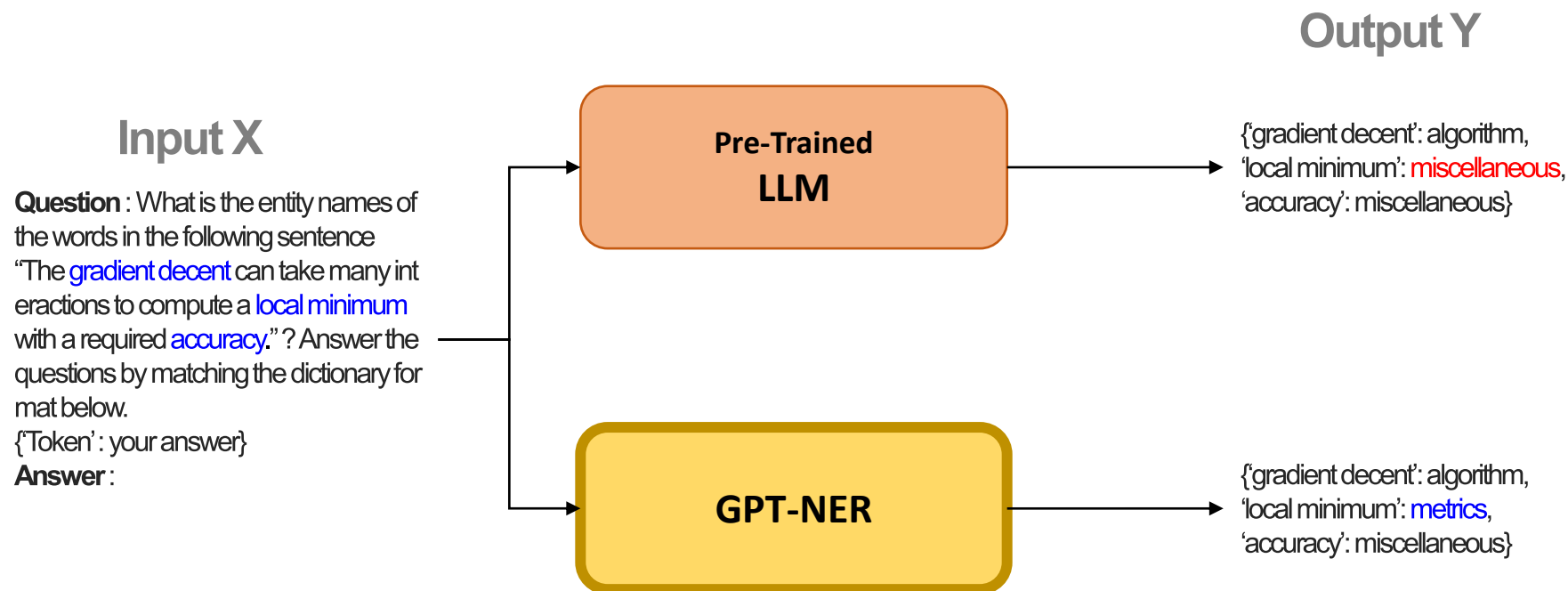
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ 연구 배경

- 기존 사전 학습된 LLM은 knowledge-intensive tasks에서 task-specific architecture 보다 낮은 성능을 보임

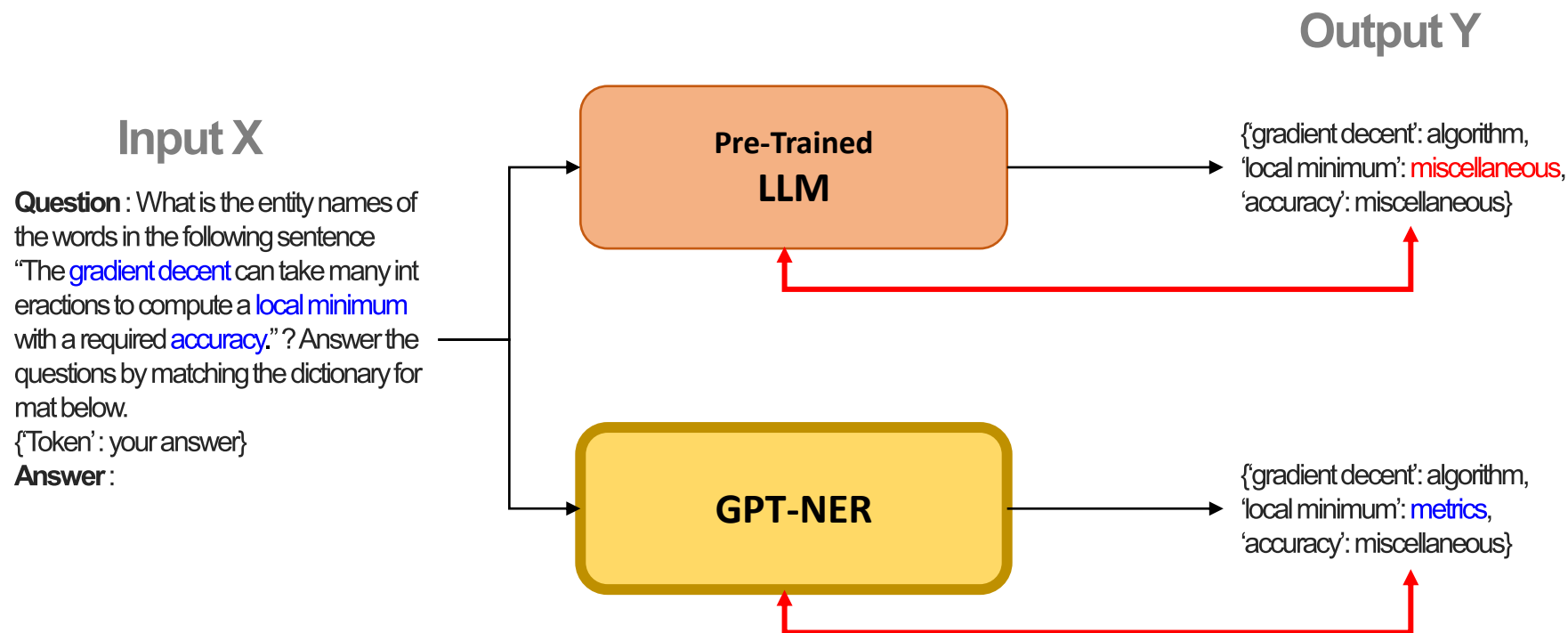


Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ 연구 배경

- 기존 사전 학습된 LLM은 knowledge-intensive tasks에서 task-specific architecture 보다 낮은 성능을 보임
- 모델의 output이 어떤 정보(파라미터)를 참고하는지 알기 어려움

