In-Context Retrieval-Augmented Language Models

Seongeun Baek

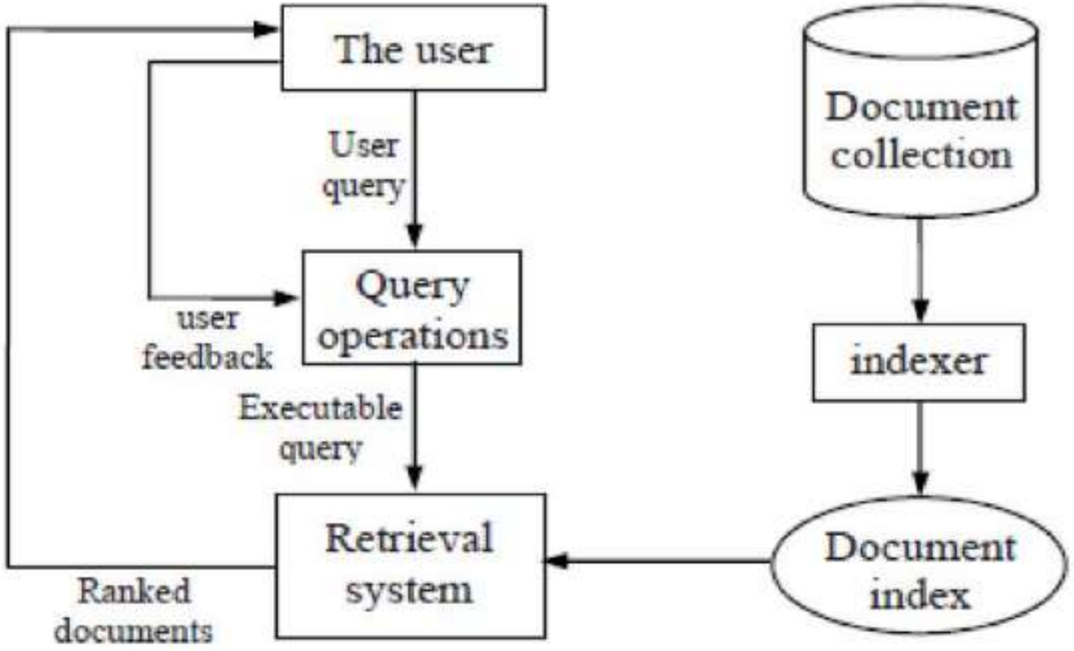
2024. 2. 05.

고려대학교
KOREA UNIVERSITY

Main point

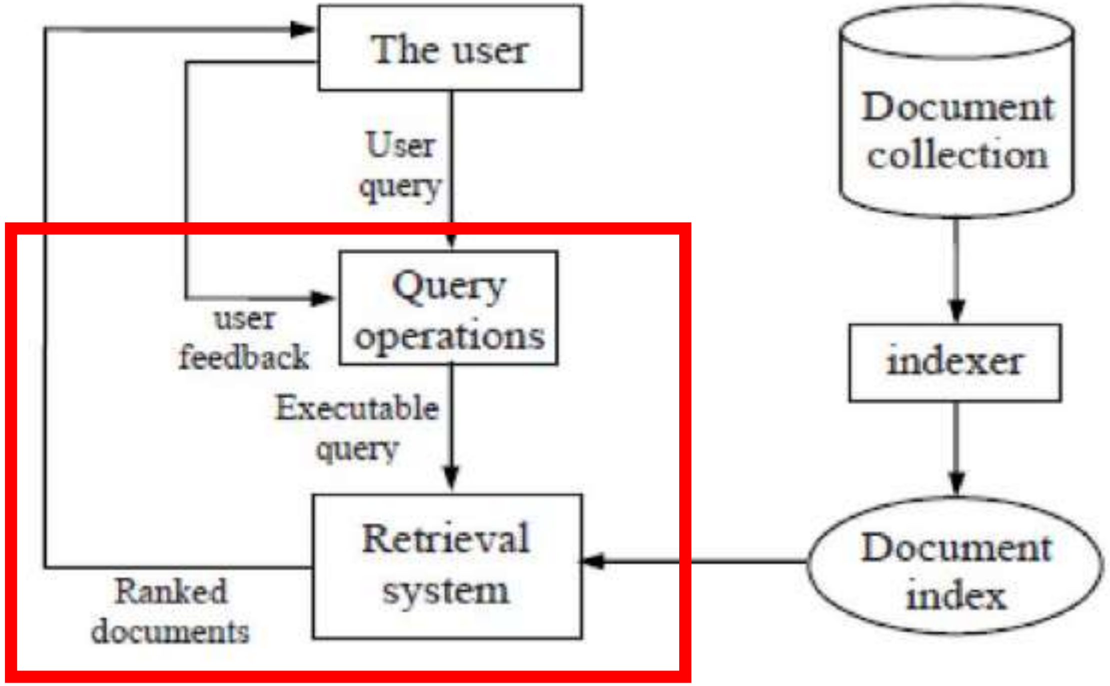
In-Context learning을 활용한 Retrieval Augmented Language Model

