Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection

Akari Asai[†], Zeqiu Wu[†], Yizhong Wang[†§], Avirup Sil[‡], Hannaneh Hajishirzi[†§]

[†]University of Washington [§]Allen Institute for AI [‡]IBM Research AI

ICLR 2024

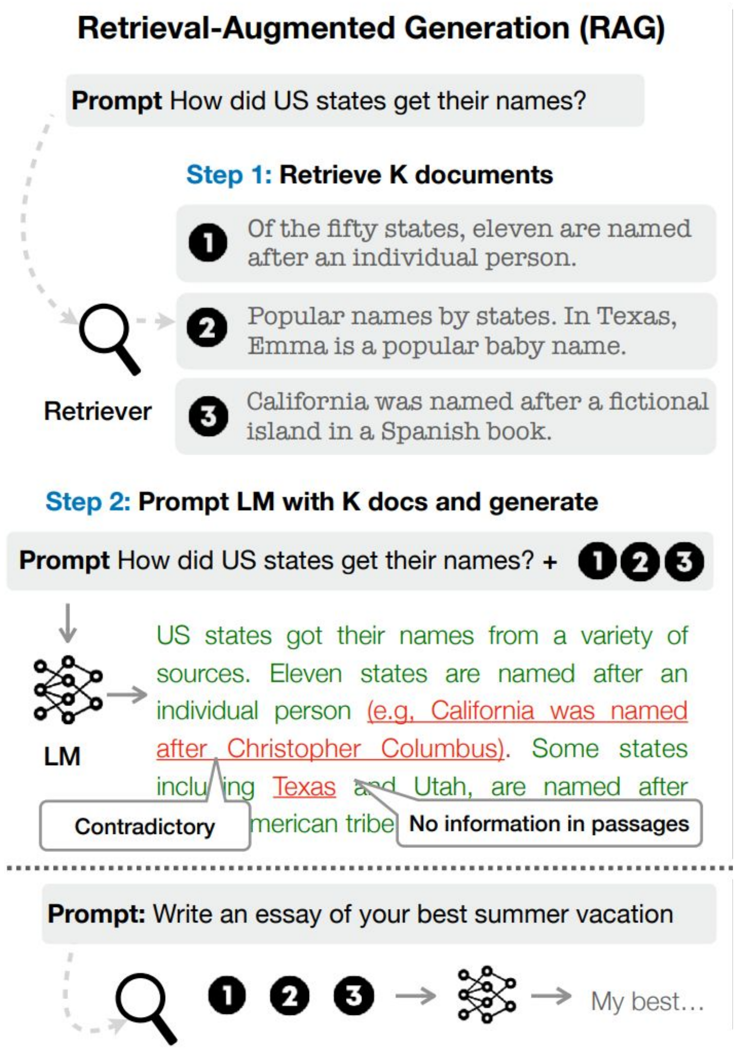
2024.10.25

발제자: 윤예준
yeayen789@gmail.com

Background

- LLMs은 parametric knowledge에 의존하기 때문에 factual inaccuracies를 포함한 response를 자주 생성
- 이에 대안으로 나온 RAG는 검색한 관련 passage로 LLMs의 입력을 증강하여 knowledge-intensive task에서의 factual errors를 줄임

