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
{akari,zeqiuwu,yizhongw,hannaneh}@cs.washington.edu, avi@us.ibm.com

ICLR2024 8886

Motivation & Preliminary

- Why is Retrieval Augmented Generation (RAG) important?
  - Reduce factual errors
  - Improve helpfulness/usefulness
- Current Challenges of RAG

  1. May hinder the versatility of LLMs
  --- **Learn** when to retrieve

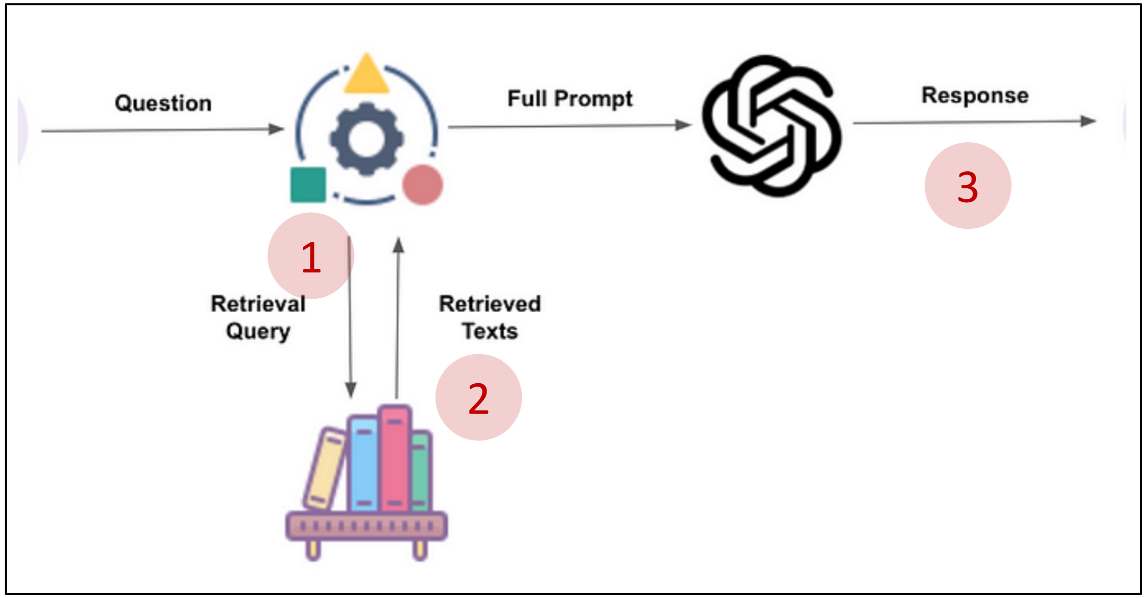
  2. May introduce unnecessary or off-topic passages

  --- **Learn** what to retrieve (Part of critique)

  3. (weak/small) LLMs may not know how to use the retrieved knowledge.

  --- **Learn** to generate (given retrieved data)

SELF-RAG: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION



Special Reflection Tokens to Learn

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.