Previous works



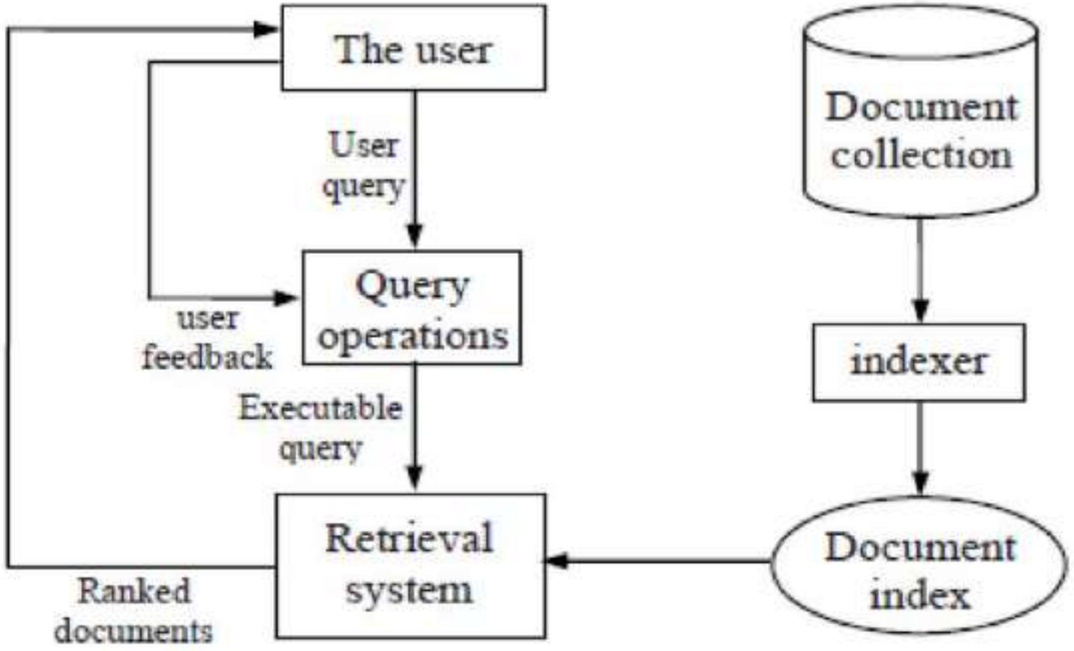
1. **Document Selection**
2. **Document Reading**

Previous works



1. **Document Selection**
2. **Document Reading**

Previous works



1. Document Selection
2. Document Reading

Previous works

1. Document Selection
2. Document Reading

Methods (In-Context RALM)

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p_\theta(x_i | x_{<i}), \qquad (1)$$

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p_\theta \left(x_i \middle| \left[\mathcal{R}_\mathcal{C}(x_{<i}); x_{<i} \right] \right), \qquad (2)$$

(1) Conditional probability -> (2) Retrieval augmented

Methods (RALM Design Choices)

$$p(x_1, ..., x_n) = \prod_{j=0}^{n_s-1} \prod_{i=1}^{s} p_\theta \left(x_{s \cdot j+i} \middle| \left[\mathcal{R}_\mathcal{C}(x_{\leq s \cdot j}); x_{<(s \cdot j+i)} \right] \right), \tag{3}$$

