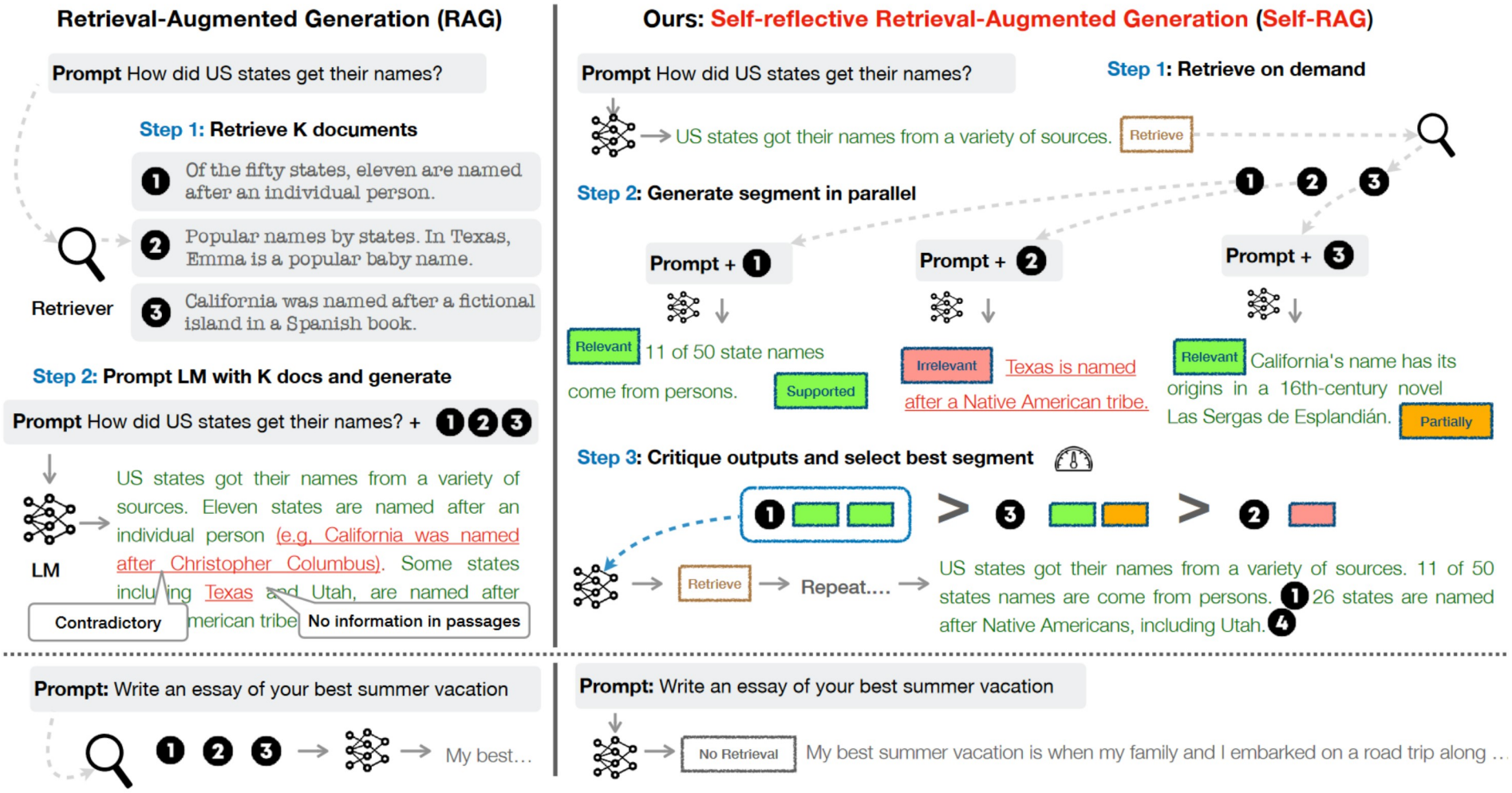
Self RAG

marii

Self-RAG High level Overview



Speaker notes

This is a speaker note

New 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.

