

# FLARE: Active Retrieval Augmented Generation

**arXiv, 2023**

330 citations

Carnegie Mellon University, Sea AI Lab, Meta

01.     **INTRODUCTION**

02.     **METHOD**

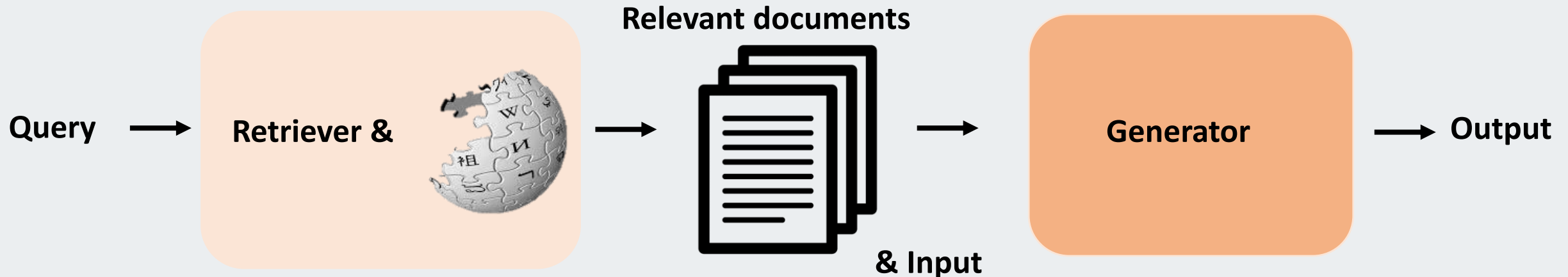
03.     **EXPERIMENTS**

04.     **LIMITATIONS**

# INTRODUCTION

## Retrieval Augmented Generation

- To prevent hallucinations, external information is incorporated
- Retriever
  - Bring knowledge related to query
  - E.g., BM25, Search Engine
- Generator
  - Generative models trained on large corpus with numerous parameters
  - E.g., BART, GPT-3



# INTRODUCTION

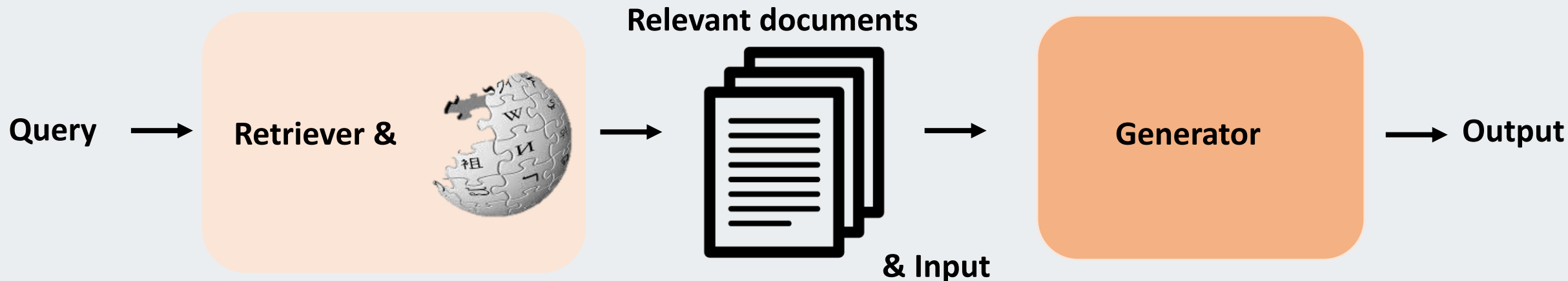
## Retrieval Augmented Generation

- To prevent hallucinations, external information is incorporated
- Retriever
  - Bring knowledge related to query
  - E.g., BM25, Search Engine
- Generator
  - Generative models trained on large corpus with numerous parameters
  - E.g., BART, GPT-3

$$TF(t, d) = \frac{\text{문서 } d \text{ 에서 단어 } t \text{ 가 등장한 횟수}}{\text{문서 } d \text{ 에 등장한 모든 단어의 수}}$$

$$IDF(t, D) = \log \left( \frac{\text{총 문서의 개수}}{\text{단어 } t \text{ 를 포함하는 문서의 수}} \right)$$

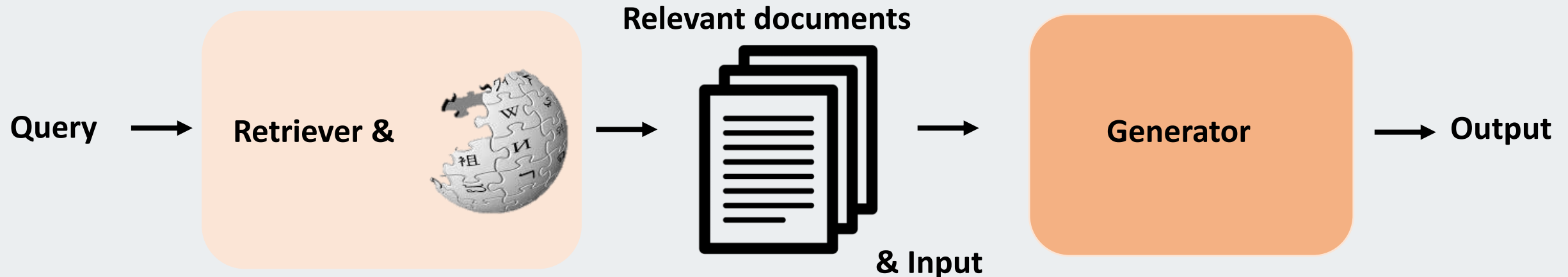
$$TF-IDF(t, d, D) = TF(t, d) * IDF(t, D)$$