Retrieval Stride

$$p(x_1, ..., x_n) = \prod_{j=0}^{n_s-1} \prod_{i=1}^{s} p_\theta \left(x_{s \cdot j+i} \middle| \left[\mathcal{R}_\mathcal{C}(q_j^{s,\ell}); x_{<(s \cdot j+i)} \right] \right). \tag{4}$$

Retrieval Query Length

Language modeling

- **WikiText-103**
  - RALM을 평가하는 데에 가장 많이 사용되는 dataset
- **The Pile**
  - **Arxiv**
  - **Stack Exchange**
  - **FreeLaw**
- **Real-News**
  - The Pile이 news에 대한 corpus가 부족

Experiments (Models)

Language Models
•GPT-2의 4개 모델
•GPT-Neo와 GPT-J의 3개 모델
•OPT의 8개 모델
•LLaMa의 3개 모델
→ 모두 open source, available

Retrievers
•sparse (word-based) retrievers
  • BM25
•dense (neural) retrievers
  • frozen BERT-base
  • Contriever and Spider

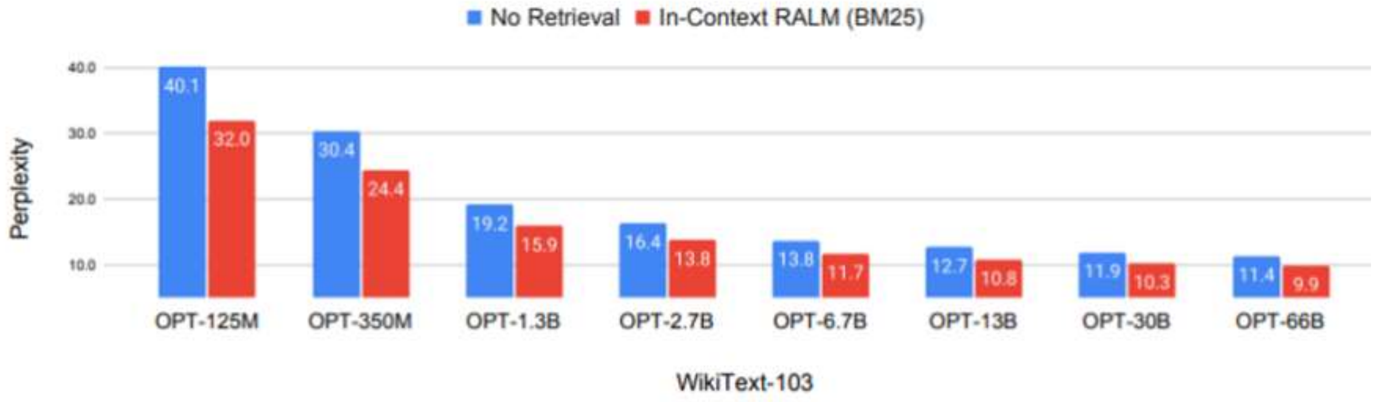
Reranking
•RoBERTa-base로 시작하여 reranker를 학습

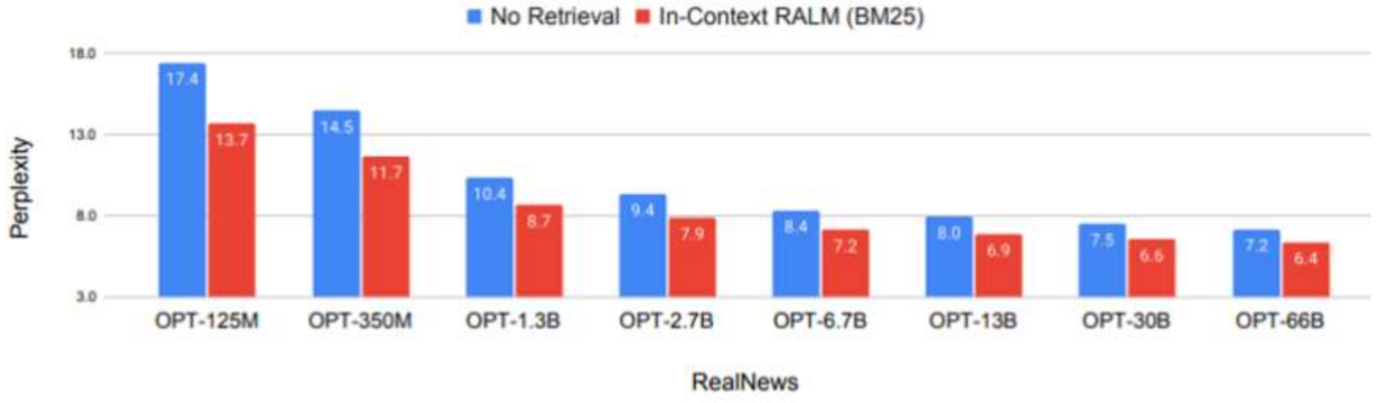
Experiments (In-context RALM with Off-the-Shelf Retrievers)