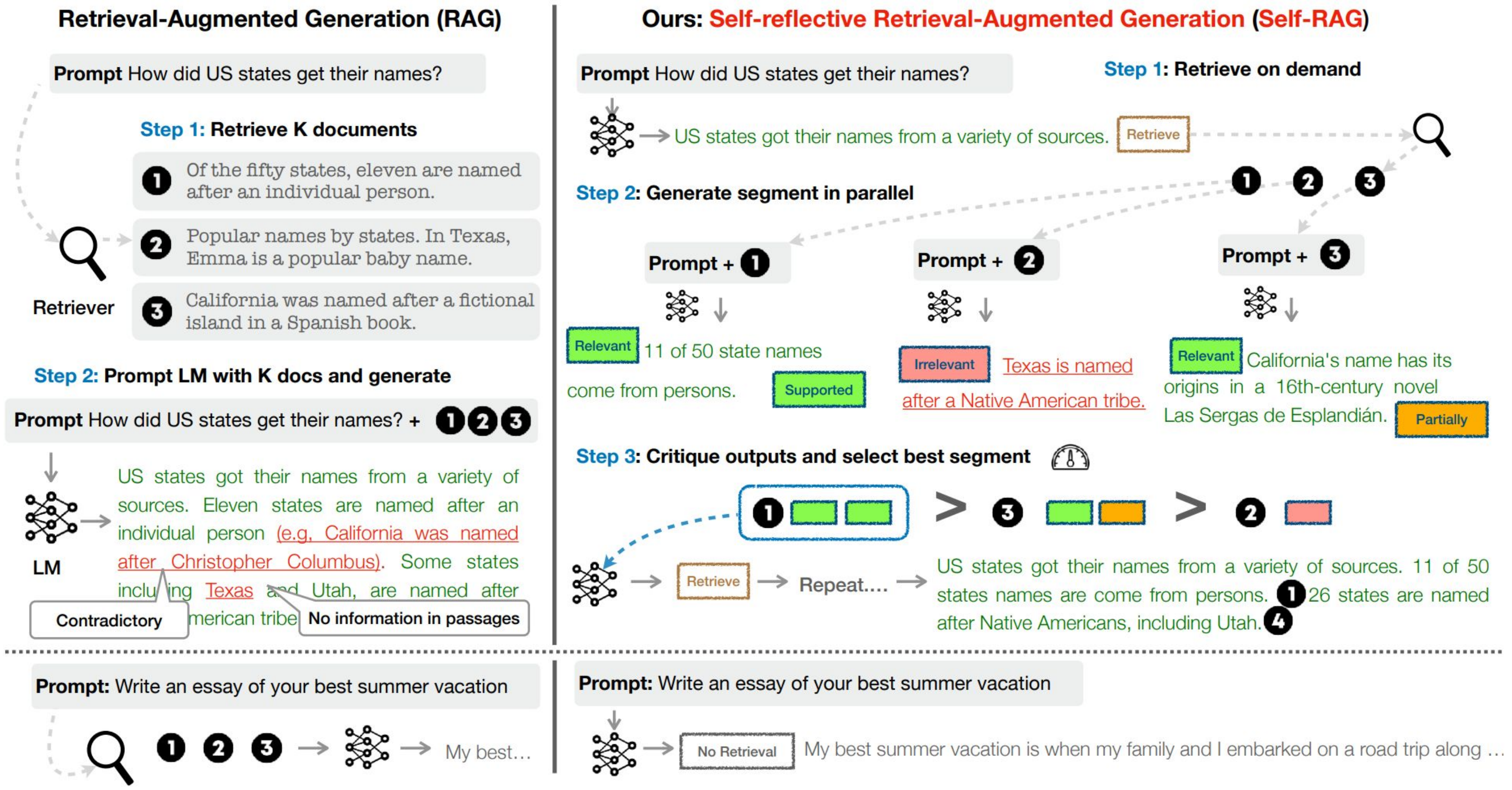
- 기존 RAG 방법들의 문제점
  - 검색된 passage가 도움이 되는지 확인 없이 무분별하게 검색
    → 관련 없는 검색된 passage를 입력에 사용할 수 있음
  - LLMs의 output은 검색된 근거와의 일관성 보장하지 않음
    → LLMs의 다양성과 품질을 저해하는 response를 생성할 수 있음
    (검색된 근거 활용하는 학습이 되어있지 않기 때문)