image.png

Inference Algorithm

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 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

image.png

Inference Example

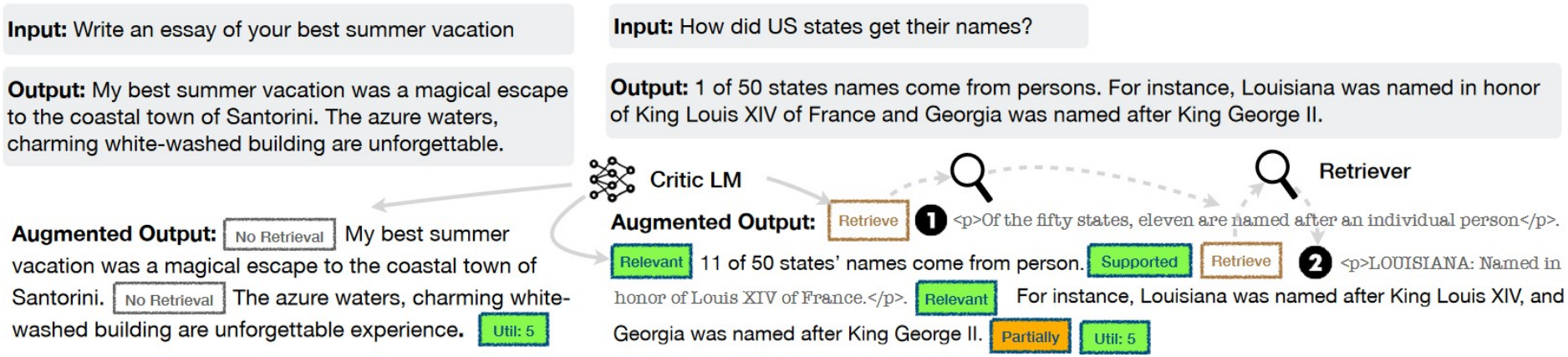


image.png

Training Algorithm

Algorithm 2 SELF-RAG Training

Input input-output data $\mathcal{D} = \{X, Y\}$, generator $\mathcal{M}$, $\mathcal{C}$ θ
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

image.png

Speaker notes

Code for critic missing(accidently deleted?) from github. It is initialized from Llama2-7B, but I think trained with a classification head.

Tree-Decoding with Critique Tokens

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

$$\mathcal{S}(\boxed{\text{Critique}}) = \sum_{G \in \mathcal{G}} w^G s_t^G \text{ for } \mathcal{G} = \{\boxed{\text{ISREL}}, \boxed{\text{ISSUP}}, \boxed{\text{ISUSE}}\},$$

image.png

Speaker notes

This is used in a beam search to evaluate beams for multiple t. So a beam's score would be similar to $\sum_{t=0}^{T} f(y_t, d, Critque)$

Example of how it works

Input when did walking dead season 7 come out (from nq)

Output [Retrieve=Yes]<p>The Walking Dead (season 7)–The Walking Dead (season 7) The seventh season of "The Walking Dead", an American post-apocalyptic horror television series on AMC, premiered on October 23, 2016, and concluded on April 2, 2017, consisting of 16 episodes. Developed for television by Frank Darabont, the series is based on the eponymous series of comic books by Robert Kirkman, Tony Moore, and Charlie Adlard. ...<p>[IsRel =Relevant]October 23 , 2016[IsSup =Fully Supported][IsUse =5]

image.png

Generating Retrieve Tokens

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.

image.png

Generating IsRel Tokens

Instruction age to run for US House of Representatives
Evidence The Constitution sets three qualifications for service in the U.S. Senate: age (at least thirty years of age); U.S. citizenship (at least nine years); and residency in the state a senator represents at the time of election.
Rating [Irrelevant]
Explanation The evidence only discusses the ages to run for the US Senate, not for the House of Representatives.