# INTRODUCTION

## Retrieval Augmented Generation

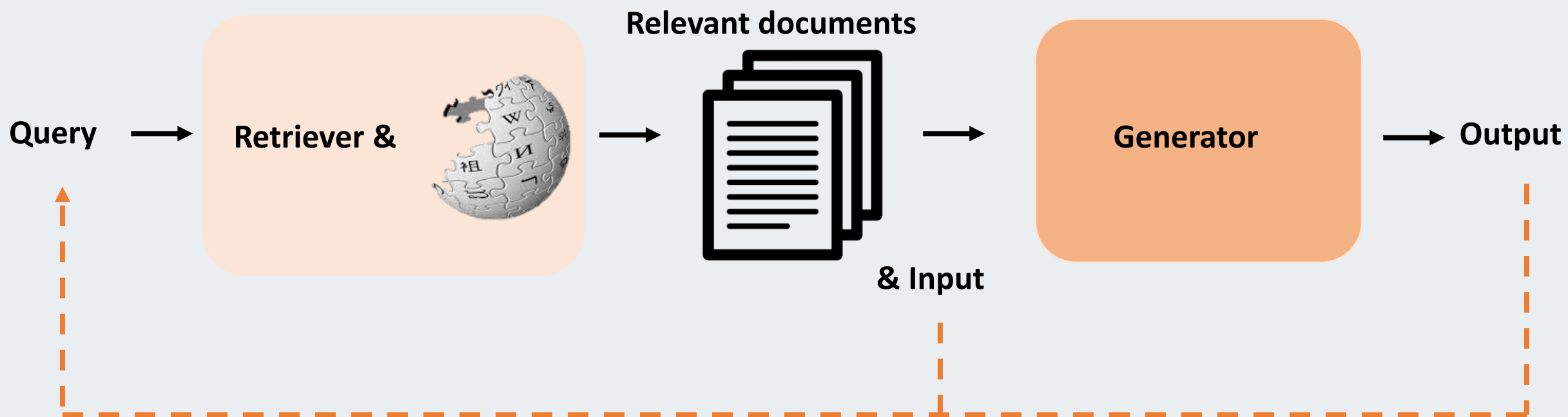
- To prevent hallucinations, external information is incorporated
- Retriever
  - Bring knowledge related to query
  - E.g., BM25, Search Engine
- Generator
  - Generative models trained on large corpus with numerous parameters
  - E.g., BART, GPT-3



# INTRODUCTION

## Retrieval Augmented Generation

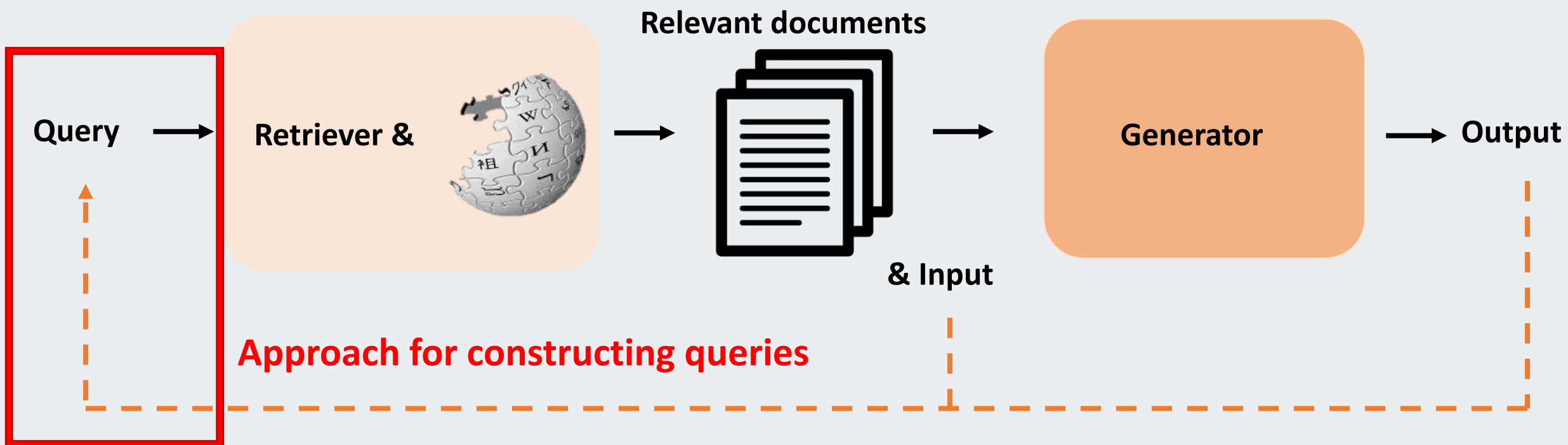
- Single-time retrieval
- **Multi-time retrieval**