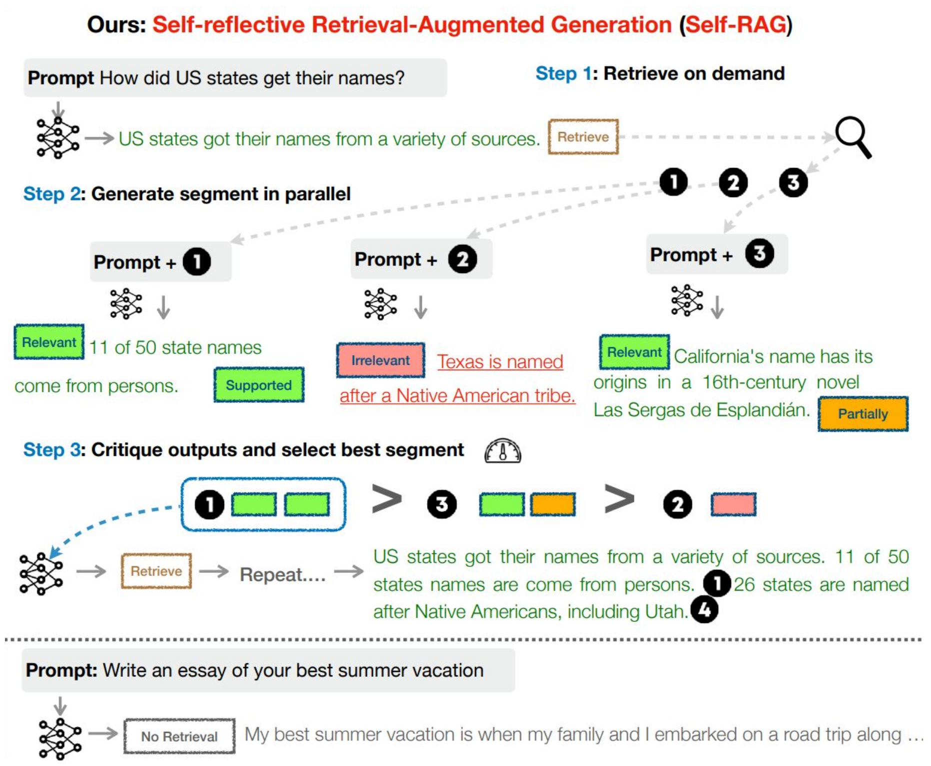
Self-RAG

- **Learning to Retrieve, Generate, and Critique through Self-Reflection**

- On-demand 검색 및 self-reflection을 통해 response의 다양성을 해치지 않으면서 factual accuracy를 포함한 LLMs의 generation quality를 개선하는 방법 제안



Self-RAG – overview



Type	Input	Output	Definitions
Retrieve	x / x, y	{yes, no, continue}	Decides when to retrieve with $\mathcal{R}$
IsRel	x, d	{**relevant**, irrelevant}	d provides useful information to solve x.
IsSup	x, d, y	{**fully supported**, partially supported, no support}	All of the verification-worthy statement in y is supported by d.
IsUse	x, y	{**5**, 4, 3, 2, 1}	y is a useful response to x.

Table 1: Four types of reflection tokens used in Self-Rag. Each type uses several tokens to represent its output values. The bottom three rows are three types of Critique tokens, and **the bold text** indicates the most desirable critique tokens. x, y, d indicate input, output, and a relevant passage, respectively.

Algorithm 1 Self-Rag Inference

Require: Generator LM $\mathcal{M}$, Retriever $\mathcal{R}$, Large-scale passage collections $\{d_1, \dots, d_N\}$

Input: input prompt x and preceding generation $y_{<t}$, **Output:** next output segment y_t

$\mathcal{M}$ predicts Retrieve given $(x, y_{<t})$

if Retrieve == `Yes` **then**

  Retrieve relevant text passages $\mathbf{D}$ using $\mathcal{R}$ given (x, y_{t-1}) ▷ Retrieve

  $\mathcal{M}$ predicts IsRel given x, d and y_t given $x, d, y_{<t}$ for each $d \in \mathbf{D}$ ▷ Generate

  $\mathcal{M}$ predicts IsSup and IsUse given x, y_t, d for each $d \in \mathbf{D}$ ▷ Critique

  Rank y_t based on IsRel, IsSup, IsUse ▷ Detailed in Section 3.3

else if Retrieve == `No` **then**

  $\mathcal{M}_{gen}$ predicts y_t given x ▷ Generate

  $\mathcal{M}_{gen}$ predicts IsUse given x, y_t ▷ Critique

Self-RAG – train dataset build

- Reflection token을 삽입한 학습 데이터셋 구축
  - Manual annotation 또는 sota LLM (GPT-4)으로 구축 할 수 있으나 cost가 높음
  - 따라서 GPT-4로 학습 데이터를 샘플링하여 reflection 데이터셋 구축 및 critic model에 지식을 distill함
  - Fine-tuned critic model로 학습 데이터셋에 대해 reflection token 각각 생성 후 결합

Instructions
Given an instruction, please make a judgment on whether finding some external documents from the web (e.g., Wikipedia) helps to generate a better response. Please answer [Yes] or [No] and write an explanation.

Demonstrations
Instruction Give three tips for staying healthy.
Need retrieval? [Yes]
Explanation There might be some online sources listing three tips for staying healthy or some reliable sources to explain the effects of different behaviors on health. So retrieving documents is helpful to improve the response to this query.

Instruction Describe a time when you had to make a difficult decision.
Need retrieval? [No]
Explanation This instruction is asking about some personal experience and thus it does not require one to find some external documents.

Instruction Write a short story in third person narration about a protagonist who has to make an important career decision.
Need retrieval? [No]
Explanation This instruction asks us to write a short story, which does not require external evidence to verify.

Instruction What is the capital of France?
Need retrieval? [Yes]
Explanation While the instruction simply asks us to answer the capital of France, which is a widely known fact, retrieving web documents for this question can still help.

Instruction Find the area of a circle given its radius. Radius = 4
Need retrieval? [No]
Explanation This is a math question and although we may be able to find some documents describing a formula, it is unlikely to find a document exactly mentioning the answer.