x: query, y: output, d: documents

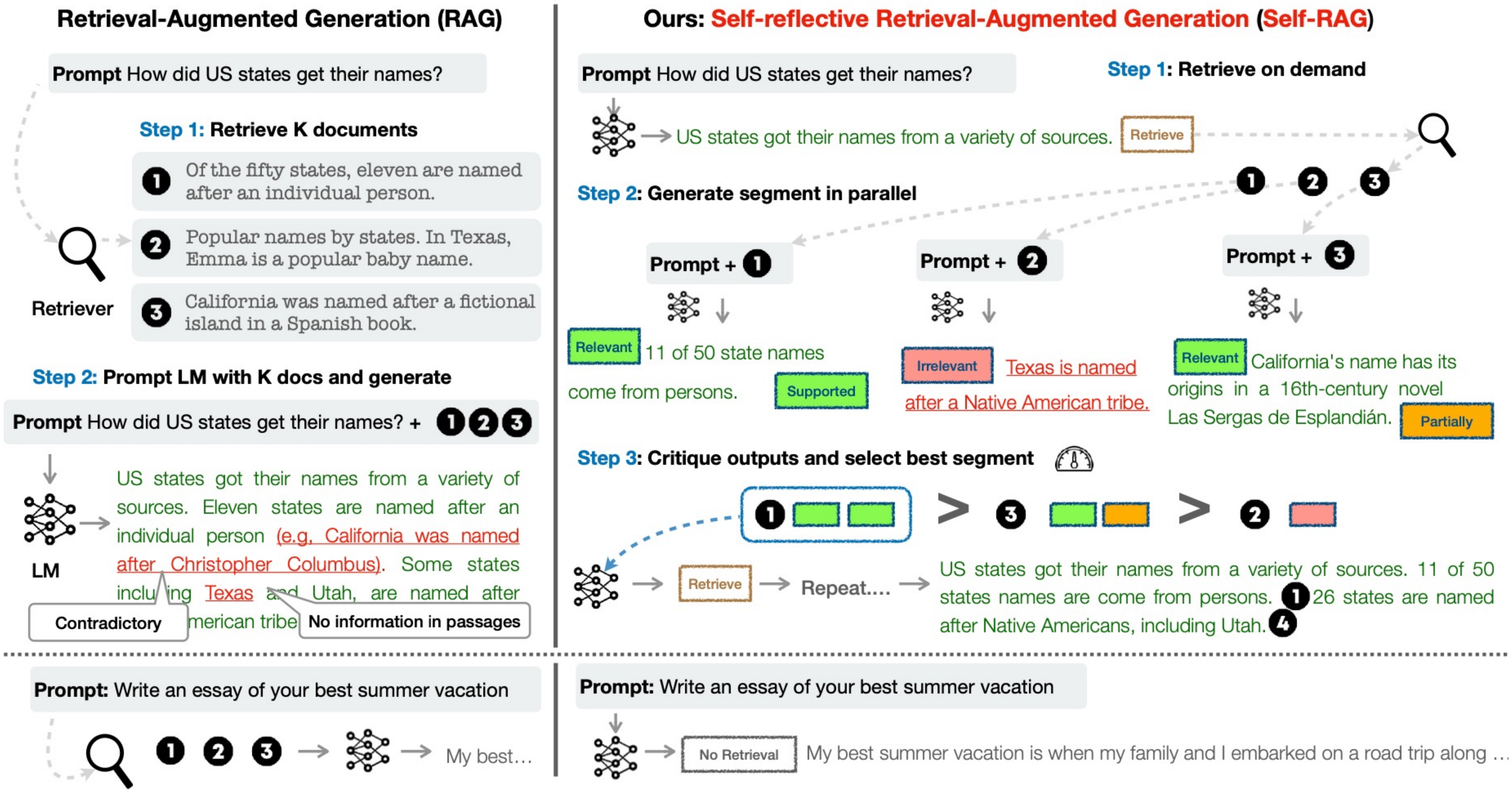


Figure 1: Overview of SELF-RAG. SELF-RAG learns to retrieve, critique, and generate text passages to enhance overall generation quality, factuality, and verifiability.