# INTRODUCTION

## Retrieval Augmented Generation

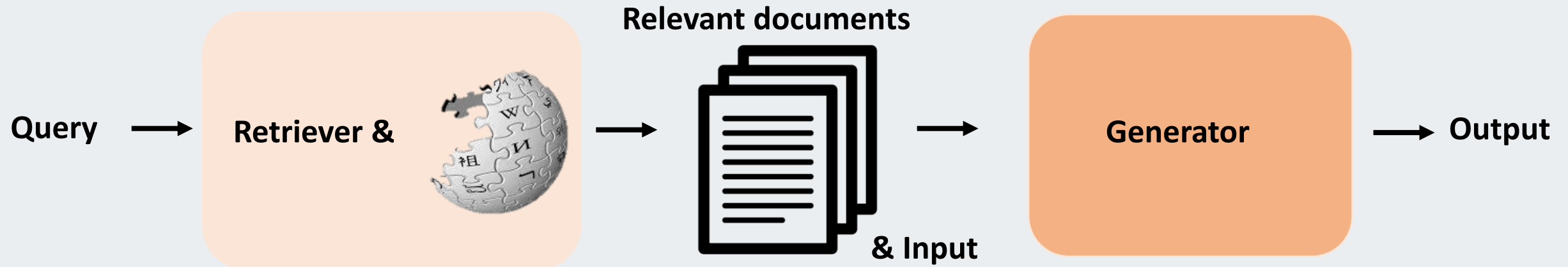
- Single-time retrieval
- **Multi-time retrieval**



# INTRODUCTION

## Retrieval Augmented Generation

- Single-time retrieval: unsuitable for long-text generation task
  - Initial retrieval cannot address all aspects/details of the topic