Instruction Arrange the words in the given sentence to form a grammatically correct sentence. quickly the brown fox jumped
Need retrieval? [No]
Explanation This task doesn't require any external evidence, as it is a simple grammatical question.

Instruction Explain the process of cellular respiration in plants.
Need retrieval? [Yes]
Explanation This instruction asks for a detailed description of a scientific concept, and is highly likely that we can find a reliable and useful document to support the response.

Table 8: Instructions and demonstrations for Retrieve aspect given the input only.

Algorithm 3 $\mathcal{M}_{gen}$ Data creation

Input Input-output data $\mathcal{D} = X, Y$
for $(x, y) \in \{X, Y\}$ **do**
  Given (x, y) $\mathcal{C}$ predicts Retrieve
  if Retrieve is predicted **then**
    Retrieve relevant passages **D** using $\mathcal{R}$ given (x, y) ▷ Retrieve passages
    for $d \in$ **D** **do**
      $\mathcal{C}$ predicts IsRel for each d ▷ Predict relevance of passages
      $\mathcal{C}$ predicts IsSup for each (y, d) ▷ Predict supports of outputs
    $\mathcal{C}$ predicts IsUse for each d ▷ Predict overall utility ($t = \mathcal{T}$ only)
    Sample d
  else if Retrieve is not predicted **then**
    $\mathcal{C}$ predicts IsUse given x, y
  Add augmented (x, y, d, r) to $\mathcal{D}_{gen}$

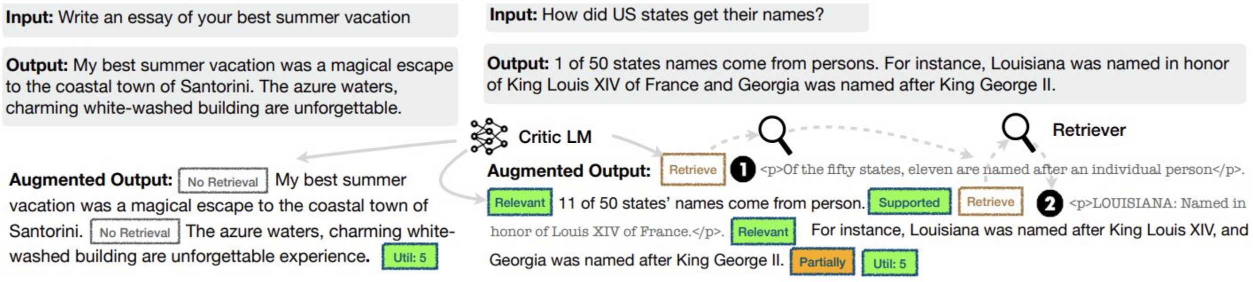


Figure 2: Self-RAG training examples. The left example does not require retrieval while the right one requires retrieval; thus, passages are inserted. More examples are in Appendix Table 4.

Self-RAG – training

Algorithm 2 SELF-RAG Training

```
**Input** input-output data $\mathcal{D} = \{X, Y\}$, generator $\mathcal{M}$, $\mathcal{C}$ $\theta$
Initialize $\mathcal{C}$ with a pre-trained LM
Sample data $\{X^{sample}, Y^{sample}\} \sim \{X, Y\}$ ▷ **Training Critic LM (Section 3.2.1)**
**for** $(x, y) \in (X^{sample}, Y^{sample})$ **do** ▷ Data collections for $\mathcal{C}$
    Prompt GPT-4 to collect a reflection token $r$ for $(x, y)$
    Add $\{(x, y, r)\}$ to $\mathcal{D}_{critic}$
Update $\mathcal{C}$ with next token prediction loss ▷ Critic learning; Eq. 1
Initialize $\mathcal{M}$ with a pre-trained LM ▷ **Training Generator LM (Section 3.2.2)**
**for** $(x, y) \in (X, Y)$ **do** ▷ Data collection for $\mathcal{M}$ with $\mathcal{D}_{critic}$
    Run $\mathcal{C}$ to predict $r$ given $(x, y)$
    Add $(x, y, r)$ to $\mathcal{D}_{gen}$
Update $\mathcal{M}$ on $\mathcal{D}_{gen}$ with next token prediction loss ▷ Generator LM learning; Eq. 2
```

Critic learning

$$\max_{\mathcal{C}} \mathbb{E}_{((x,y),r)\sim\mathcal{D}_{critic}} \log p_{\mathcal{C}}(r|x,y), \quad r \text{ for reflection tokens.}$$