image.png

Generating IsSup Tokens

Instruction Explain the use of word embeddings in Natural Language Processing.
Preceding sentences Word embeddings are one of the most powerful tools available for Natural Language Processing (NLP). They are mathematical representations of words or phrases in a vector space, allowing similarities between words and the context in which they are used to be measured.
Output Word embeddings are useful for tasks such as sentiment analysis, text classification, predicting the next word in a sequence, and understanding synonyms and analogies.
Evidence Word embedding
Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. Conceptually it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with a much lower dimension. Methods to generate this mapping include neural networks, dimensionality reduction on the word co-occurrence matrix, probabilistic models, explainable knowledge base method, and explicit representation in terms of the context in which words appear. Word and phrase embeddings, when used as the underlying input representation, have been shown to boost the performance in NLP tasks such as syntactic parsing, sentiment analysis, next token predictions as well and analogy detection.
Score [Fully supported]
Explanation The output sentence discusses the application of word embeddings, and the evidence mentions all of the applications syntactic parsing, sentiment analysis, next token predictions as well as analogy detection as the applications. Therefore, the score should be [Fully supported].

image.png

Generating IsUse Tokens

Instruction Given a description of a travel destination, recommend 10 tourist attractions with detailed descriptions of each. The travel destination is Tokyo, Japan.
Output Tokyo is a vibrant city full of exciting tourist attractions. Some must-see sights include the Tokyo Skytree, Tokyo Disneyland, Sensoji Temple, Meiji Shrine, Tsukiji Fish Market, Harajuku, and Shinjuku Gyoen.
Perceived utility 3
Explanation This output doesn't provide descriptions of each attraction and the number of the attractions is also less than 10. While this output partially answers the instructions, it doesn't match the instructions strictly.

image.png

Results

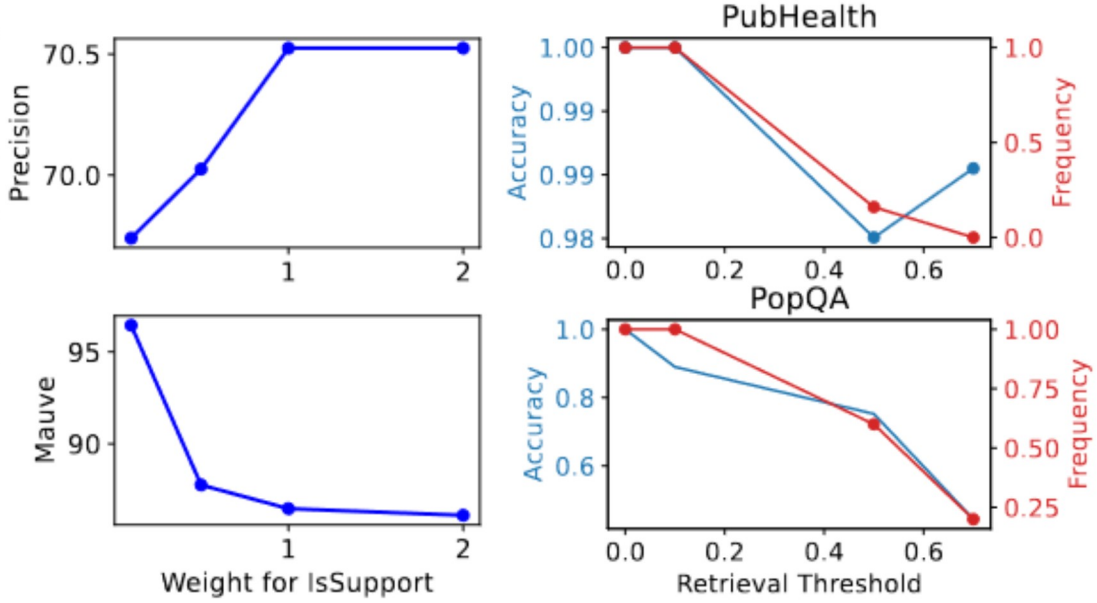
LM	Short-form PopQA (acc)	Short-form TQA (acc)	Closed-set Pub (acc)	Closed-set ARC (acc)	Long-form generations (with citations) Bio (FS)	ASQA (em)	ASQA (rg)	ASQA (mau)	ASQA (pre)	ASQA (rec)
LMs with proprietary data										
Llama2-c$_{13B}$	20.0	59.3	49.4	38.4	55.9	22.4	29.6	28.6	–	–
Ret-Llama2-c$_{13B}$	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**