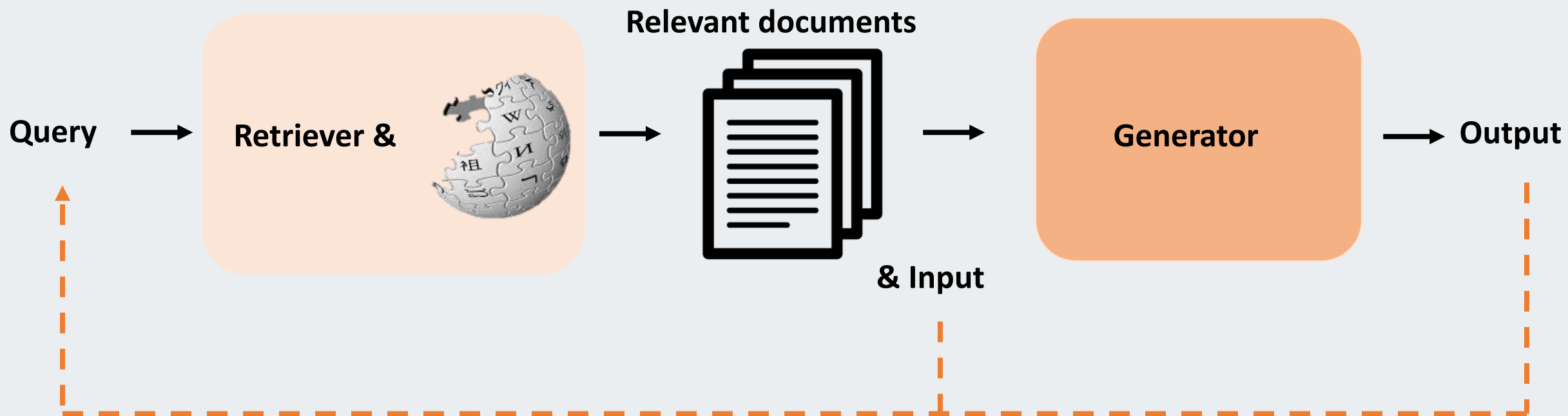




# INTRODUCTION

## Retrieval Augmented Generation

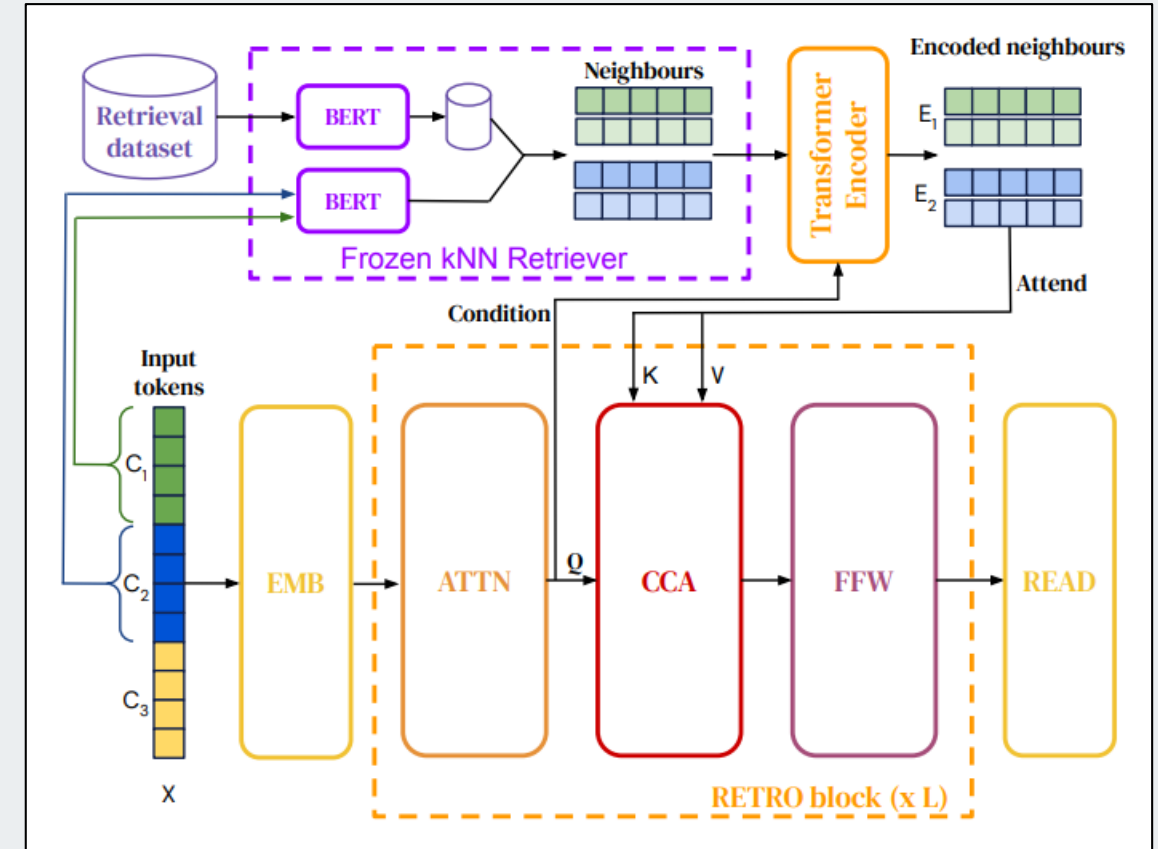
- Multi-time retrieval
  - Previous window
  - Previous sentence
  - Question decomposition



# INTRODUCTION

## Retrieval Augmented Generation

- Previous window (RETRO, 2022, DeepMind)
  - Given window size, use input sequence in window size as the queries



# INTRODUCTION

## Retrieval Augmented Generation

- Previous window (RETRO, 2022, DeepMind)
  - Given window size, use input sequence in window size as the queries

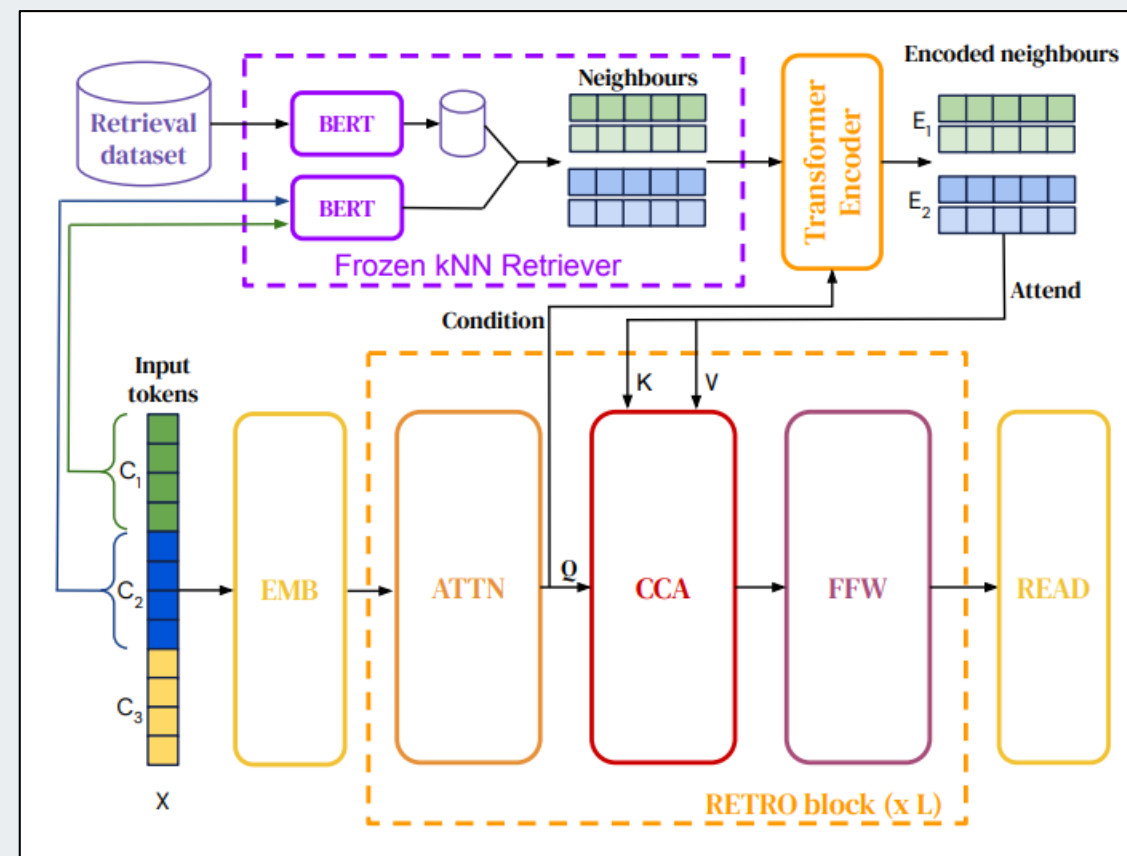
“Given window size, use input sequence in ...”