Generator learning

$$\max_{\mathcal{M}} \mathbb{E}_{(x,y,r)\sim\mathcal{D}_{gen}} \log p_{\mathcal{M}}(y, r|x).$$

Self-RAG – inference

- Adaptive retrieval with threshold
  - 생성된 토큰이 Retrieve=yes인 경우 검색 시도
  - Retrieve=yes의 prob이 threshold를 넘는 경우

- Tree-decoding with critique tokens
  - segment-level beam search
  - critic score와 결합하여 최종 segment 선택

$$f(y_t, d, \boxed{\text{Critique}}) = p(y_t|x, d, y_{<t})) + \mathcal{S}(\boxed{\text{Critique}}), \text{where}$$

$$\mathcal{S}(\boxed{\text{Critique}}) = \sum_{G \in \mathcal{G}} w^G s_t^G \text{ for } \mathcal{G} = \{\boxed{\text{IsRel}}, \boxed{\text{IsSup}}, \boxed{\text{IsUse}}\},$$

Algorithm 1 SELF-RAG Inference

Require: Generator LM $\mathcal{M}$, Retriever $\mathcal{R}$, Large-scale passage collections $\{d_1, \dots, d_N\}$

Input: input prompt x and preceding generation $y_{<t}$, **Output:** next output segment y_t
$\mathcal{M}$ predicts $\boxed{\text{Retrieve}}$ given $(x, y_{<t})$
if $\boxed{\text{Retrieve}}$ == `Yes` **then**
  Retrieve relevant text passages **D** using $\mathcal{R}$ given (x, y_{t-1}) ▷ Retrieve
  $\mathcal{M}$ predicts $\boxed{\text{IsRel}}$ given x, d and y_t given $x, d, y_{<t}$ for each $d \in \mathbf{D}$ ▷ Generate
  $\mathcal{M}$ predicts $\boxed{\text{IsSup}}$ and $\boxed{\text{IsUse}}$ given x, y_t, d for each $d \in \mathbf{D}$ ▷ Critique
  Rank y_t based on $\boxed{\text{IsRel}}$, $\boxed{\text{IsSup}}$, $\boxed{\text{IsUse}}$ ▷ Detailed in Section 3.3
else if $\boxed{\text{Retrieve}}$ == `No` **then**
  $\mathcal{M}_{gen}$ predicts y_t given x ▷ Generate
  $\mathcal{M}_{gen}$ predicts $\boxed{\text{IsUse}}$ given x, y_t ▷ Critique

Experimental settings

- Generator model: Llama2 7B, 13B
- Critic model: Llama2 7B

- Inference settings
    - Weight (IsREL, IsSUP, IsUse): 1, 1, 0.5
    - Retrieval threshold: 0.2, 0 (ALCE만)
    - Beam width: 2
    - Retriever: Contriever-MS MARCO
        - Top-k: 5

Dataset name	Category	Data source	# of instances	% of $\boxed{\text{Retrieve}}$=Yes
GPT-4 Alpaca	Instruction-following	Open-Instruct	26,168	53.2
Stanford Alpaca	Instruction-following	Open-Instruct	25,153	48.0
FLAN-V2	Instruction-following	Open-Instruct	17,817	15.8
ShareGPT	Instruction-following	Open-Instruct	13,406	76.8
Open Assistant 1	Instruction-following	Open-Instruct	9,464	77.1
Wizard of Wikipedia	Knowledge-intensive	KILT	17,367	22.7
Natural Questions	Knowledge-intensive	KILT	15,535	87.7
FEVER	Knowledge-intensive	KILT	9,966	63.2
OpenBoookQA	Knowledge-intensive	HF Dataset	4,699	2.3
Arc-Easy	Knowledge-intensive	HF Dataset	2,147	11.0
ASQA	Knowledge-intensive	ASQA	3,897	91.5

Table 3: The generator LM $\mathcal{M}$ training data statistics.

Results

- Comparison against baselines without retrieval
  - ChatGPT 보다 PopQA, Pub, Bio, ASQA(Rouge and MAUVE)에서 높은 성능 달성
  - Pre-trained LLMs 보다 높은 성능
  - Bio에서 CoVE 모델보다 성능 능가