Model	Retrieval	Reranking	WikiText-103 word ppl	RealNews token ppl	ArXiv token ppl	Stack Exch. token ppl	FreeLaw token ppl
GPT-2 S	–	–	37.5	21.3	12.0	12.8	13.0
	BM25 §5	–	29.6	16.1	10.9	11.3	9.6
	BM25	Zero-shot §6.1	28.6	15.5	10.1	10.6	8.8
	BM25	Predictive §6.2	26.8	–	–	–	–
GPT-2 M	–	–	26.3	15.7	9.3	8.8	9.6
	BM25 §5	–	21.5	12.4	8.6	8.1	7.4
	BM25	Zero-shot §6.1	20.8	12.0	8.0	7.7	6.9
	BM25	Predictive §6.2	19.7	–	–	–	–
GPT-2 L	–	–	22.0	13.6	8.4	8.5	8.7
	BM25 §5	–	18.1	10.9	7.8	7.8	6.8
	BM25	Zero-shot §6.1	17.6	10.6	7.3	7.4	6.4
	BM25	Predictive §6.2	16.6	–	–	–	–
GPT-2 XL	–	–	20.0	12.4	7.8	8.0	8.0
	BM25 §5	–	16.6	10.1	7.2	7.4	6.4
	BM25	Zero-shot §6.1	16.1	9.8	6.8	7.1	6.0
	BM25	Predictive §6.2	15.4	–	–	–	–

Table 1: Perplexity on the test set of WikiText-103, RealNews and three datasets from the Pile. For each LM, we report: (a) its performance without retrieval, (b) its performance when fed the top-scored passage by BM25 (§5), and (c) its performance when applied on the top-scored passage of each of our two suggested rerankers (§6). All models share the same vocabulary, thus token-level perplexity (*token ppl*) numbers are comparable. For WikiText we follow prior work and report word-level perplexity (*word ppl*).

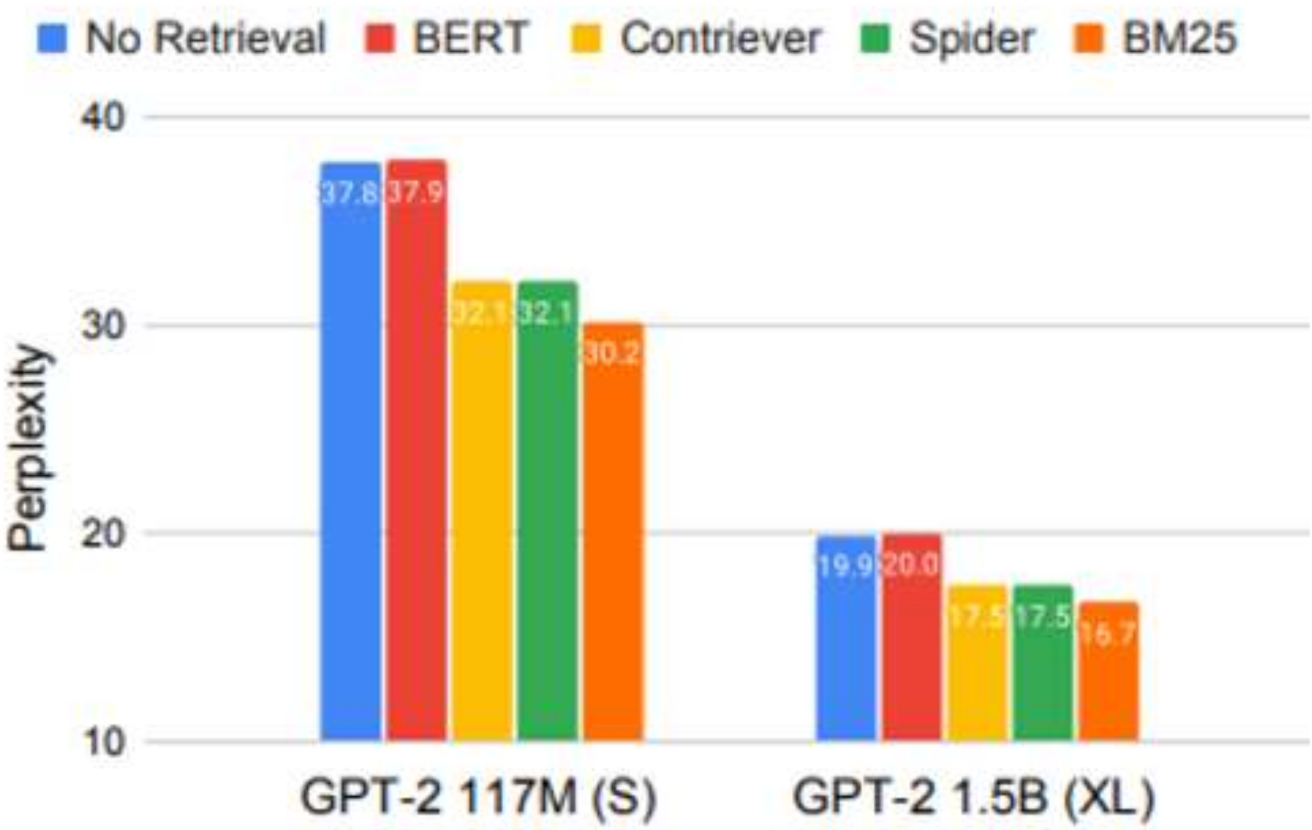
Experiments (In-context RALM with Off-the-Shelf Retrievers)