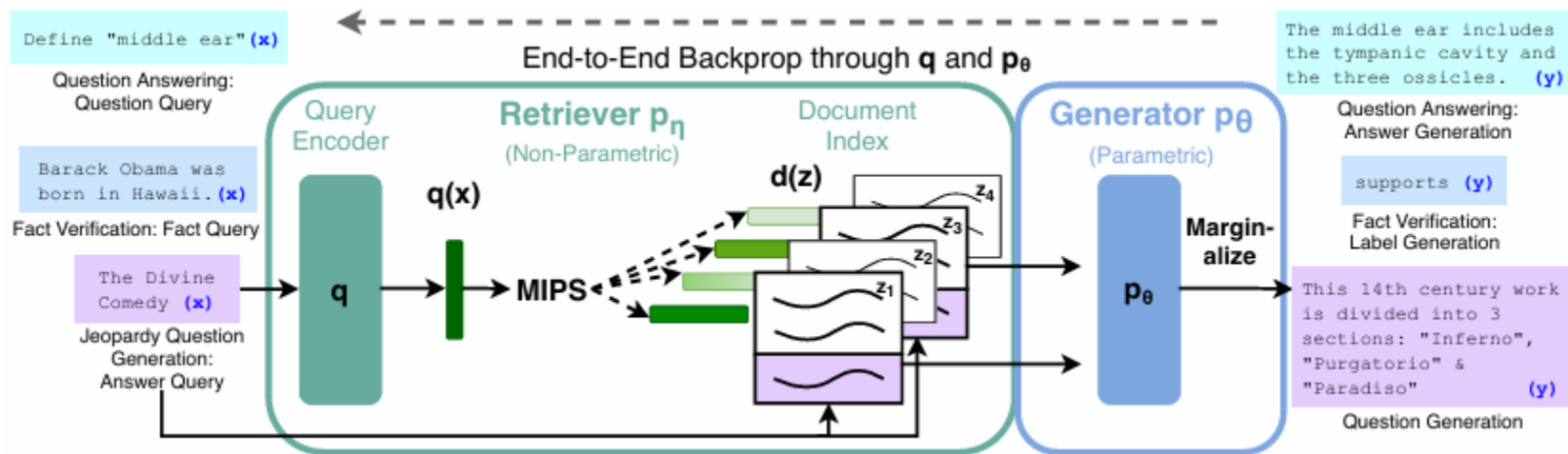


Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ RAG Framework

- Retriever, Generator 두 모델로 구성



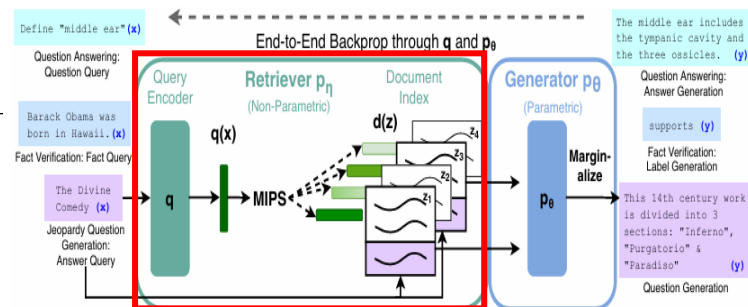
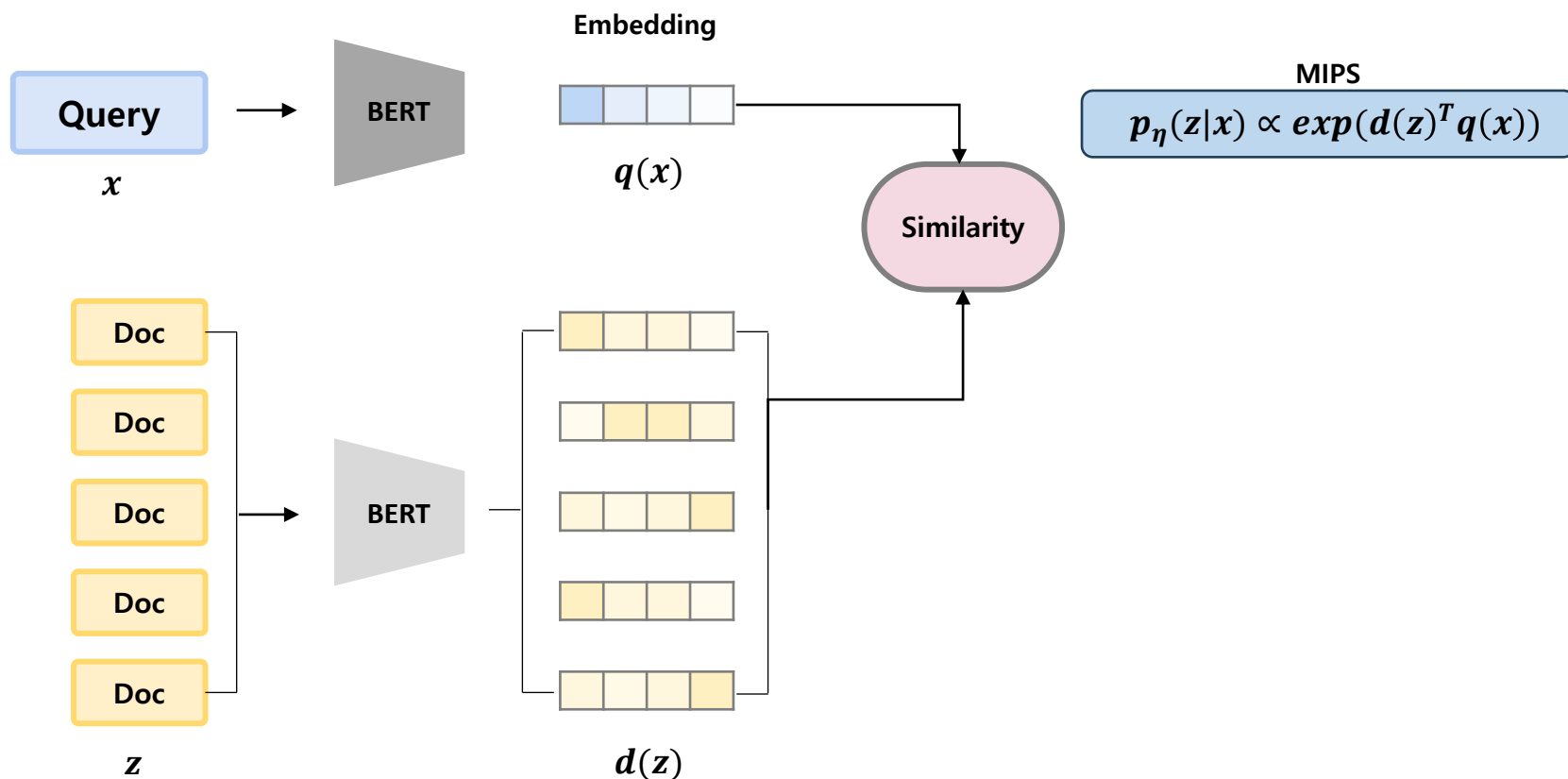
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Retriever: DPR

- 사전 학습 된 DPR 모델 사용



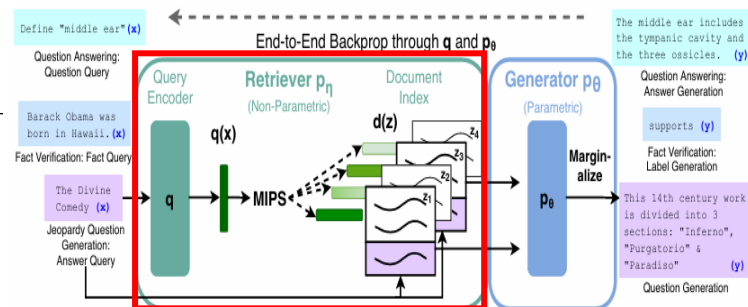
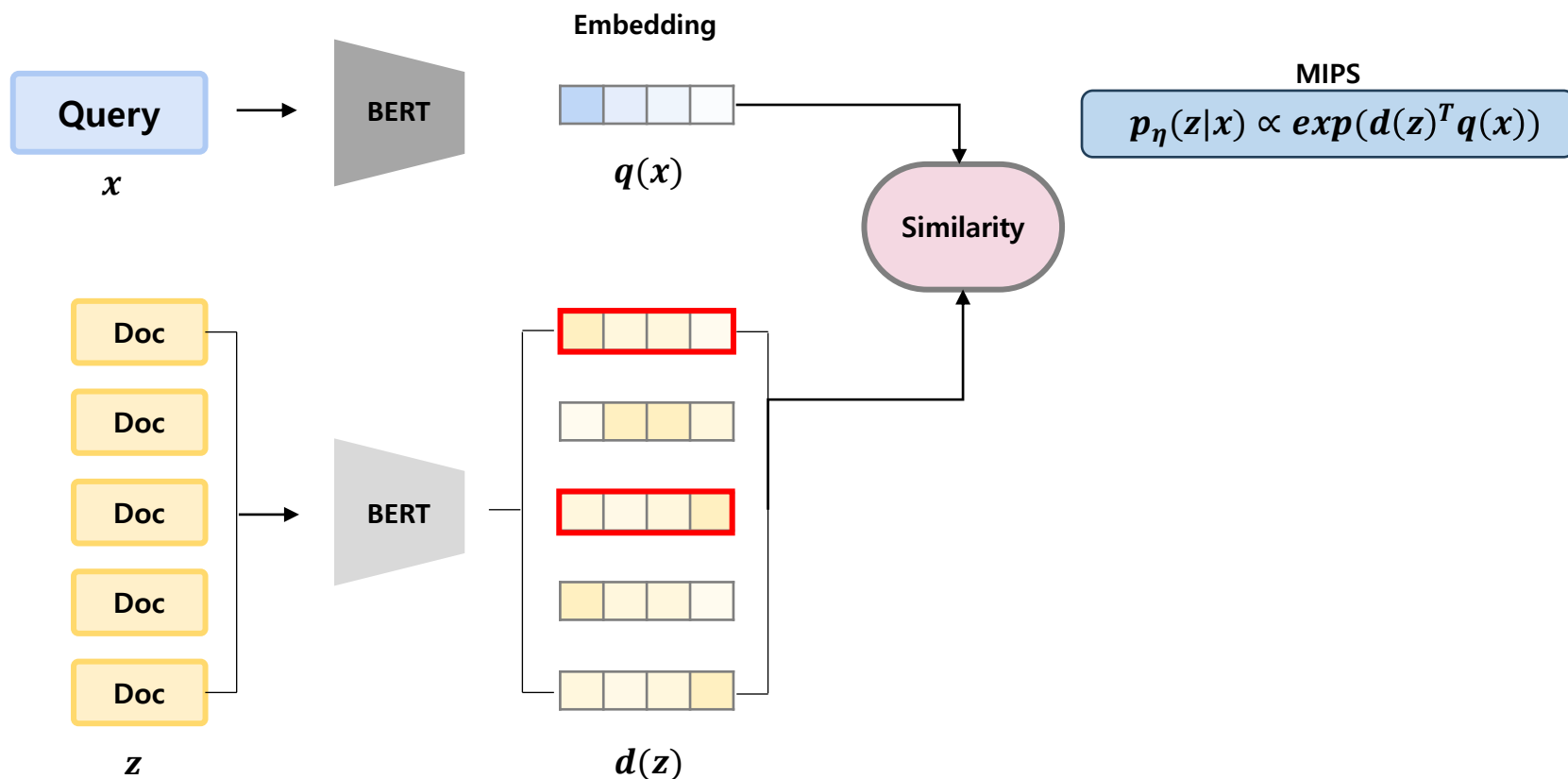
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Retriever: DPR