	Short-form		Closed-set		Long-form generations (with citations)					
	PopQA	TQA	Pub	ARC	Bio			ASQA		
LM	(acc)	(acc)	(acc)	(acc)	(FS)	(em)	(rg)	(mau)	(pre)	(rec)
LMs with proprietary data										
Llama2-c$_{13B}$	20.0	59.3	49.4	38.4	55.9	22.4	29.6	28.6	–	–
Ret-Llama2-c$_{13B}$	51.8	59.8	52.1	37.9	79.9	32.8	34.8	43.8	19.8	36.1
ChatGPT	29.3	74.3	70.1	75.3	71.8	35.3	36.2	68.8	–	–
Ret-ChatGPT	50.8	65.7	54.7	75.3	–	40.7	39.9	79.7	65.1	76.6
Perplexity.ai	–	–	–	–	71.2	–	–	–	–	–
Baselines without retrieval										
Llama2$_{7B}$	14.7	30.5	34.2	21.8	44.5	7.9	15.3	19.0	–	–
Alpaca$_{7B}$	23.6	54.5	49.8	45.0	45.8	18.8	29.4	61.7	–	–
Llama2$_{13B}$	14.7	38.5	29.4	29.4	53.4	7.2	12.4	16.0	–	–
Alpaca$_{13B}$	24.4	61.3	55.5	54.9	50.2	22.9	32.0	70.6	–	–
CoVE$_{65B}$ *	–	–	–	–	71.2	–	–	–	–	–
Baselines with retrieval										
Toolformer*$_{6B}$	–	48.8	–	–	–	–	–	–	–	–
Llama2$_{7B}$	38.2	42.5	30.0	48.0	78.0	15.2	22.1	32.0	2.9	4.0
Alpaca$_{7B}$	46.7	64.1	40.2	48.0	76.6	30.9	33.3	57.9	5.5	7.2
Llama2-FT$_{7B}$	48.7	57.3	64.3	65.8	78.2	31.0	35.8	51.2	5.0	7.5
SAIL*$_{7B}$	–	–	69.2	48.4	–	–	–	–	–	–
Llama2$_{13B}$	45.7	47.0	30.2	26.0	77.5	16.3	20.5	24.7	2.3	3.6
Alpaca$_{13B}$	46.1	66.9	51.1	57.6	77.7	**34.8**	36.7	56.6	2.0	3.8
Our SELF-RAG $_{7B}$	54.9	66.4	72.4	67.3	**81.2**	30.0	35.7	**74.3**	66.9	67.8
Our SELF-RAG $_{13B}$	**55.8**	**69.3**	**74.5**	**73.1**	80.2	31.7	**37.0**	71.6	**70.3**	**71.3**

Results

- Comparison against baselines with retrieval
  - 많은 task에서 기존 RAG 모델 성능 능가
  - 검색 기능 포함 시 instruct-tuned LMs은 검색 없는 모델에 비해 큰 성능 향상
  - 대부분 RAG 모델들은 인용 정확도 떨어짐
  - ChatGPT와의 성능 격차 해소하였으며 인용 정밀도에서 성능 능가

LM	Short-form PopQA (acc)	Short-form TQA (acc)	Closed-set Pub (acc)	Closed-set ARC (acc)	Long-form generations (with citations) Bio (FS)	ASQA (em)	ASQA (rg)	ASQA (mau)	ASQA (pre)	ASQA (rec)
LMs with proprietary data										
Llama2-c$_{13B}$	20.0	59.3	49.4	38.4	55.9	22.4	29.6	28.6	–	–
Ret-Llama2-c$_{13B}$	51.8	59.8	52.1	37.9	79.9	32.8	34.8	43.8	19.8	36.1
ChatGPT	29.3	74.3	70.1	75.3	71.8	35.3	36.2	68.8	–	–
Ret-ChatGPT	50.8	65.7	54.7	75.3	–	40.7	39.9	79.7	65.1	76.6
Perplexity.ai	–	–	–	–	71.2	–	–	–	–	–
Baselines without retrieval										
Llama2$_{7B}$	14.7	30.5	34.2	21.8	44.5	7.9	15.3	19.0	–	–
Alpaca$_{7B}$	23.6	54.5	49.8	45.0	45.8	18.8	29.4	61.7	–	–
Llama2$_{13B}$	14.7	38.5	29.4	29.4	53.4	7.2	12.4	16.0	–	–
Alpaca$_{13B}$	24.4	61.3	55.5	54.9	50.2	22.9	32.0	70.6	–	–
CoVE$_{65B}$ *	–	–	–	–	71.2	–	–	–	–	–
Baselines with retrieval										
Toolformer*$_{6B}$	–	48.8	–	–	–	–	–	–	–	–
Llama2$_{7B}$	38.2	42.5	30.0	48.0	78.0	15.2	22.1	32.0	2.9	4.0
Alpaca$_{7B}$	46.7	64.1	40.2	48.0	76.6	30.9	33.3	57.9	5.5	7.2
Llama2-FT$_{7B}$	48.7	57.3	64.3	65.8	78.2	31.0	35.8	51.2	5.0	7.5
SAIL*$_{7B}$	–	–	69.2	48.4	–	–	–	–	–	–
Llama2$_{13B}$	45.7	47.0	30.2	26.0	77.5	16.3	20.5	24.7	2.3	3.6
Alpaca$_{13B}$	46.1	66.9	51.1	57.6	77.7	**34.8**	36.7	56.6	2.0	3.8
Our SELF-RAG $_{7B}$	54.9	66.4	72.4	67.3	**81.2**	30.0	35.7	**74.3**	66.9	67.8
Our SELF-RAG $_{13B}$	**55.8**	**69.3**	**74.5**	**73.1**	80.2	31.7	**37.0**	71.6	**70.3**	**71.3**