Chunk 1

Chunk 2

Chunk 3

**Retrieval**



# INTRODUCTION

## Retrieval Augmented Generation

- Previous window (RETRO, 2022, DeepMind)
  - Given window size, use input sequence in window size as the queries
  - Predict next token using previous tokens and external info from previous chunk

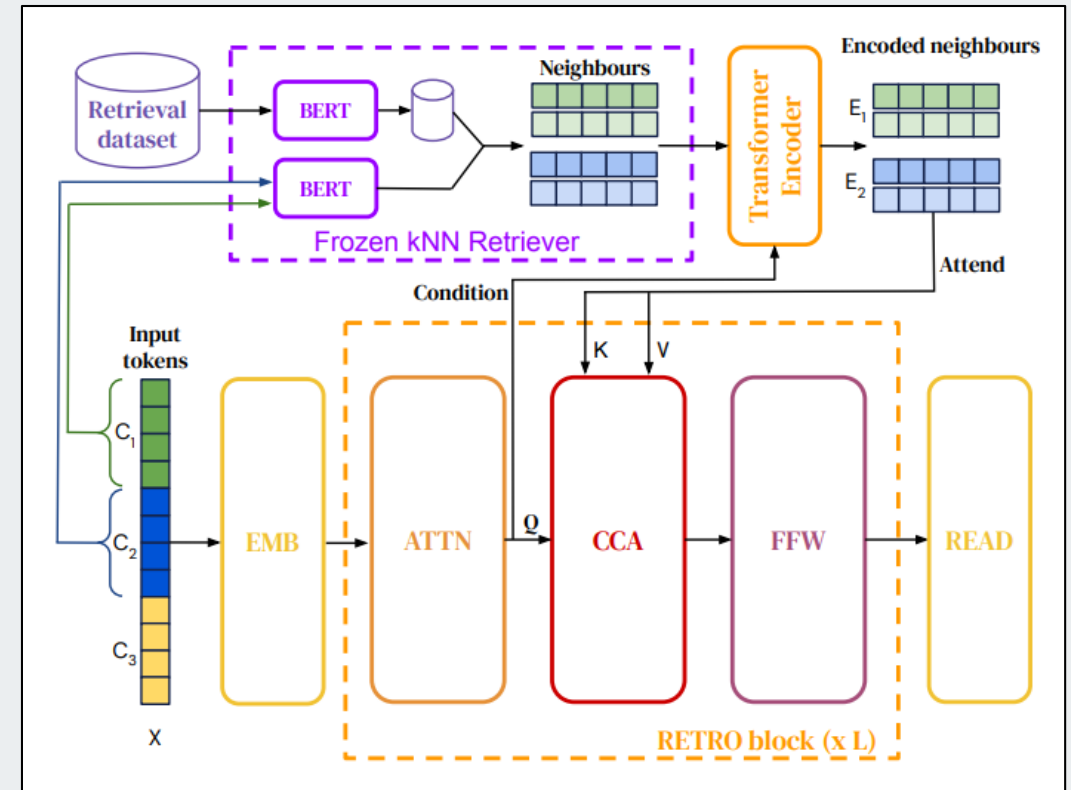
“Given window size, use input sequence in ...”