- 사전 학습 된 DPR 모델 사용



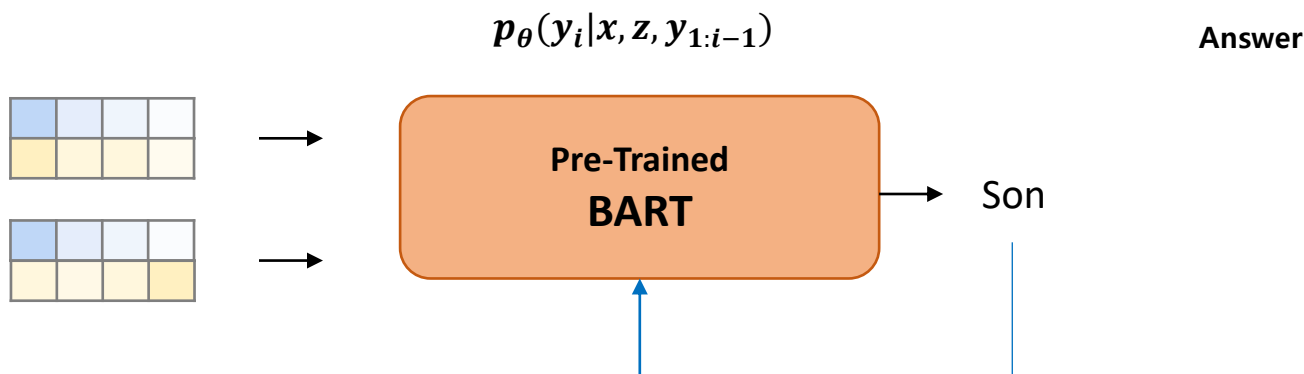
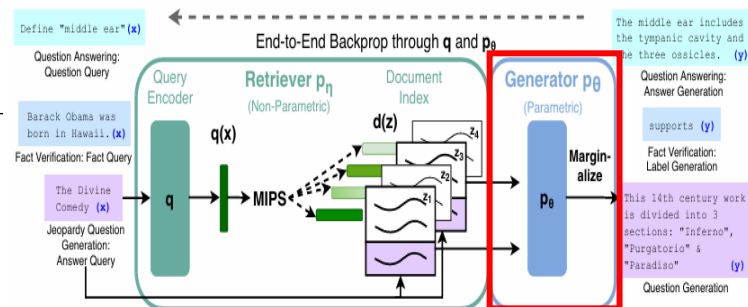
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Generator: BART

- 당시 SOTA 언어 모델 이었던 BART 모델 사용

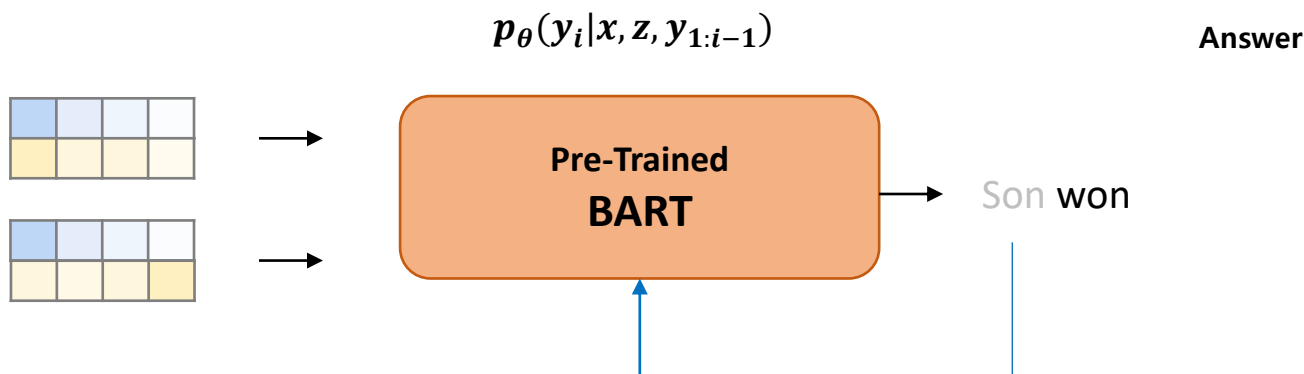
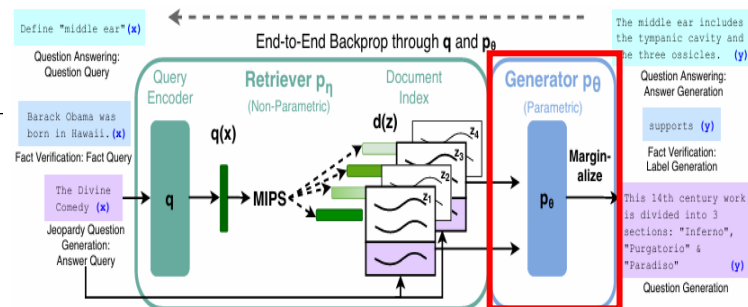


Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Generator: BART

- 당시 SOTA 언어 모델 이었던 BART 모델 사용

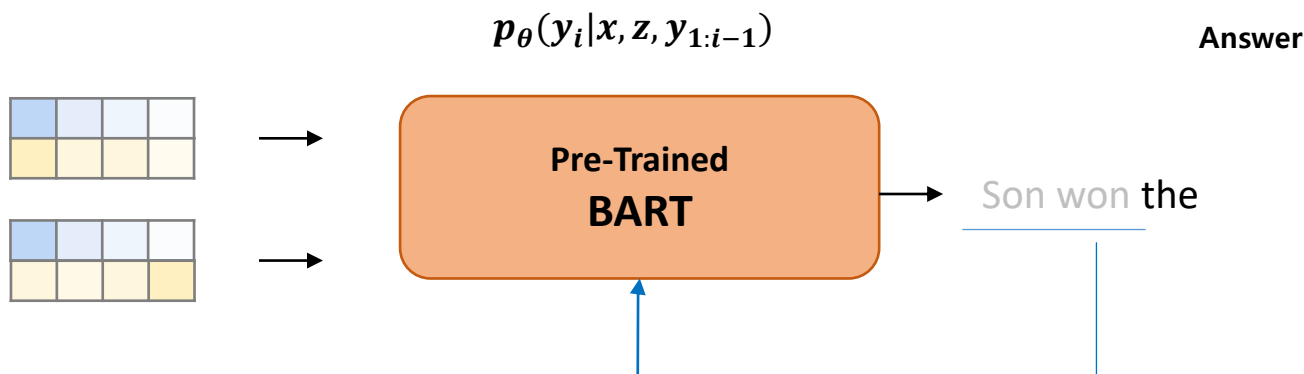
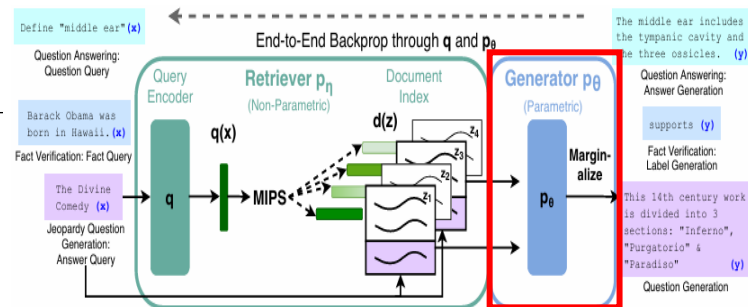


Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Generator: BART

- 당시 SOTA 언어 모델 이었던 BART 모델 사용

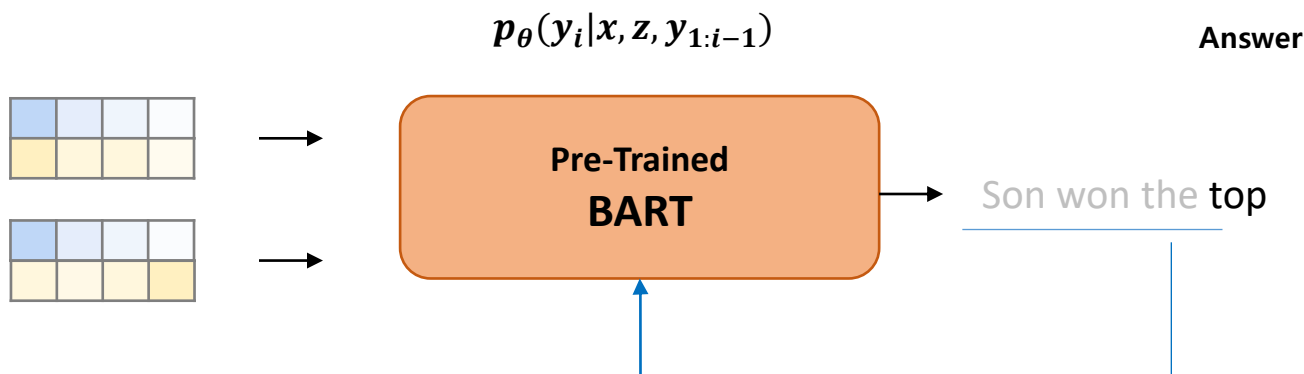
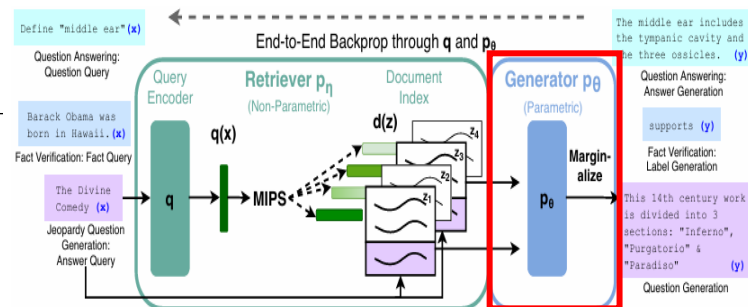


Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Generator: BART

- 당시 SOTA 언어 모델 이었던 BART 모델 사용



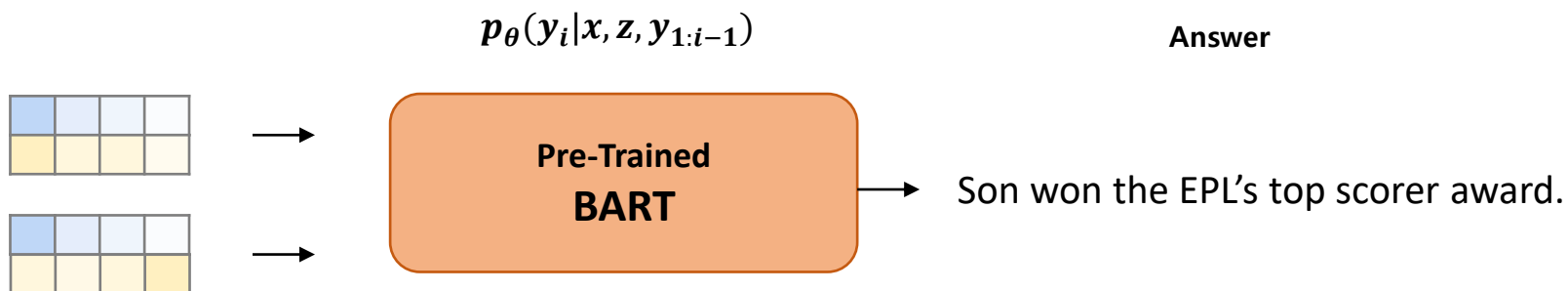
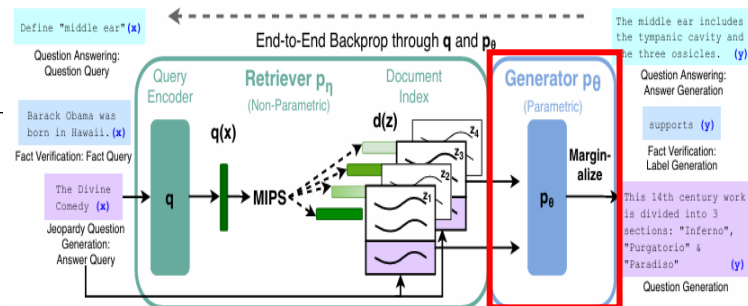
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Generator: BART

- 당시 SOTA 언어 모델 이었던 BART 모델 사용

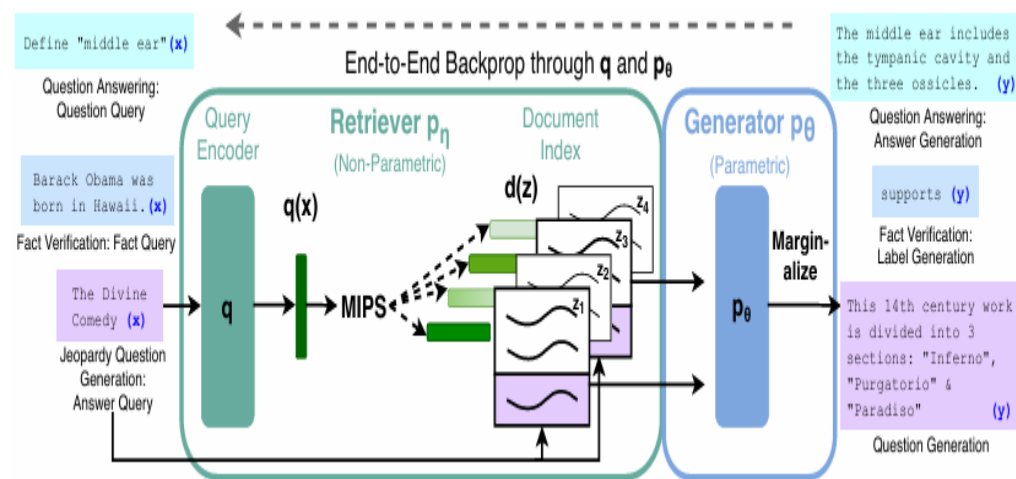


Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ RAG-Sequence Model vs RAG-Token Model

- RAG-Sequence Model : 각 output token 마다 **같은** retrieved document 사용
- RAG-Token Model : 각 output token 마다 **다른** retrieved document 사용



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

[RAG-Sequence Model]

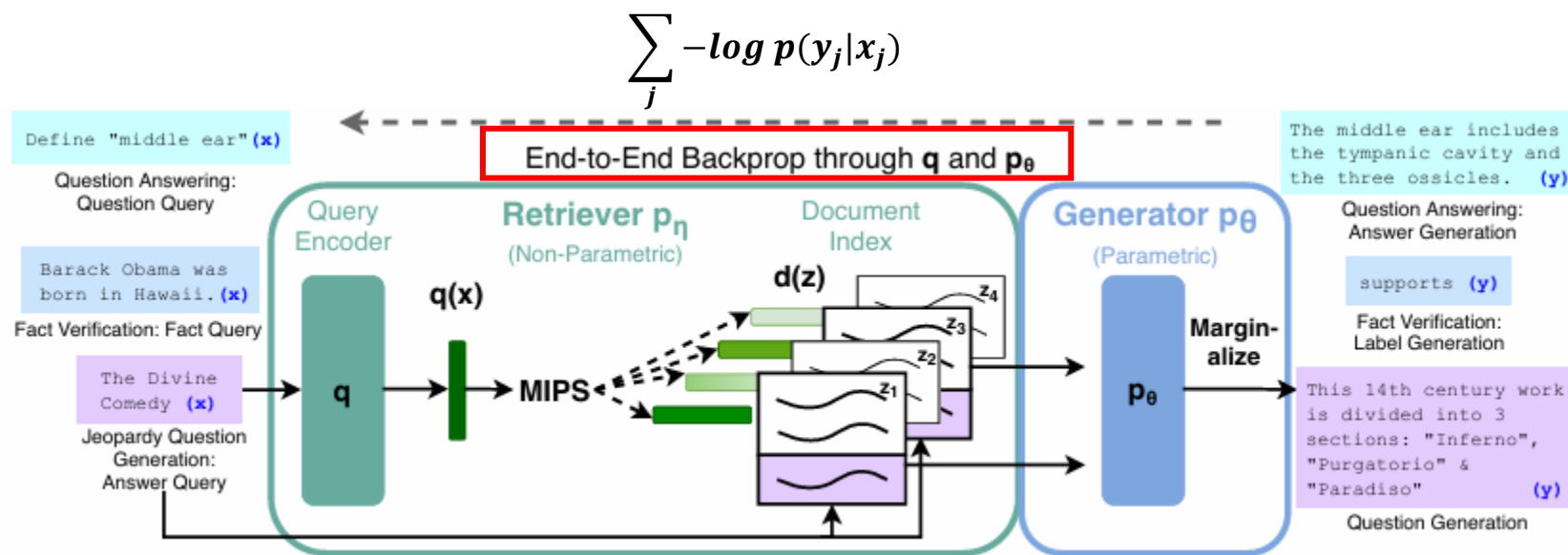
$$\prod_i^N \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})$$

[RAG-Token Model]

Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Training



Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Experiments – Main results

	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

[Open-Domain QA]

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	<u>89.5</u>
RAG-Seq.	14.7	21.4	<u>40.8</u>	<u>44.2</u>		

[Abstractive QA]

Table 4: Human assessments for the Jeopardy Question Generation Task.

	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%

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Experiments – 정성적 평가

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 Generation	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"

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

Retrieval with Language Models

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

❖ Experiments – Ablation study

Table 6: Ablations on the dev set. As FEVER is a classification task, both RAG models are equivalent.

Model	NQ	TQA Exact Match	WQ	CT	Jeopardy-QGen B-1 QB-1	MSMarco R-L B-1	FVR-3 Label Accuracy	FVR-2 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		

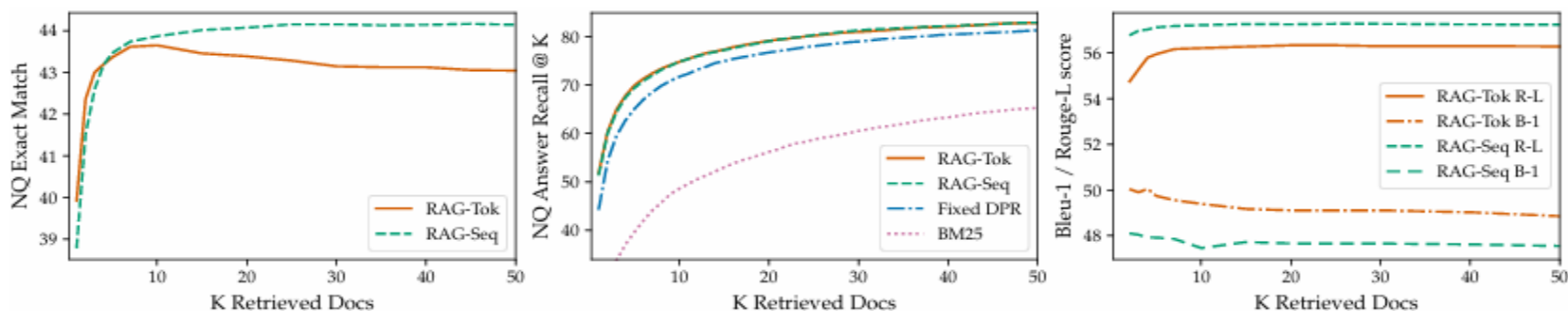


Figure 3: Left: NQ performance as more documents are retrieved. Center: Retrieval recall performance in NQ. Right: MS-MARCO Bleu-1 and Rouge-L as more documents are retrieved.

Retrieval with Language Models

Paper

❖ Improving Language Models by Retrieving from Trillions of Tokens (2022, PMLR)

- 2 trillion token database를 통해 성능을 향상시킨 RETRO(Retrieval-Enhanced Transformer) 제안

Improving Language Models by Retrieving from Trillions of Tokens

Sebastian Borgeaud^{*} Arthur Mensch^{*} Jordan Hoffmann^{*} Trevor Cai Eliza Rutherford Katie Millican
George van den Driessche Jean-Baptiste Lespiau Bogdan Damoc Aidan Clark Diego de Las Casas
Aurelia Guy Jacob Menick Roman Ring Tom Hennigan Saffron Huang Loren Maggiore Chris Jones
Albin Cassirer Andy Brock Michela Paganini Geoffrey Irving Oriol Vinyals Simon Osindero
Karen Simonyan Jack W. Rae[†] Erich Elsen[†] Laurent Sifre^{*,†}

Abstract

We enhance auto-regressive language models by conditioning on document chunks retrieved from a large corpus, based on local similarity with preceding tokens. With a 2 trillion token database, our Retrieval-Enhanced Transformer (RETRO) obtains comparable performance to GPT-3 and Jurassic-1 on the Pile, despite using $25\times$ fewer parameters. After fine-tuning, RETRO performance translates to downstream knowledge-intensive tasks such as question answering. RETRO combines a frozen BERT retriever, a differentiable encoder and a chunked cross-attention mechanism to predict tokens based on an order of magnitude more data than what is typically consumed during training. We typically train RETRO from scratch, yet can also rapidly RETROfit pre-trained transformers with retrieval and still achieve good performance. Our work opens up new avenues for improving language models through explicit memory at unprecedented scale.

rameters. Transformers have been scaled from millions of parameter models in seminal work to over hundred billion parameters (Brown et al., 2020), which has led to models that do well on a wide array of tasks in a zero or few-shot formulation. Increasing model size predictably improves performance on downstream tasks (Kaplan et al., 2020). Increasing the number of parameters is beneficial in two ways: additional computations at training and inference time, and increased memorization of the training data.