image.png

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

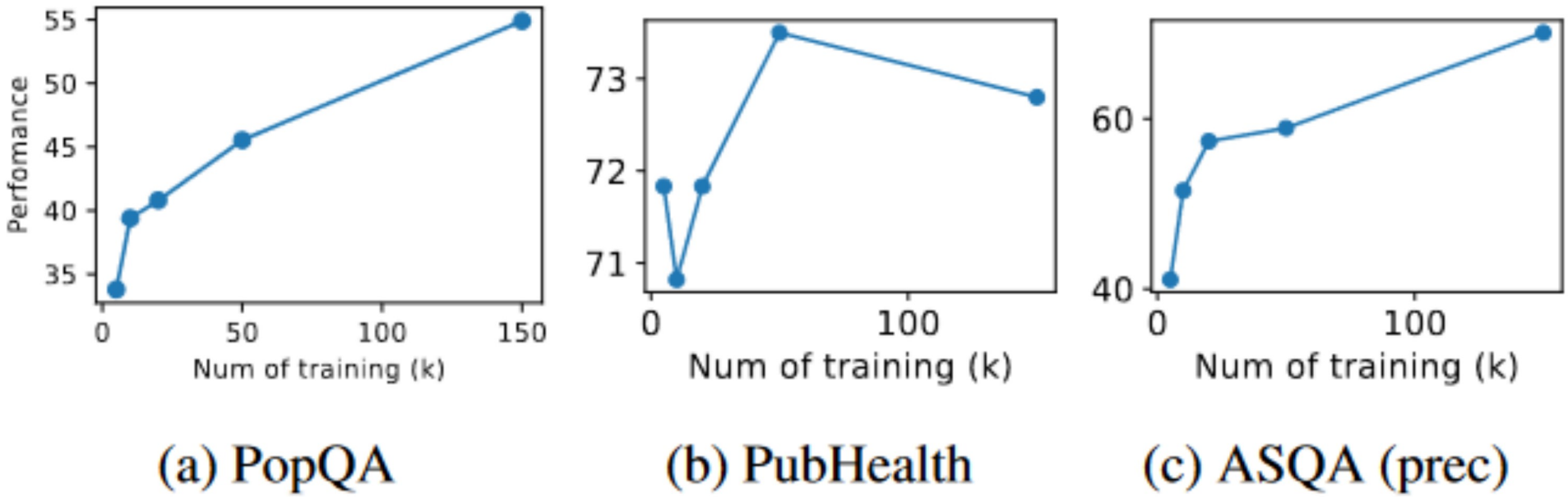
Figure 3: **Analysis on SELF-RAG:** (a) **Ablation studies** for key components of SELF-RAG training

Speaker notes

Training were ablations done during training. No retriever is if we just don't do RAG. No Critic, we only retrieve the first document asked for.

Test is done during inference. No retrieval is forcing Self-Rag model to not use retrieval. Hard constraints means model must predict exactly the "Retrieve" token out of all tokens, instead of at a particular probability threshold. Retrieve1 always retrieves top1 document, similar to regular RAG. Remove "IsSup" removes the token saying if the retrieved documented doesn't,partially, or fully supports the answer.

Supports more training



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

image-2.png

Speaker notes

Somtimes I get questions on how long to train models or should we collect more data. You can select small subsets of your data, or select smaller number of training steps. This can be used to determine optimal next steps.

Human Eval

	Pop	Bio.
S & P	92.5	70.0
IsRel	95.0	90.0
IsSup	90.0	85.0

Speaker notes

S&P is plausible and supported

IsREL+IsSup, human evaluators found these to both rated correctly.

IsSup- is interesting, suggesting the model actually outputs that it did not get good support from the retriever.

Questions?