Chunk 1

Chunk 2

Chunk 3

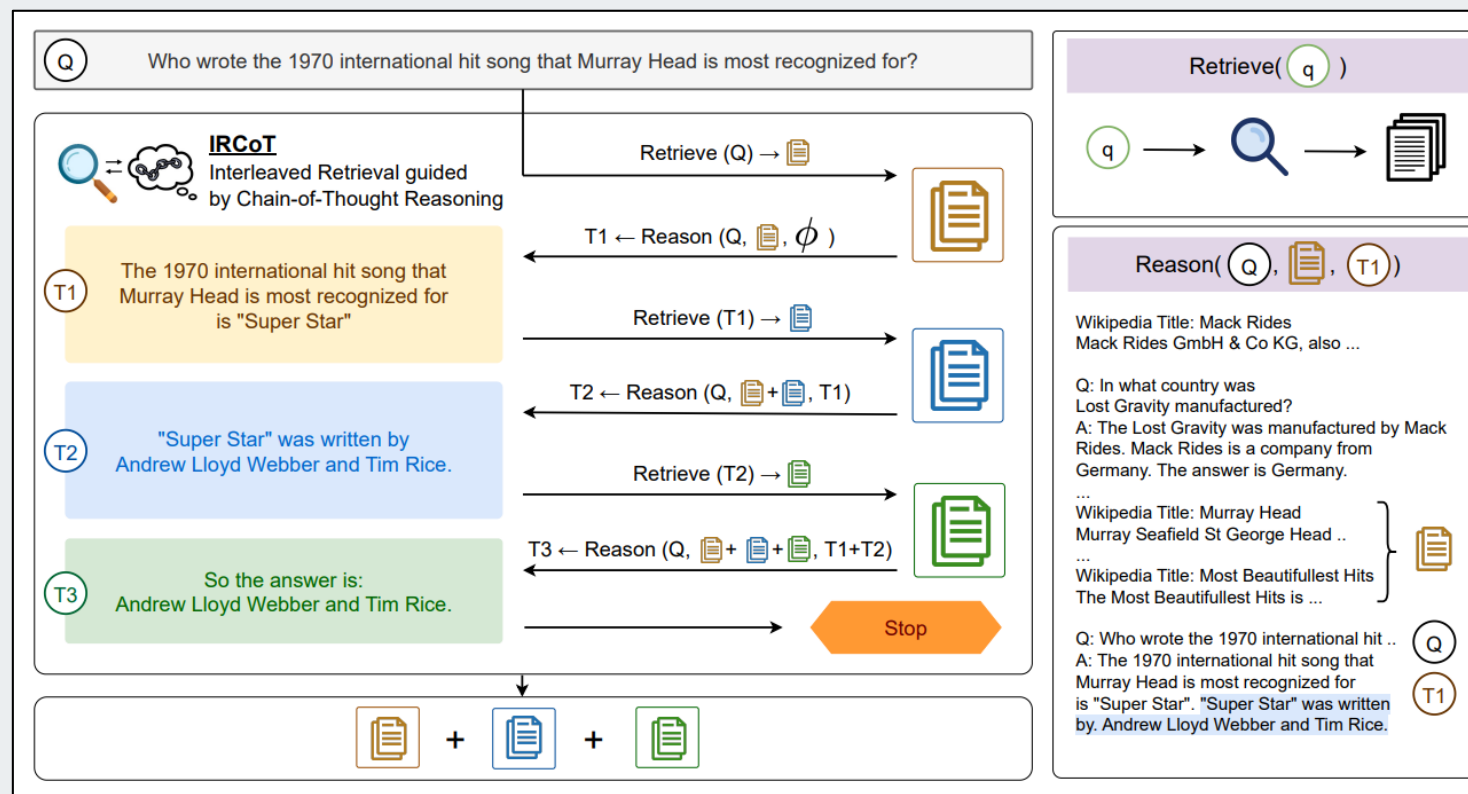
**Retrieval**



# INTRODUCTION

## Retrieval Augmented Generation

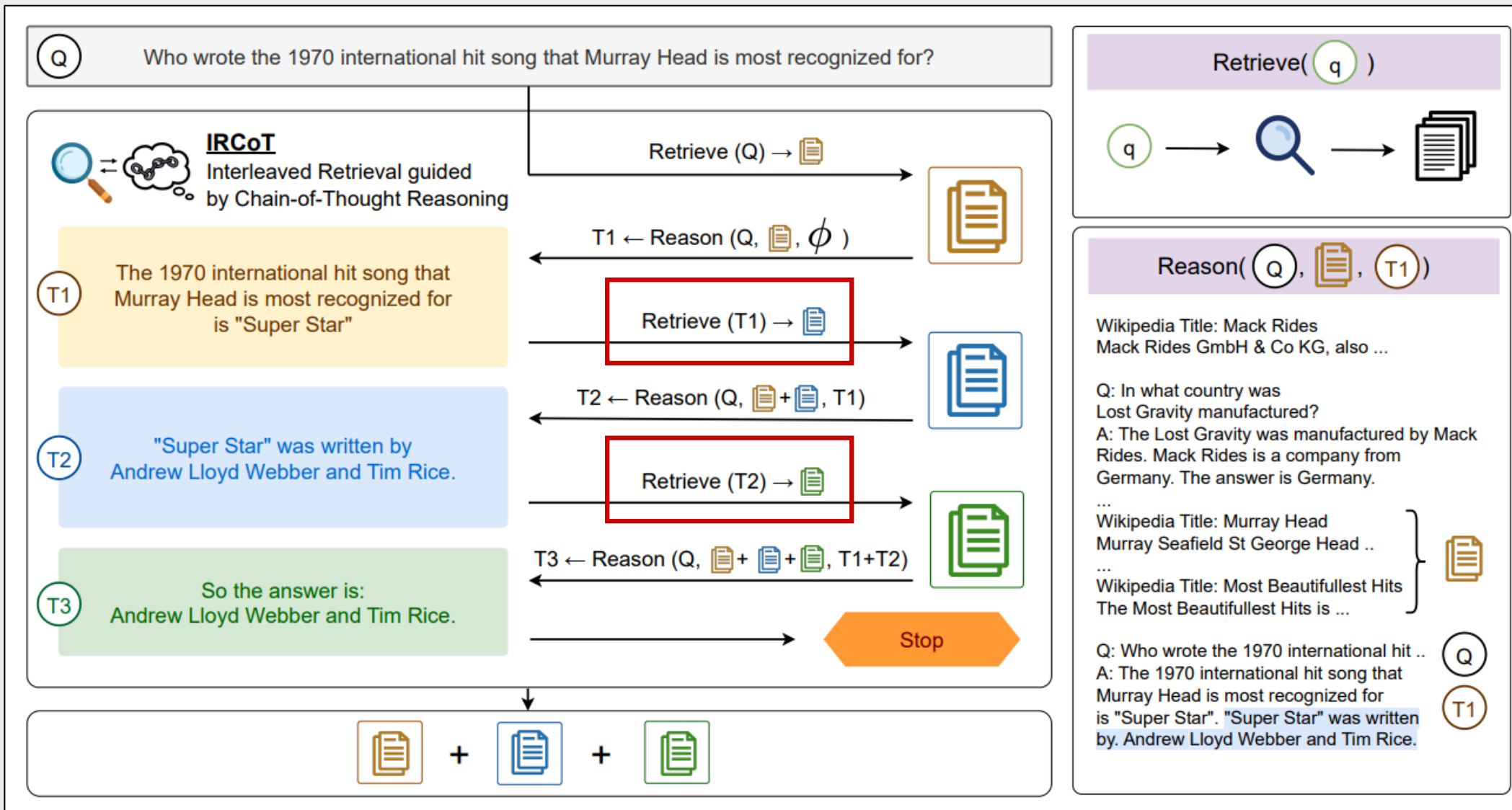
- Previous sentence (IR-CoT, 2023, Stony Brook University)
  - Use previous sentences as queries
  - Specifically, retrieve-step is based on the last CoT sentence