We endeavor to decouple these improvements, by efficiently augmenting language models with a massive-scale memory without significantly increasing computations. Specifically, we suggest retrieval from a large text database as a complementary path to scaling language models. Instead of increasing the model size and training on more data, we equip models with the ability to directly access a large database to perform predictions—a semi-parametric approach. At a high level, our Retrieval Transformer (RETRO) model splits the input sequence into chunks and retrieves text similar to the previous chunk to improve the predictions in the current chunk. Existing retrieval for language modelling work only

Retrieval with Language Models

Improving Language Models by Retrieving from Trillions of Tokens

❖ 연구 배경

- Computation resource를 증가시키지 않으면서 대규모 text database를 통해 언어 모델의 **효율성과 성능**을 향상시키고자 함

**Pre-Trained
GPT-1**

**Pre-Trained
GPT-2**

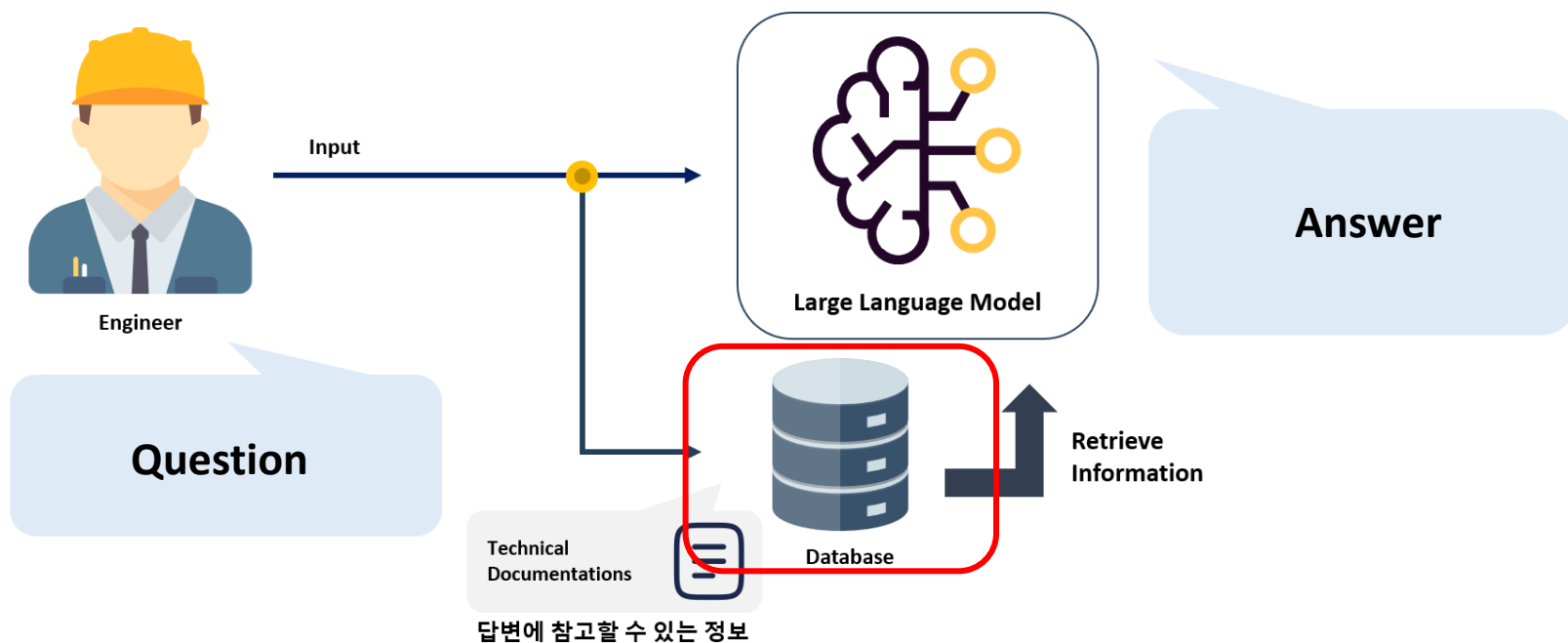
**Pre-Trained
GPT-3**

Retrieval with Language Models

Improving Language Models by Retrieving from Trillions of Tokens

❖ 연구 배경

- Computation resource를 증가시키지 않으면서 대규모 text database를 통해 언어 모델의 **효율성과 성능**을 향상시키고자 함

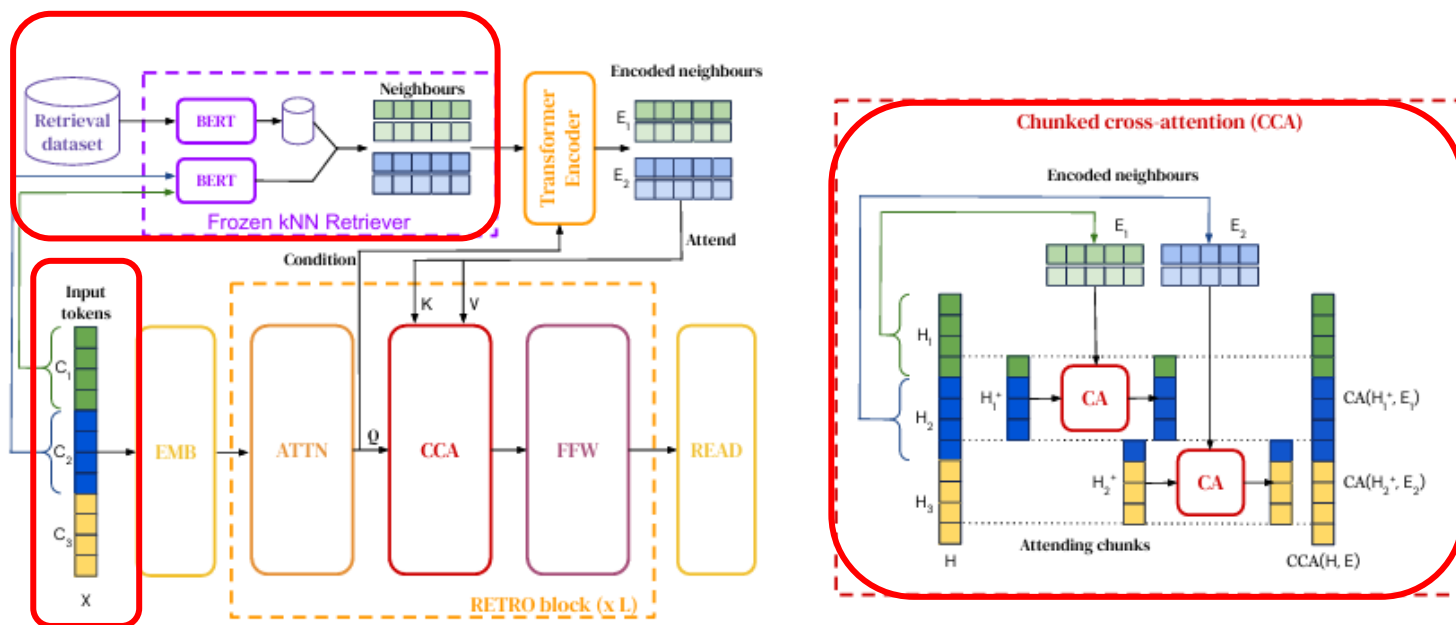


Retrieval with Language Models

Improving Language Models by Retrieving from Trillions of Tokens

❖ RETRO Framework

- Chunk: 특정 개수의 token을 묶어 입력 값으로 사용
- Continuation: Retrieval dataset에서 특정 passage 바로 다음에 나오는 연속된 passage
- Chunked Cross-Attention: Chunk 단위의 input에 대해서 retrieval한 정보와 cross-attention 수행

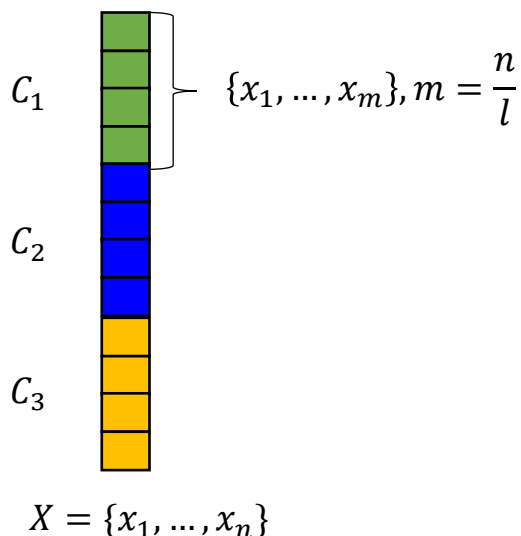
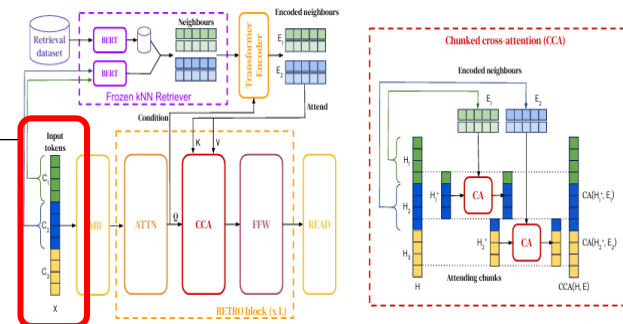