Method: Training

Learning to Retrieve\Critique

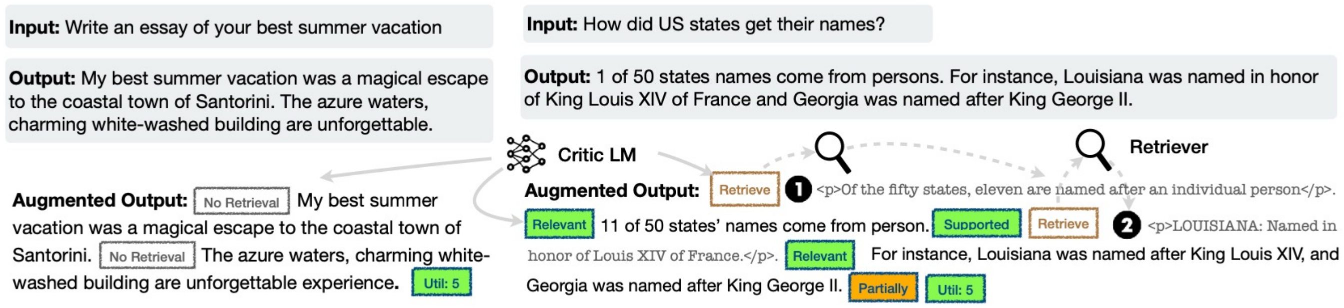
- Why?
  - Call GPT4 API is too expensive and hard to reproduce.
- Solution
  - Knowledge Distill from GPT4

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

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.

Learning to Generate

- Use the learned Critic to augment output
- Then SFT $\quad \max_{\mathcal{M}} \mathbb{E}_{(x,y,r)\sim\mathcal{D}_{gen}} \log p_{\mathcal{M}}(y, r|x).$



Dataset name	category	Data source	the number of instances
GPT-4 Alpaca	Instruction-following	Open-Instruct	26,168
Stanford Alpaca	Instruction-following	Open-Instruct	25,153
FLAN-V2	Instruction-following	Open-Instruct	17,817
ShareGPT	Instruction-following	Open-Instruct	13,406
Open Assistant 1	Instruction-following	Open-Instruct	9,464
Wizard of Wikipedia	Knowledge-intensive	KILT	17,367
Natural Questions	Knowledge-intensive	KILT	15,535
FEVER	Knowledge-intensive	KILT	9,966
OpenBoookQA	Knowledge-intensive	HF Dataset	4,699
Arc-Easy	Knowledge-intensive	HF Dataset	2,147
ASQA	Knowledge-intensive	ASQA	3,897

Total: 150k

Algorithm 2 Self-Rag Training

```
1: Input input-output data D = {X, Y}, generator M, C θ
2: Initialize C with a pre-trained LM
3: Sample data {X^sample, Y^sample} ~ {X, Y}        ▷ Training Critic LM (Section 3.2.1)
4: for (x, y) ∈ (X^sample, Y^sample) do             ▷ Data collections for C
5:     Prompt GPT-4 to collect a reflection token r for (x, y)
6:     Add {(x, y, r)} to D_critic
7: Update C with next token prediction loss          ▷ Critic learning; Eq. 1
8: Initialize M with a pre-trained LM                ▷ Training Generator LM (Section 3.2.2)
9: for (x, y) ∈ (X, Y) do                           ▷ Data collection for M with D_critic
10:    Run C to predict r given (x, y)
11:    Add (x, y, r) to D_gen
12: Update M on D_gen with next token prediction loss   ▷ Generator LM learning; Eq. 2
```

Method: Inference

- Step1: Adaptive retrieval with threshold.
- Step2: Call retriever R
- Step3: Tree-decoding with critique tokens.
  - Beam-Search

$$f(y_t, d, \boxed{\text{Critique}}) = p(y_t|x, d, y_{<t})) + \mathcal{S}(\boxed{\text{Critique}}), \text{where} \tag{3}$$

$$\mathcal{S}(\boxed{\text{Critique}}) = \sum_{G \in \mathcal{G}} w^G s_t^G \text{ for } \mathcal{G} = \{\boxed{\text{IsRel}}, \boxed{\text{IsSup}}, \boxed{\text{IsUse}}\}, \tag{4}$$

where $s_t^G = \frac{p_t(\hat{r})}{\sum_{i=1}^{N^G} p_t(r_i)}$ stands for the generation probability of the most desirable reflection token $\hat{r}$ (e.g., $\boxed{\text{IsRel}}$ =Relevant) for the critique token type G with N^G distinct tokens (that represent