# INTRODUCTION

## Retrieval Augmented Generation

- 



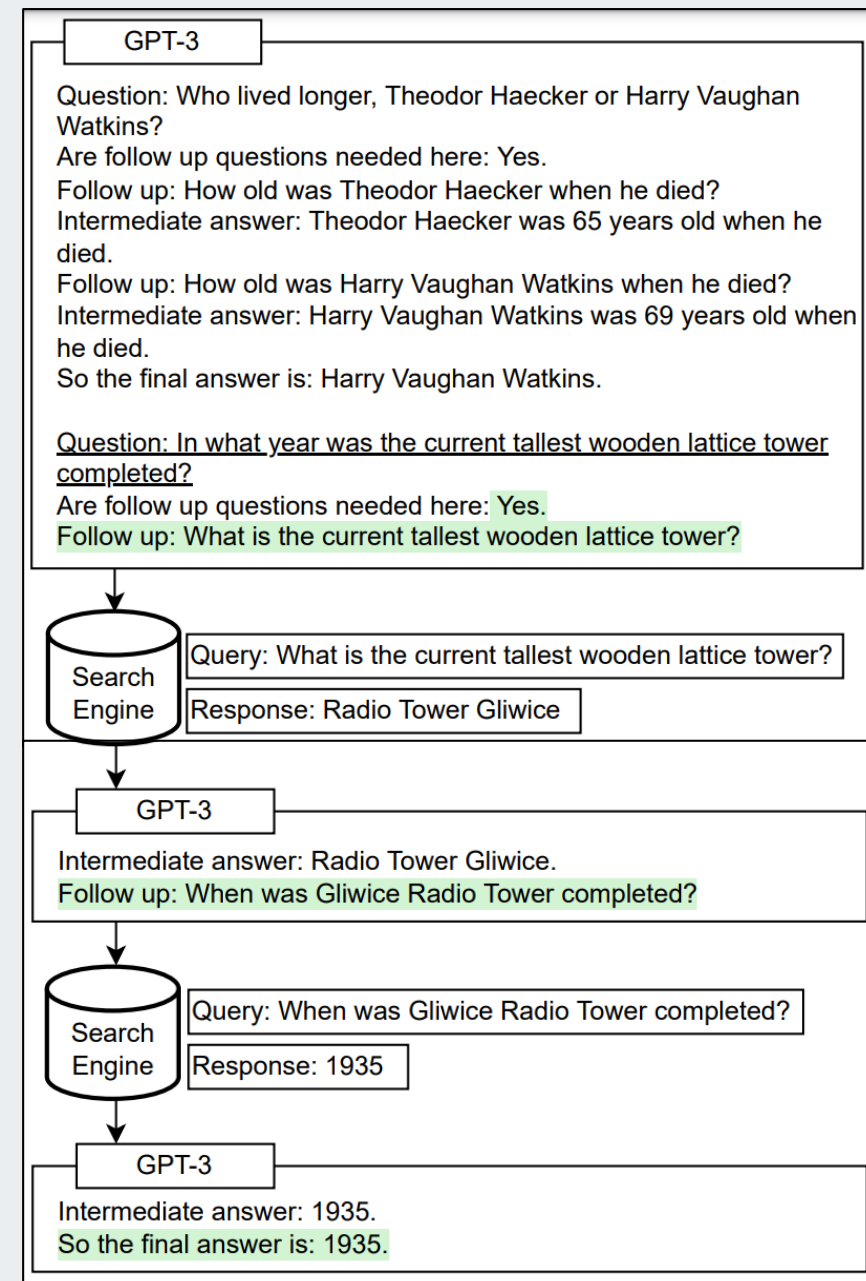
# INTRODUCTION

## Retrieval Augmented Generation

- Question decomposition

(Self-Ask, 2023, University of Washington & Meta)