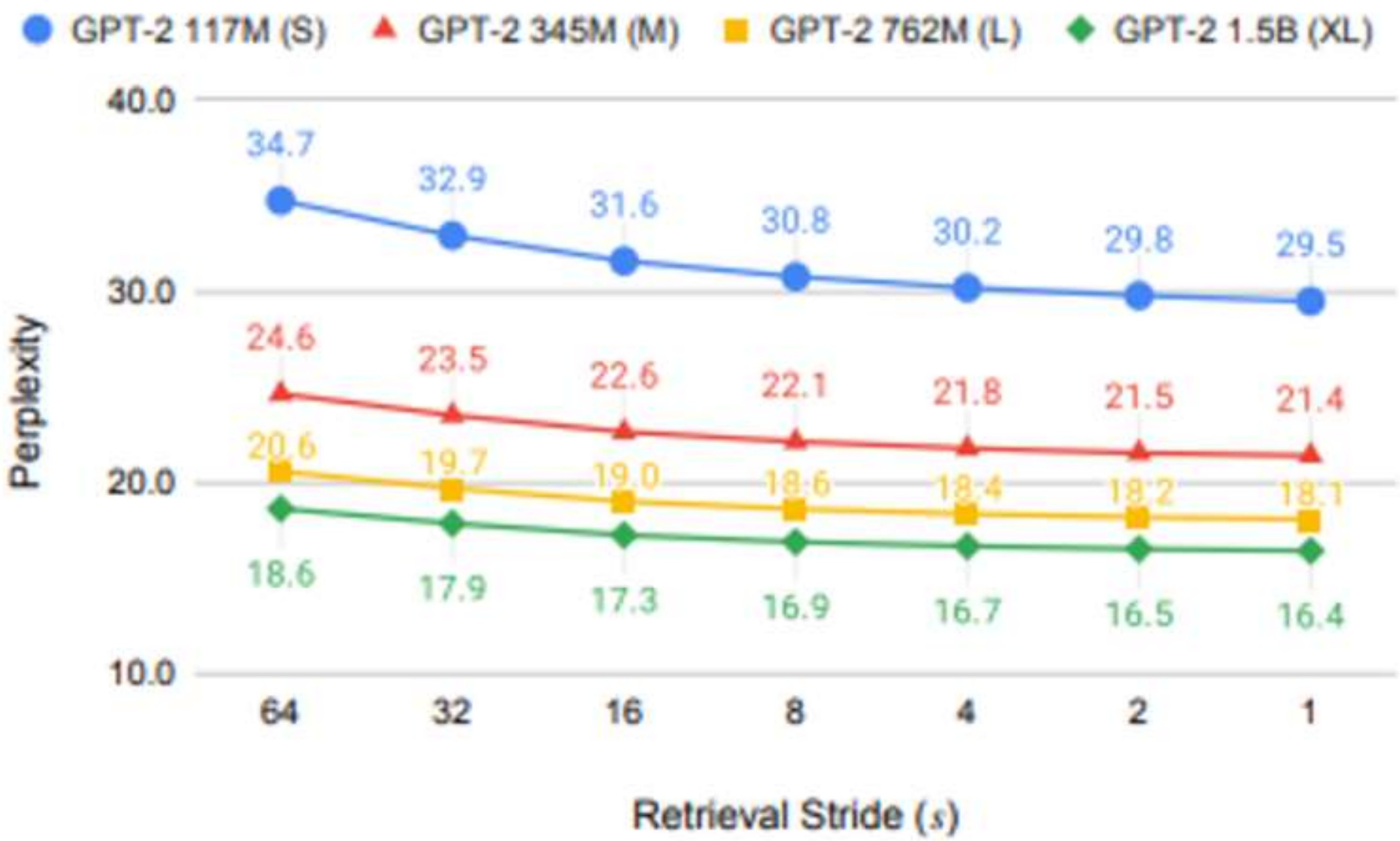


Model	Retrieval	WikiText-103
		word ppl
LLaMA-7B	-	9.9
	BM25, §5	8.8
LLaMA-13B	-	8.5
	BM25, §5	7.6
LLaMA-33B	-	6.3
	BM25, §5	6.1

Experiments (Best Retriever)