Ablation study

- No Retriever: 검색된 문서 없이 instruction-output pair로 학습
- No Critic: reflection token 없이 항상 검색된 top-1을 제공하여 학습
- Hard constraints: Retrieve=Yes 예측시 검색
- Remove IsSUP: Beam search시 IsSUP score 제거

$$f(y_t, d, \boxed{\text{Critique}}) = p(y_t|x, d, y_{<t})) + \mathcal{S}(\boxed{\text{Critique}}), \text{where}$$

$$\mathcal{S}(\boxed{\text{Critique}}) = \sum_{G \in \mathcal{G}} w^G s_t^G \text{ for } \mathcal{G} = \{\boxed{\text{IsRel}}, \boxed{\text{IsSup}}, \boxed{\text{IsUse}}\},$$

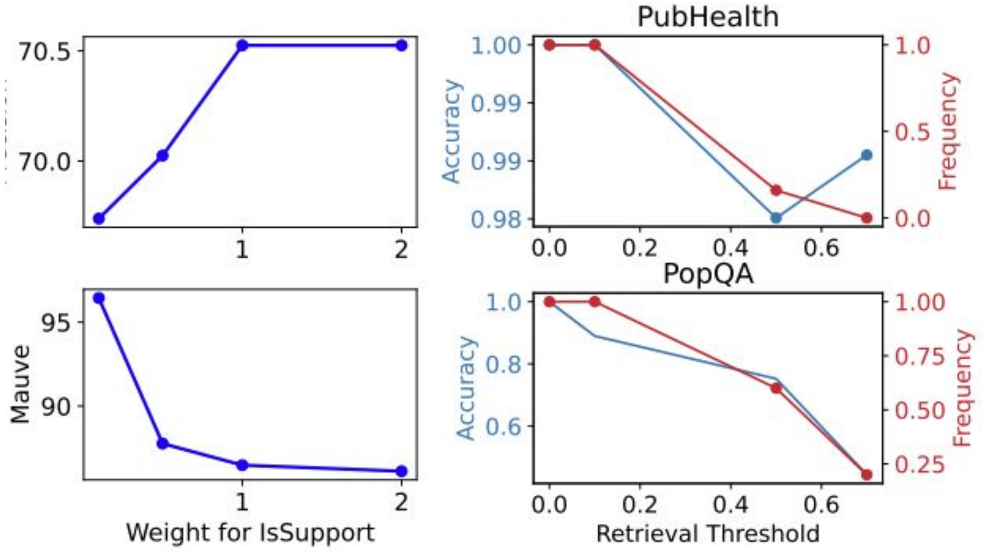
- 각 구성 요소들 모두 성능 향상에 기여하는 것을 확인
- 학습에서의 Retriever와 Critic이 Self-RAG의 성능 향상에 크게 기여

	PQA (acc)	Med (acc)	AS (em)
SELF-RAG (50k)	45.5	73.5	32.1
Training			
No Retriever $\mathcal{R}$	43.6	67.8	31.0
No Critic $\mathcal{C}$	42.6	72.0	18.1
Test			
No retrieval	24.7	73.0	–
Hard constraints	28.3	72.6	–
Retrieve top1	41.8	73.1	28.6
Remove IsSUP	44.1	73.2	30.6

(a) Ablation

Ablation study

- Effects of inference-time customization
  - IsSUP의 가중치 늘리면 증거에 의해 support한지를 강조하기 때문에 인용 정확도 향상
  - IsSUP의 가중치가 클 수록 MAUVE 점수 낮아짐
    → 생성 길어지고 유창해지면 인용과 관련 없는 응답이 더 생성되기 때문
- Efficiency and accuracy trade-off
  - reward token의 prob을 사용하여 검색 빈도 조정 가능
  - threshold 높을 수록 검색 빈도 낮아져 성능 저하 발생