Retrieval with Language Models

Improving Language Models by Retrieving from Trillions of Tokens

❖ Chunk

- 특정 개수의 token을 묶어 입력 값으로 사용
 - Ex. Samsung released a new phone this year with an interpreter function.
- 연산 부담을 감소시킴



$$L(X|\theta, D) \triangleq \sum_{i=1}^n l_{\theta}(x_i | (x_j)_{j < i}, (\text{RET}_D(C_v))_{v < u_i})$$

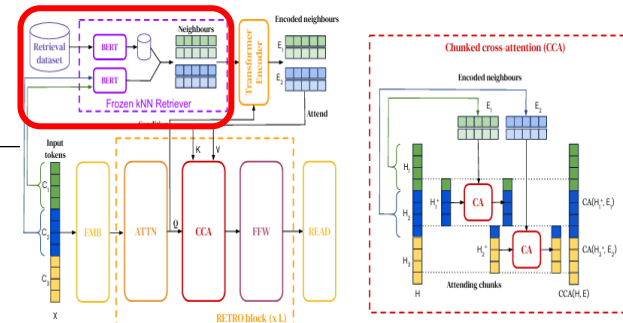
이전 chunk들이
retrieval 하는 정보

Retrieval with Language Models

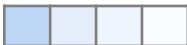
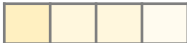
Improving Language Models by Retrieving from Trillions of Tokens

❖ Continuation

- Retrieval dataset에서 특정 passage 바로 다음에 나오는 연속된 passage 사용
- Key: frozen BERT Embeddings
- Value: Raw chunks of text tokens



Key에 해당하는 original text

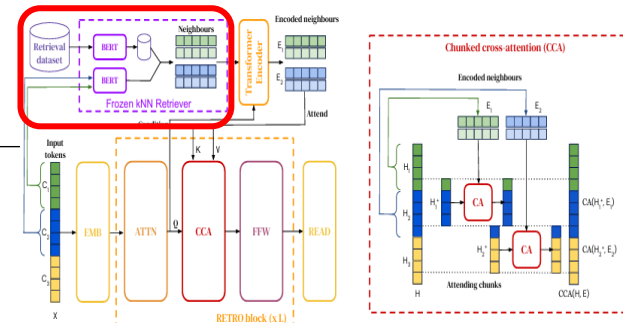
Retrieval Dataset	Key (BERT embedding)	Value (Text / Neighbor and continuation chunks)	
		Dune is a 2021 American epic science fiction film directed by Denis Villeneuve	Neighbor (N)
		It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert	Continuation (F)
		Dune is a 1965 science fiction novel by American author Frank Herbert	Neighbor (N)
		Originally published as two separate serials in Analog magazine	Continuation (F)

Retrieval with Language Models

Improving Language Models by Retrieving from Trillions of Tokens

❖ Continuation

- Retrieval dataset에서 특정 passage 바로 다음에 나오는 연속된 passage 사용
- Key: frozen BERT Embeddings
- Value: Raw chunks of text tokens



Original text 다음 text

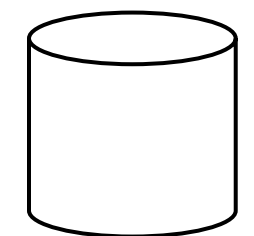
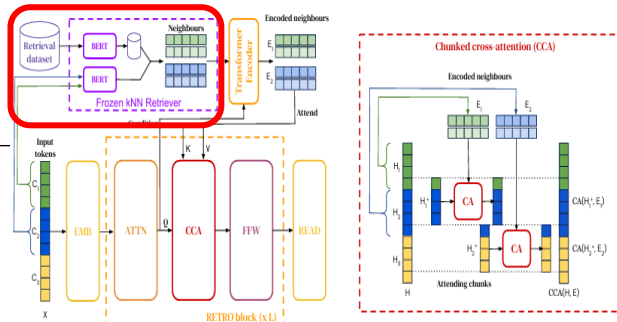
Retrieval Dataset	Key (BERT embedding)	Value (Text / Neighbor and continuation chunks)	
		Dune is a 2021 American epic science fiction film directed by Denis Villeneuve	Neighbor (N)
		It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert	Continuation (F)
		Dune is a 1965 science fiction novel by American author Frank Herbert	Neighbor (N)
		Originally published as two separate serials in Analog magazine	Continuation (F)

Retrieval with Language Models

Improving Language Models by Retrieving from Trillions of Tokens

❖ Continuation

- Retrieval dataset에서 특정 passage 바로 다음에 나오는 연속된 passage 사용
- Key: frozen BERT Embeddings
- Value: Raw chunks of text tokens



Retrieval Dataset

$[N, F]$



Input query chunk

C

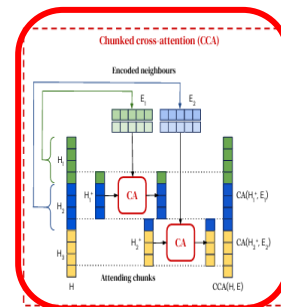
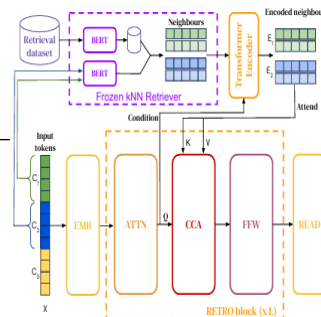
$$d(C, N) = ||\mathbf{BERT}(C) - \mathbf{BERT}(N)||_2^2$$

Retrieval with Language Models

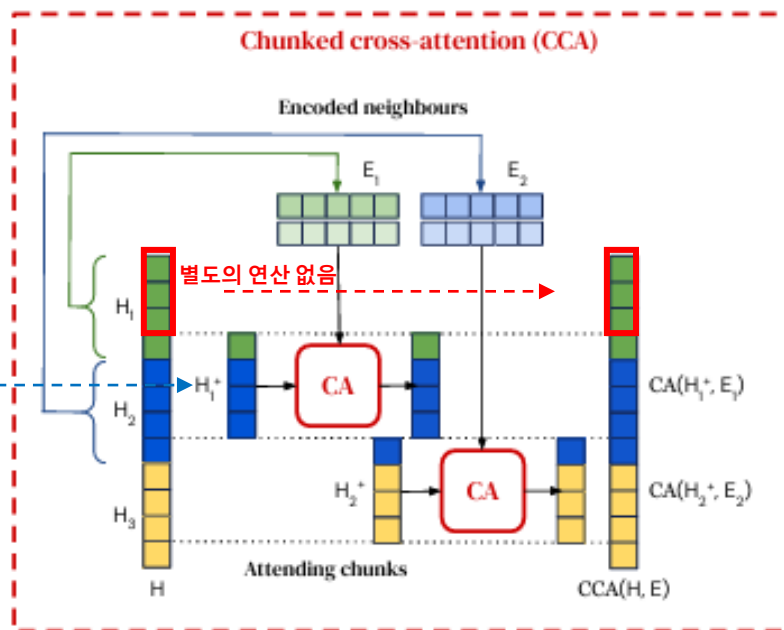
Improving Language Models by Retrieving from Trillions of Tokens

❖ Chunked Cross-Attention(CCA)

- Chunk 단위의 input에 대해서 retrieval한 정보와 cross-attention 수행
- 보다 낮은 연산 비용으로 retrieval한 정보와 attention 수행
- Self-attention을 통해 마지막 token은 이전 token(chunk)들의 정보를 가지고 있음
 - Chunk간의 causality 보존



Causality 보존하기 위해서
마지막 token을 다음 chunk에 합침

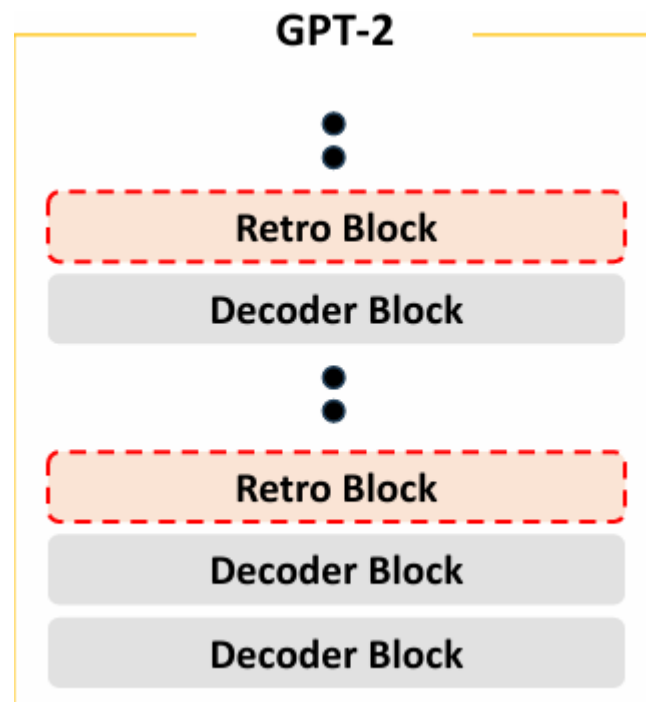
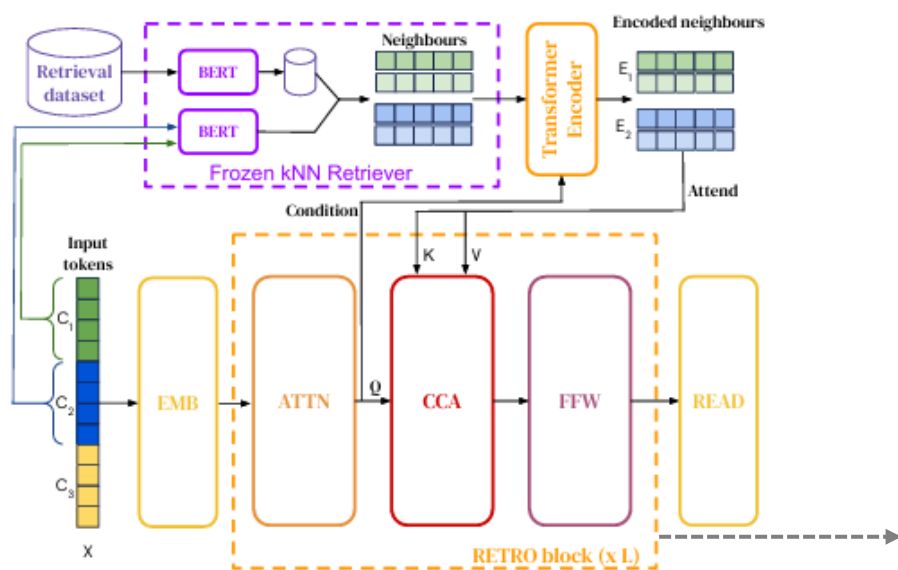