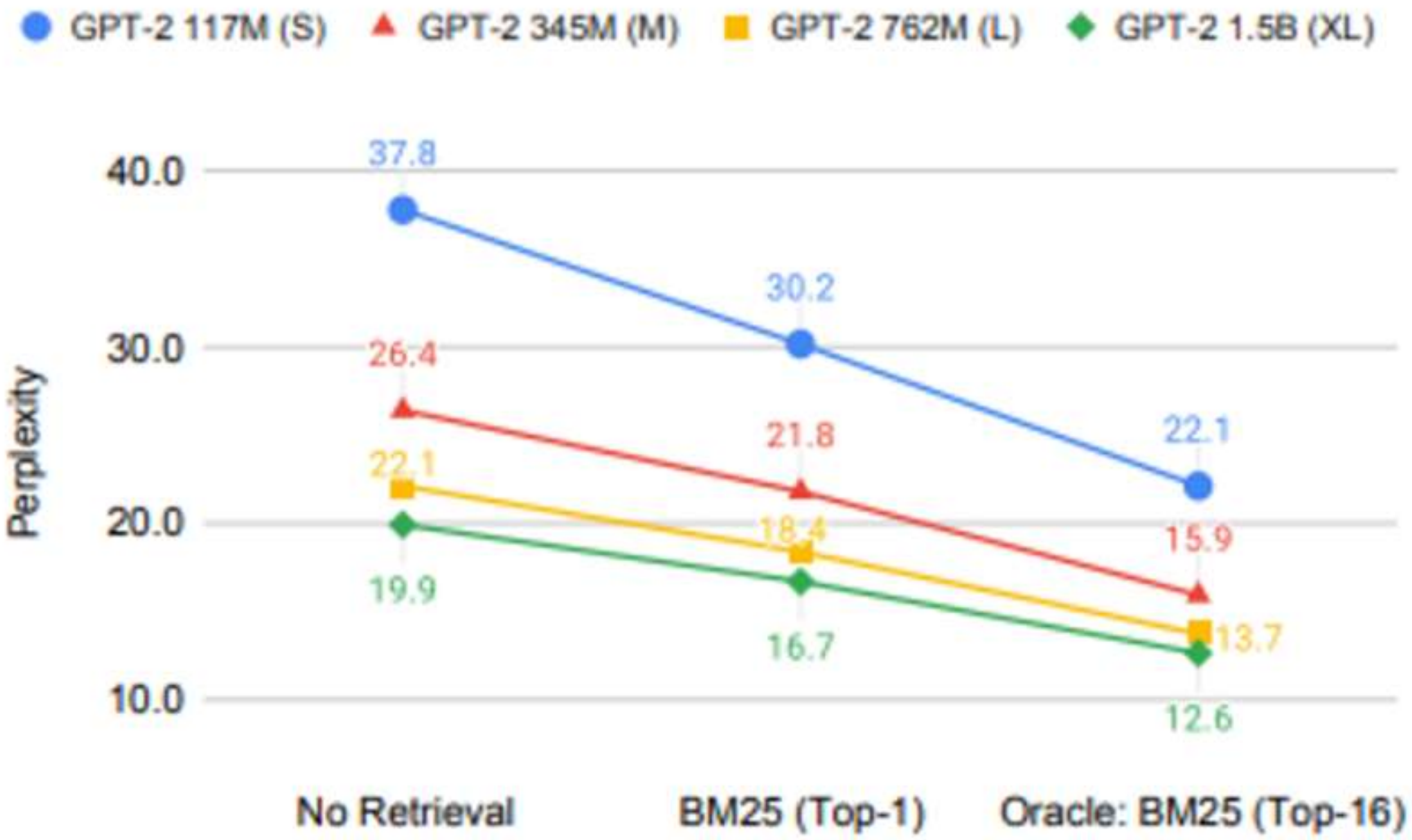


Experiments (Frequent Retrieval)



Experiments (Contextualization vs Recency Trade-off)

Experiments (In-Context RALM with LM-Oriented Reranking)



Experiments (LMs as Zero-Shot Rerankers)

$$y := x_{s \cdot j+1}, \ldots, x_{s \cdot j+s}$$

$$i^* = \underset{i \in [k]}{\arg\max}\, p_\theta(y |\, [d_i; x_{\le s \cdot j}]). \qquad (5)$$

$$y' := x_{s \cdot j - s' + 1}, \ldots, x_{s \cdot j}$$

$$\hat{i} = \underset{i \in [k]}{\arg\max}\, p_\phi(y' |\, [d_i; x_{\le (s \cdot j - s')}]). \qquad (6)$$

Experiments (LMs as Zero-Shot Rerankers)

$$y := x_{s \cdot j+1}, \ldots, x_{s \cdot j+s}$$

$$i^* = \arg\max_{i \in [k]} p_\theta(y | [d_i; x_{\leq s \cdot j}]). \quad (5)$$

$$y' := x_{s \cdot j - s' + 1}, \ldots, x_{s \cdot j}$$

$$\hat{i} = \arg\max_{i \in [k]} p_\phi(y' | [d_i; x_{\leq (s \cdot j - s')}]). \quad (6)$$

Experiments (LMs as Zero-Shot Rerankers)

Model	Reranking Model	WikiText-103 word ppl	RealNews token ppl
GPT-2 345M (M)	GPT-2 110M (S)	20.8	12.1
	GPT-2 345M (M)	20.8	12.0