(b) Customization (c) Retrieval

Conclusion

- Self-RAG라는 새로운 프레임워크를 제안. LLMs의 생성 quality와 factuality를 검색과 self-reflection을 통해 향상시킴
- LM이 새로 추가한 reflection token을 예측하여 검색, 생성, 비평하는 방법을 학습하도록 훈련
- Self-RAG는 reflection token을 활용하여 inference time에 LM의 동작을 customize할 수 있음
- 더 많은 파라미터를 가진 LLM이나 기존의 검색 증강 생성 방식을 사용하는 모델보다 뛰어난 성능을 보임

느낀 점

- 장점
  - 기존 RAG 방법에서는 고정된 수의 retrieved된 passage를 무조건 넣는 형태로 구성되어 있는데 self-reflection token으로 이를 해결
  - 다양한 reflection token을 활용하여 다양한 benchmark dataset에서 baseline보다 높은 성능을 보임

- 단점
  - 어떤 reflection token이 성능에 영향을 미쳤는지 보이지 않음: IsSUP만 포함되어있음
  - ChatGPT와 성능이 높은 경우도 있지만 10개 평가중 6개가 ChatGPT 기반 모델보다 성능이 높음
  - 이 방법이 좋다는 것을 더 보여주려면 Atlas와 같은 다른 RAG 방법들과 비교가 추가되었으면 함

Open review summary

- 장점
  - 기존 RAG의 한계 개선
  - self-reflection token을 활용한 Self-RAG 방법의 참신성
  - manual annotation에 의존하지 않음
  - 대부분의 경우 baseline보다 좋은 성능
- 단점
  - 실험 결과 보충이 필요하다.
    - Self-reflection과 RAG 중 어떤 것이 기여했는지 안보임
    - Self-reflection token의 ablation study 부족
    - 검색 threshold, self-reflection token 등 예측 성능 비교가 없음

Open Questions

- Self-RAG 방법(검색, 생성, 자기 반성 측면에서)을 프롬프트 엔지니어링을 통해 학습 없이 적용 가능한가?

Appendix

Details of beam-search score calculations. We first compute scores for each critique type by taking the normalized probabilities of desirable tokens. For $\boxed{\text{ISREL}}$, we compute the score as follows:

$$s(\boxed{\text{ISREL}}) = \frac{p(\boxed{\text{ISREL}} = \text{RELEVANT})}{p(\boxed{\text{ISREL}} = \text{RELEVANT}) + p(\boxed{\text{ISREL}} = \text{IRRELEVANT})}.$$

For $\boxed{\text{ISSUP}}$, we compute the score as follows:

$$s(\boxed{\text{ISREL}}) = \frac{p(\boxed{\text{ISSUP}} = \text{FULLY})}{S} + 0.5 \times \frac{p(\boxed{\text{ISSUP}} = \text{PARTIALLY})}{S},$$

where $S = \sum_{t \in \{\text{FULLY}, \text{PARTIALLY}, \text{NO}\}} p(\boxed{\text{ISSUP}} = t)$. For $\boxed{\text{ISUSE}}$ where we have a five-scale score, we compute the weighted sum of the scores. We assigns weighted scores of $w = \{-1, -0.5, 0, 0.5, 1\}$ to the tokens $\boxed{\text{ISUSE}}$ $=\{1, 2, 3, 4, 5\}$, and compute the final scores as follows:

$$s(\boxed{\text{ISUSE}}) = \sum_{i}^{5} w_i \frac{p(\boxed{\text{ISUSE}} = i)}{S},$$

where $S = \sum_{t \in \{1,2,3,4,5\}} p(\boxed{\text{ISUSE}} = t)$.

Details of adaptive retrieval. For retrieval based on soft constraints, we trigger retrieval if the following condition is satisfied:

$$\frac{p(\boxed{\text{Retrieve}} = \text{YES})}{p(\boxed{\text{Retrieve}} = \text{YES}) + p(p(\boxed{\text{Retrieve}} = \text{NO})} > \delta.$$

Appendix

IsRel	IsSup	IsUse	sequence	PopQA performance (acc.)
x	x	x	x	0.538
	x	x	x	0.536
		x	x	0.512
		x		0.416
			x	0.512

The number of passages (N=2)	PopQA performance (acc.)
2	0.498
3	0.504
5	0.504
7	0.540
10	0.538
15	0.528
20	0.520

	FLAN	Stanford Alpaca	NQ	FEVER
% of instances with [Retrieve]	15.8%	53.3%	87.7%	63.2%