Senstive to w, a customizable feature

Algorithm 1 SELF-RAG Inference

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

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

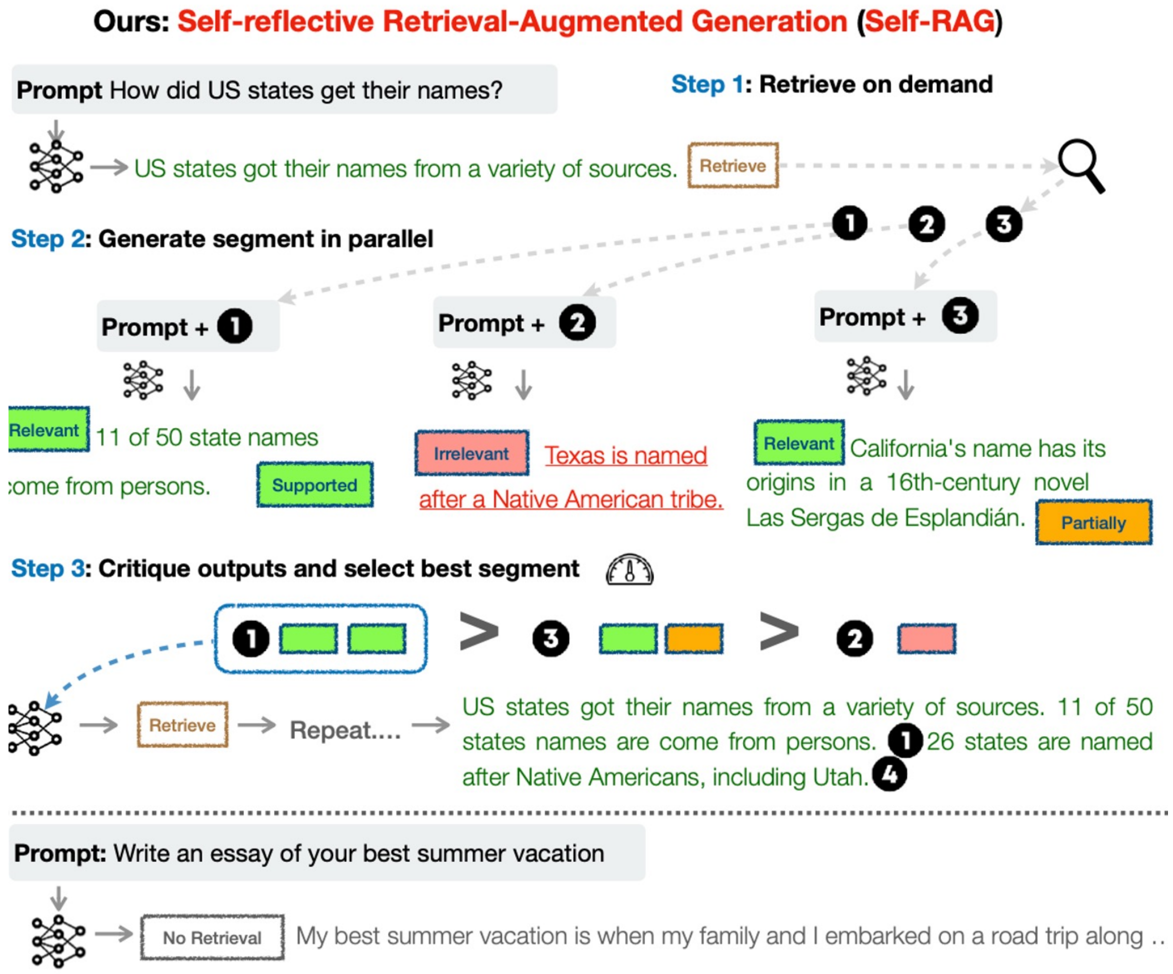
$\mathcal{M}$ predicts $\boxed{\text{IsSup}}$ and $\boxed{\text{IsUse}}$ given x, y_t, d for each $d \in \mathbf{D}$ ▷ Critique