Retrieval with Language Models

Improving Language Models by Retrieving from Trillions of Tokens

❖ Decoder block

- 최초로 decoder-only 모델 구조(GPT-2)를 generator로 사용
- GPT-2 기반의 decoder에 retro block을 중간 중간에 삽입



Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022, June). Improving language models by retrieving from trillions of tokens. In International conference on machine learning (pp. 2206-2240). PMLR.

Retrieval with Language Models

Improving Language Models by Retrieving from Trillions of Tokens

❖ Experiments

- Retrieval dataset을 키움에 따라 우수한 성능 도출
- FID: 사전 학습된 T5 모델이 encoder output에 보다 많이 의존

Table 2. Perplexities on Wikitext103. When using the Wikipedia dataset for retrieval, RETRO performs similarly to our implementation of k NN-LM. As we scale the retrieval dataset, RETRO performs better, in part due to exploiting chunk-level leakage.

Model / Database	#tokens	#keys	Valid	Test
Adapt. inputs / -	-	-	17.96	18.65
SPALM / Wiki	3B	3B	17.20	17.60
k NN-LM / Wiki	3B	3B	16.06	16.12
Megatron / -	-	-	-	10.81
Baseline / -	-	-	21.53	22.96
k NN-LM / Wiki	4B	4B	18.52	19.54
RETRO / Wiki	4B	0.06B	18.46	18.97
RETRO / C4	174B	2.9B	12.87	10.23
RETRO / MT 1%	18B	0.8B	18.92	20.33
RETRO / MT 10%	179B	4B	13.54	14.95
RETRO / MT 100%	1.8T	28B	3.21	3.92

Table 3. Question answering results on Natural Questions.

Model	Test Accuracy
REALM (Gua et al., 2020)	40.4
DPR (Karpukhin et al., 2020)	41.5
RAG (Lewis et al., 2020)	44.5
EMDR ² (Sachan et al., 2021)	52.5
FID (Izacard & Grave, 2021)	51.4
FID + Distill. (Izacard et al., 2020)	54.7
Baseline 7B (closed book)	30.4
RETRO 7.5B (DPR retrieval)	45.5

Retrieval with Language Models

Improving Language Models by Retrieving from Trillions of Tokens

❖ Experiments

- RETROfit: 사전 학습된 GPT-2에서 decoder block 파라미터는 고정 / retro block 파라미터만 업데이트

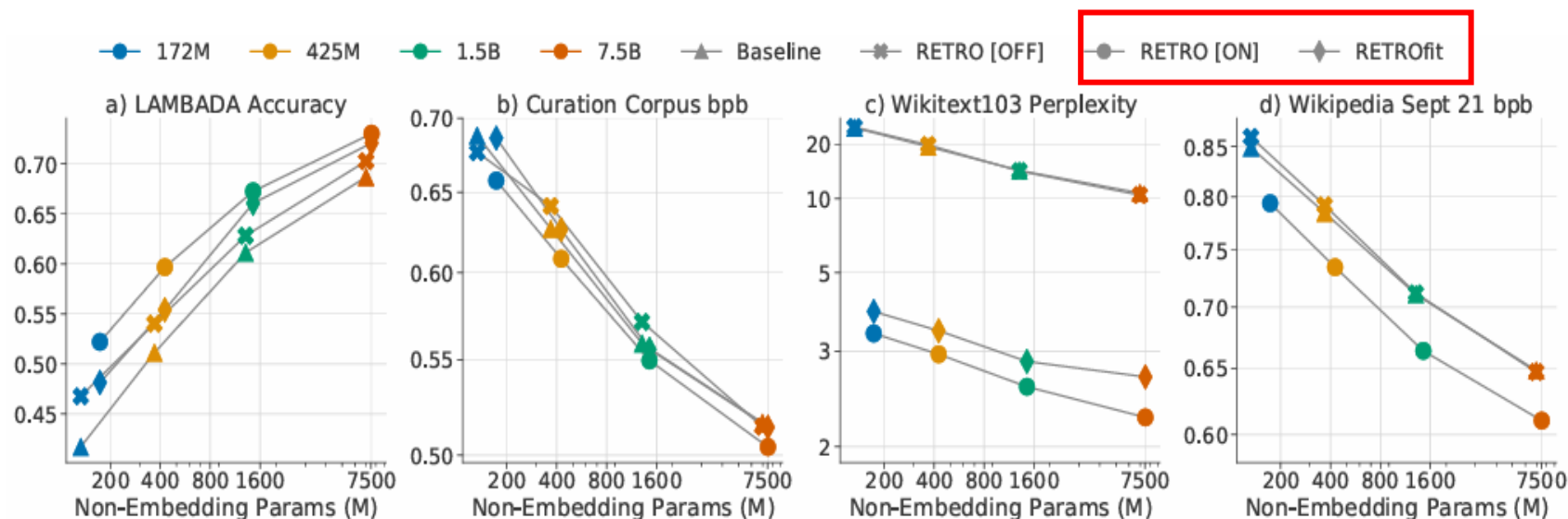
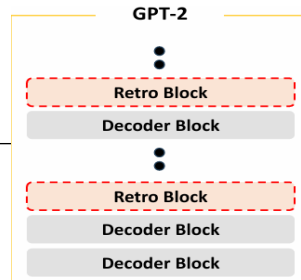


Figure 3. Scaling with respect to model size. (a) LAMBADA top-1 accuracy. (b) Evaluation loss on curation corpus. (c) Perplexity on Wikitext103 valid. (d) Bits-per-byte on selected Wikipedia articles from September 2021.

Retrieval with Language Models

Paper

❖ Shall We Pretrain Autoregressive Language Models with Retrieval? (2023, EMNLP)

- 기존 RETRO는 코드가 공개되어 있지 않아 NVIDIA 및 대학 연구진들이 구현 및 추가 실험 진행
- RETRO++: RETRO에 추가적인 fine-tuning을 적용하여 구현

Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study

Boxin Wang*^{†1}

Wei Ping*^{†2}

Peng Xu*²

Lawrence McAfee²

Zihan Liu²

Mohammad Shoeybi²

Yi Dong²

Oleksii Kuchaiev²

Bo Li¹

Chaowei Xiao^{2,3}

Anima Anandkumar²

Bryan Catanzaro²

Abstract

Large decoder-only language models (LMs) can be largely improved in terms of perplexity by retrieval (*e.g.*, RETRO), but its impact on text generation quality and downstream task accuracy is unclear. Thus, it is still an open question: *shall we pretrain large autoregressive LMs with retrieval?* To answer it, we perform a comprehensive study on a *scalable pre-trained* retrieval-augmented LM (*i.e.*, RETRO) compared with standard GPT and retrieval-augmented GPT incorporated at fine-tuning or inference stages. We first provide the recipe to reproduce RETRO up to 9.5B parameters while retrieving a text corpus with 330B tokens. Based on that, we have the following novel findings: *i)* RETRO outperforms GPT on text generation with much less degeneration (*i.e.*, repetition), moderately higher factual accuracy, and slightly lower toxicity with a nontoxic retrieval database. *ii)* On the LM Evaluation Harness benchmark, RETRO largely outperforms GPT on knowledge-intensive tasks, but is on par with GPT on other tasks. Furthermore, we introduce a simple variant of the model, RETRO++, which largely improves open-domain QA results of original RETRO (*e.g.*, EM score +8.6 on Natural Question) and significantly outperforms retrieval-augmented GPT in both fine-tuning and zero-shot evaluation settings. Our findings highlight the promising direction of pretraining autoregressive LMs with retrieval as future foundation models. We release our code and model at: <https://github.com/NVIDIA/Megatron-LM/blob/main/tools/retro/README.md>.

2020), BART (Lewis et al., 2020a)), have obtained state-of-the-art results for various NLP tasks. Among them, the autoregressive LMs like GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) demonstrate noticeable in-context learning ability and excellent long-form text generation results. Due to its importance, the community has spent considerable efforts to scale up such autoregressive generative LMs with more data and parameters and observed significant breakthroughs in a variety of real-world applications (*e.g.*, Brown et al., 2020), including open-ended text generation and various downstream tasks (*e.g.*, question answering). The successful public examples include GPT-3 (w/ 170B parameters) (Brown et al., 2020), Gopher (280B) (Rae et al., 2021), Megatron-Turing (530B) (Smith et al., 2022), and PaLM (540B) (Chowdhery et al., 2022).