GPT-2 762M (L)	GPT-2 110M (S)	17.7	10.7
	GPT-2 762M (L)	17.6	10.6
GPT-2 1.5B (XL)	GPT-2 110M (S)	16.2	9.9
	GPT-2 1.5B (XL)	16.1	9.8

Experiments (Training LM-dedicated Rerankers)

$$\boldsymbol{x_{\leq s \cdot j} + d_i \rightarrow f\left(x_{\leq s \cdot j}, d_i\right) = scalar}$$

$$p_{\text{rank}}(d_i|x_{\leq s \cdot j}) = \frac{\exp(f(x_{\leq s \cdot j}, d_i))}{\sum_{i'=1}^{k} \exp(f(x_{\leq s \cdot j}, d_{i'}))}, \quad (7)$$

$$\hat{i} = \arg\max_{i \in [k]} \; p_{\text{rank}}(d_i|x_{\leq s \cdot j}). \quad (8)$$

Experiments (Training LM-dedicated Rerankers)

$$x_{\leq s \cdot j} + d_i \rightarrow f(x_{\leq s \cdot j}, d_i) = scalar$$

$$p_{\text{rank}}(d_i | x_{\leq s \cdot j}) = \frac{\exp(f(x_{\leq s \cdot j}, d_i))}{\sum_{i'=1}^{k} \exp(f(x_{\leq s \cdot j}, d_{i'}))}, \quad (7)$$

$$\boxed{\hat{i} = \underset{i \in [k]}{\arg\max}\ p_{\text{rank}}(d_i | x_{\leq s \cdot j}). \quad (8)}$$

Experiments (Training LM-dedicated Rerankers)

Collecting Training Examples + Training

predictive reranker를 훈련하기 위해 다음과 같은 training example을 모아야 한다.