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



Experiments: Main results

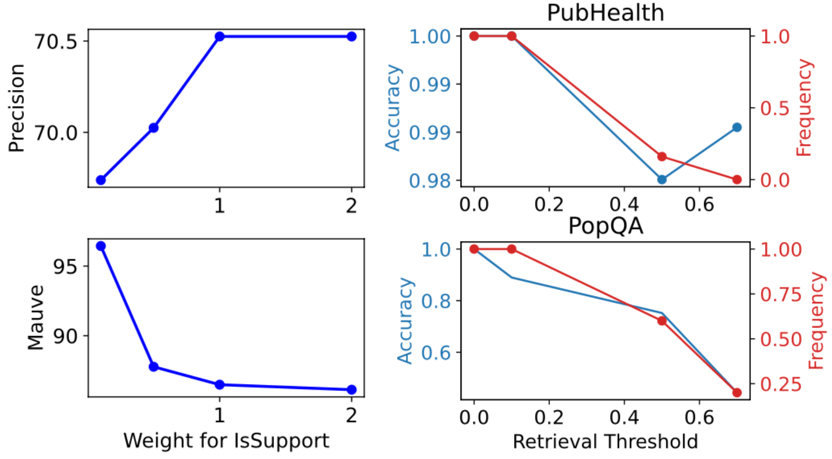
Setting:
Critic C: 7b-base
Generater M: 7b and 13b

	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**

Experiments: Ablations

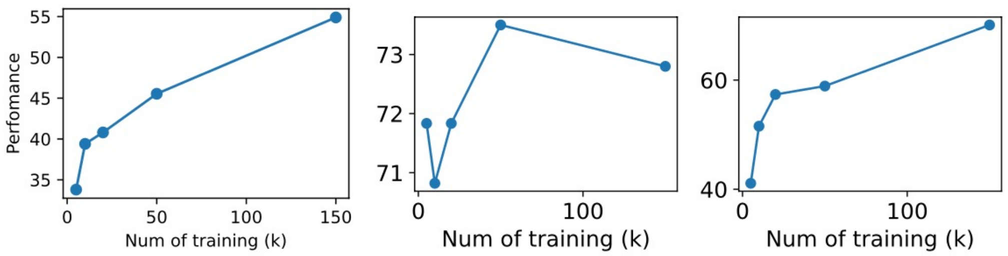
	PQA (acc)	Med (acc)	AS (em)
SELF-RAG (50k)	45.5	73.5	32.1
Training			
No Retriever $\mathcal{R}$	43.6	67.8	31.0
No Critic $\mathcal{C}$	42.6	72.0	18.1
Test			
No retrieval	24.7	73.0	–
Hard constraints	28.3	72.6	–
Retrieve top1	41.8	73.1	28.6
Remove ISSUP	44.1	73.2	30.6

(a) Ablation



(b) Customization (c) Retrieval

Figure 3: **Analysis on SELF-RAG:** (a) **Ablation studies** for key components of SELF-RAG training and inference based on our 7B model. (b) **Effects of soft weights** on ASQA citation precision and Mauve (fluency). (c) **Retrieval frequency** and *normalized* accuracy on PubHealth and PopQA.



(a) PopQA (b) PubHealth (c) ASQA (prec)

	Pop	Bio.
S & P	92.5	70.0
ISREL	95.0	90.0
ISSUP	90.0	85.0

(d) Human evaluation on PopQA and Bio generation.

Figure 4: **Training scale and Human analysis:** (a) (b) (c) **Training scale analysis** shows the effect of the training data scale on PopQA, PubHealth and ASQA (citation precision), respectively. (d) **Human analysis** on SELF-RAG outputs as well as reflection tokens.