Although large-scale autoregressive LMs have achieved huge successes, they also suffer from several weaknesses. First, it requires a huge number of model parameters to memorize the world knowledge, which makes it costly for deployment. Second, it is difficult to safeguard factual accuracy, which may provide users with incorrect information (Lee et al., 2022). Third, it is expensive to update the model knowledge learned during pre-training with up-to-date facts (Meng et al., 2022), yielding outdated answers (Lewis et al., 2020b).

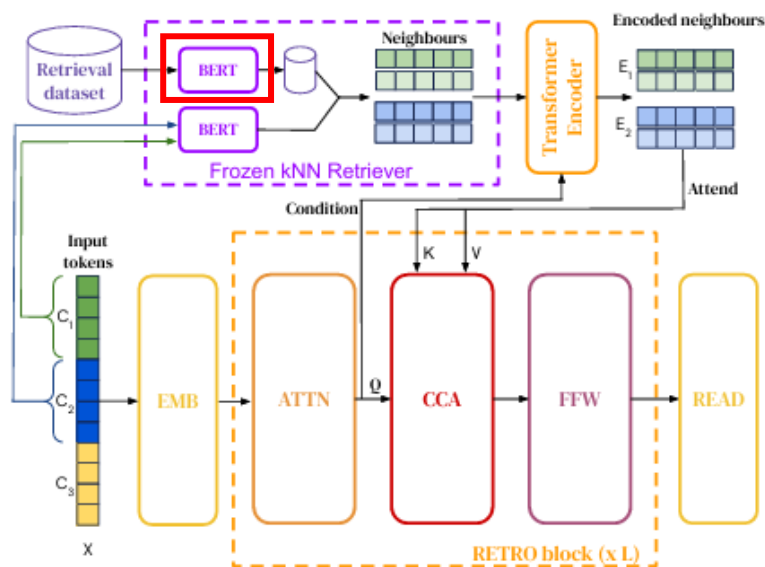
To mitigate the problems above, one line of research proposes to improve language models with retrieval. The retrieval process can be inte-

Retrieval with Language Models

Shall We Pretrain Autoregressive Language Models with Retrieval?

❖ RETRO++

- 사전 학습된 RETRO 모델에 retrieval task에 대해 fine-tuning 수행
- Retrieval된 정보들 중 최상위 정보는 Template A에 추가하고 나머지는 encoder에 저장



Template A

title: {title}source: {source}
\n question: {question}
\n answer: {answer}

Retrieval with Language Models

Shall We Pretrain Autoregressive Language Models with Retrieval?

❖ Experiments

- 전체적으로 가이드라인 실험들을 진행

Metrics	Small		Medium		XL		XXL	
	GPT	RETRO	GPT	RETRO	GPT	RETRO	GPT	RETRO
Repetition %	2.86%	2.26%	1.70%	1.50%	1.44%	0.96%	1.40%	1.12%
Self-BLEU	0.29	0.3	0.29	0.3	0.29	0.29	0.31	0.31
Zipf Coefficient	0.98	0.98	0.96	0.98	0.97	0.98	0.96	0.96

Table 3: Automatic evaluation on text generation quality for RETRO and GPT across different sizes.

Models	Retrieval Database	Exp. Max. Toxicity (↓)			Toxicity Prob. (↓)		
		Full	Toxic	Nontoxic	Full	Toxic	Nontoxic
GPT	-	0.44	0.64	0.39	37%	74%	27%
RETRO (top- $N = 2$, top- $K = 2$)	Pretraining	0.46	0.66	0.40	40%	76%	30%
RETRO (top- $N = 5$, top- $K = 2$)	Pretraining	0.46	0.66	0.40	39%	77%	29%
RETRO (top- $N = 10$, top- $K = 2$)	Pretraining	0.46	0.66	0.40	39%	76%	29%
RETRO (top- $N = 2$, top- $K = 2$)	Wiki	0.43	0.64	0.38	35%	73%	25%
RETRO (top- $N = 5$, top- $K = 2$)	Wiki	0.43	0.64	0.38	35%	71%	26%
RETRO (top- $N = 10$, top- $K = 2$)	Wiki	0.43	0.64	0.38	35%	71%	26%

Table 5: Evaluation of LM toxicity for GPT (XL) and RETRO (XL). Model toxicity is evaluated on REALTOXICITYPROMPTS. **Full** refers to the full set of prompts, **Toxic** and **Nontoxic** refer to the toxic and nontoxic subsets of prompts. ↓ means the lower, the better. RETRO can filter from top- N nearest neighbors and select the top- K nontoxic neighbors for retrieval.

Wang, B., Ping, W., Xu, P., McAfee, L., Liu, Z., Shoeybi, M., ... & Catanzaro, B. (2023). Shall we pretrain autoregressive language models with retrieval? a comprehensive study. arXiv preprint arXiv:2304.06762.

Retrieval with Language Models

Shall We Pretrain Autoregressive Language Models with Retrieval?

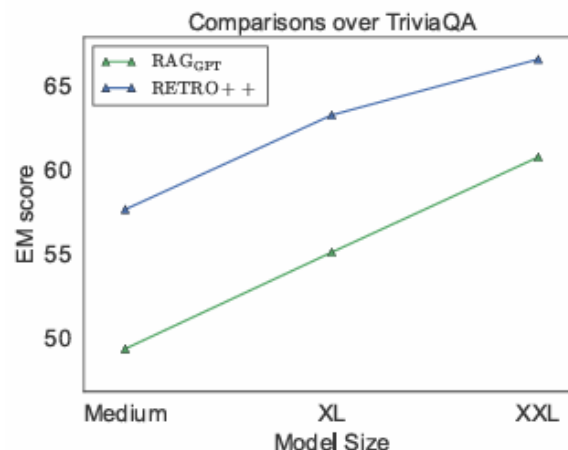
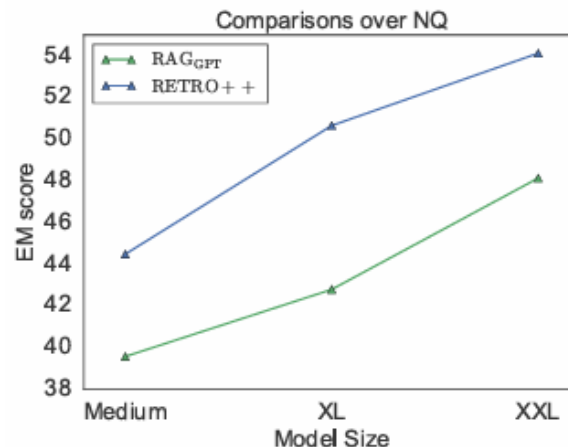
❖ Experiments

- 전체적으로 가이드라인 실험들을 진행

RAG에 **decoder-only** 모델
적용 가능성 확인

Method	NQ	TriviaQA
GPT (close book)	36.1	45.1
REALM (Gua et al., 2020)	40.4	-
DPR (Karpukhin et al., 2020)	41.5	56.8
RAG _{BART} (Lewis et al., 2020b)	44.5	56.1
RAG _{GPT}	50.9	60.9
FiD _{Large} (Izacard and Grave, 2021)	51.4	67.6
RETRO (Ours)	40.9	59.9
RETRO (Borgeaud et al., 2022)	45.5	-
RETRO++ (Ours)	54.1	66.7

Table 7: Comparisons of our RETRO and existing QA models. We report the best results with the largest model configuration respectively.



Retrieval with Language Models

Shall We Pretrain Autoregressive Language Models with Retrieval?

❖ Experiments

- 추가적인 Instruction-tuning에 따른 실험 결과
- 공개된 Instruction-tuning dataset을 조합하여 학습

	RAG _{GPT}	RETRO++
w/o Instruction tuning	24.43	25.93
w/ Instruction tuning	29.75	31.16

RAG fine-tuning 진행 후
추가적인 Instruction-tuning도 가능

Table 8: Exact Match (EM) scores for the *zero-shot evaluation* of RAG_{GPT} and RETRO++ on the NQ dataset before and after instruction tuning.

Conclusion

Conclusion

❖ RAG

- 외부 database를 통한 Retriever을 활용하여 언어 모델의 성능을 향상시킴
- 기존 사전 학습된 언어 모델을 generator로 사용 및 연산 측면에서 많은 개선 여지를 보임

❖ RETRO

- Chunk, Continuation, CCA, Decoder-only generator 활용하여 성능을 향상시킴
- CCA를 통해 연산적인 문제도 완화시킴

❖ RETRO++

- RETRO를 직접 구현 및 추가적인 fine-tuning을 수행한 RETRO++ 제안
- 다양한 실험을 통해 RAG 및 RETRO 프레임워크의 가이드 라인 제시

References

References

- ❖ Yu, W., Wu, L., Deng, Y., Mahindru, R., Zeng, Q., Guven, S., & Jiang, M. (2020, October). A technical question answering system with transfer learning. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 92-99).
- ❖ Xiao, S., Liu, Z., Zhang, P., & Muennighof, N. (2023). C-pack: Packaged resources to advance general chinese embedding. arXiv preprint arXiv:2309.07597.
- ❖ Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.
- ❖ Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.
- ❖ Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022, June). Improving language models by retrieving from trillions of tokens. In International conference on machine learning (pp. 2206-2240). PMLR.
- ❖ Wang, B., Ping, W., Xu, P., McAfee, L., Liu, Z., Shoenybi, M., ... & Catanzaro, B. (2023). Shall we pretrain autoregressive language models with retrieval? a comprehensive study. arXiv preprint arXiv:2304.06762.

고맙습니다