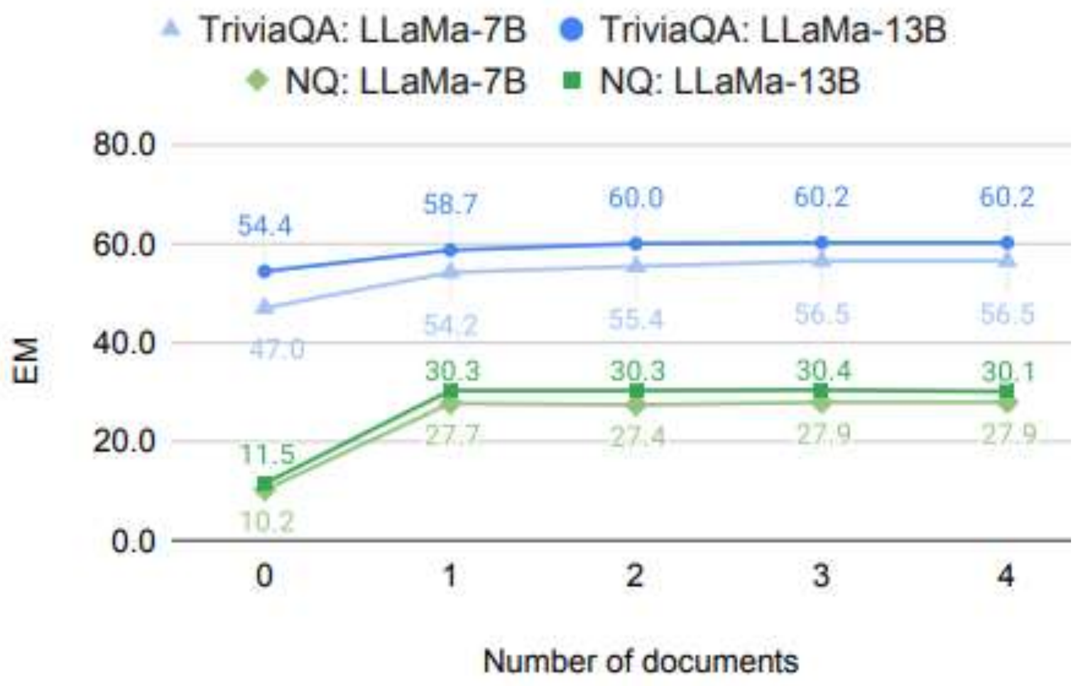
1. $x_{\leq s \cdot j}$를 훈련 데이터로부터 우리가 샘플링한 prefix라고 하자. $y := x_{s \cdot j+1}, \ldots, x_{s \cdot j+s}$는 다음 stride에 올 generation text라고 하자.
2. BM25를 $x_{\leq s \cdot j}$로부터 query $q_j^{s,\ell}$를 얻고 k개의 document를 얻는다.
3. 각 document d_i에 대해서 LM을 이용하여 $p_\theta(y|[d_i; x_{\leq s \cdot j}])$을 계산한다.

$$-log \sum_{i=1}^{k} p_{rank}(d_i|x_{\leq s \cdot j}) \cdot p_\theta(y|[d_i; x_{\leq s \cdot j}]) \quad (9)$$

4. (9)의 식을 가지고 전체 loss function을 정의하여 training한다.

이 때, training은 RoBERTa-base로 fine-tuning하는 방식으로 진행한다.

Experiments (Open-Domain Question Answering)



Model	Retrieval	NQ	TriviaQA
LLaMA-7B	-	10.3	47.5
	DPR	28.0	56.0
LLaMA-13B	-	12.0	54.8
	DPR	31.0	60.1
LLaMA-33B	-	13.7	58.3
	DPR	32.3	62.7

Discussion

- **In-Context RALM**
  - **Various Model (possible 'only API access' model)**
  - **In-Context Retrieval**
  - **Cost efficiency**
- **Limitation**
  - **Only one documents**
  - **Retrieval Stride & selective retrieval**