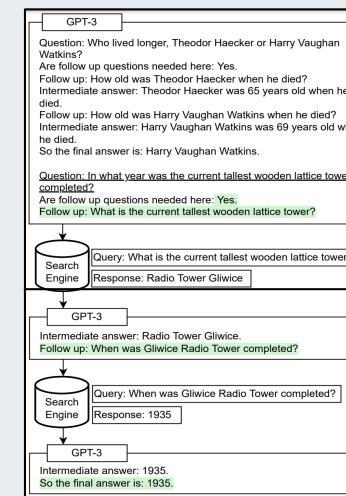
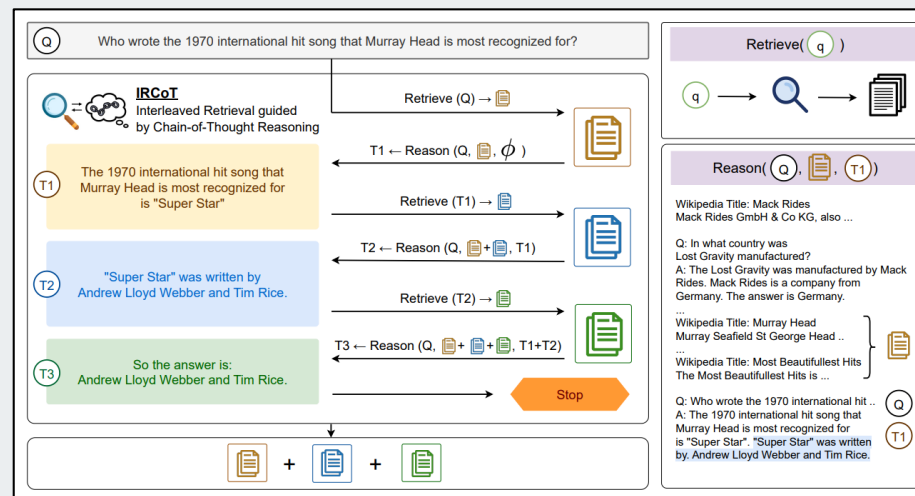
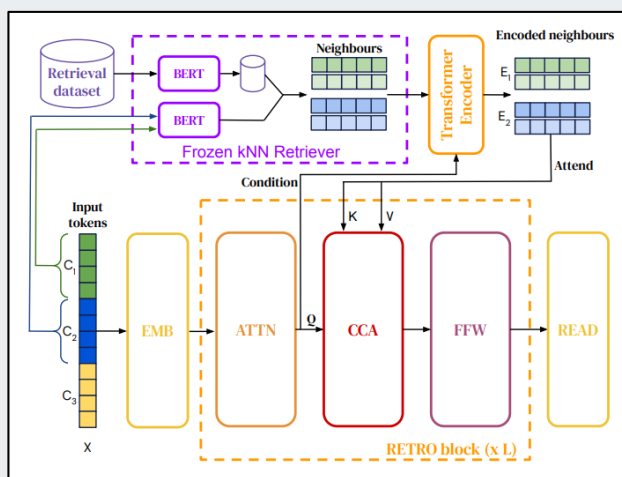
- To guide LMs to generate decomposed sub-questions, use manually annotated task-specific examples
- “Follow up” generated by model is query



# INTRODUCTION

## Retrieval Augmented Generation

- Multi-time retrieval
  - Previous window: retrieve at inappropriate times due to fixed interval
  - Previous sentence
    - ✓ queries (last sentence of CoT) may fail to reflect the content to be retrieved
  - Question decomposition
    - ✓ since task-specific prompt engineering is required, it is difficult to generalize





# METHOD

## FLARE: Forward-Looking Active Retrieval Augmented Generation

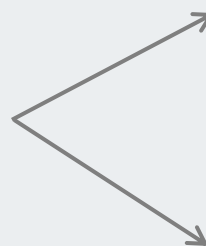
- Intuition
  - LMs retrieve information only when additional knowledge is needed
  - Search queries reflect the intention of generation
  - It's similar to how humans create papers, essays, and books

*FLARE<sub>instruct</sub>*

*FLARE<sub>direct</sub>*

Implicit query

Explicit query



# METHOD

## *FLARE<sub>instruct</sub>*

- When LMs generate “[Search (query)]”:
  1. Stop generation
  2. Retrieve documents using query

How to generate “[Search (query)]”

Instruction about downstream task

Combine skill 1 & 2

### Prompt 3.1: retrieval instructions

Skill 1. An instruction to guide LMs to generate search queries.  
Several search-related exemplars.

Skill 2. An instruction to guide LMs to perform a specific downstream task (e.g., multihop QA).  
Several task-related exemplars.

An instruction to guide LMs to combine skills 1 and 2 for the test case.  
The input of the test case.

# METHOD

## *FLARE<sub>instruct</sub>*

Skill 1. Use the Search API to look up relevant information by writing “[Search(term)]” where “term” is the search term you want to look up. For example:

Question: But what are the risks during production of nanomaterials?

Answer (with Search): [Search(nanomaterial production risks)] Some nanomaterials may give rise to various kinds of lung damage.

Skill 2. Answer questions by thinking step-by-step. First, write out the reasoning steps, then draw the conclusion. For example:

Question: When did the director of film Hypocrite (Film) die?

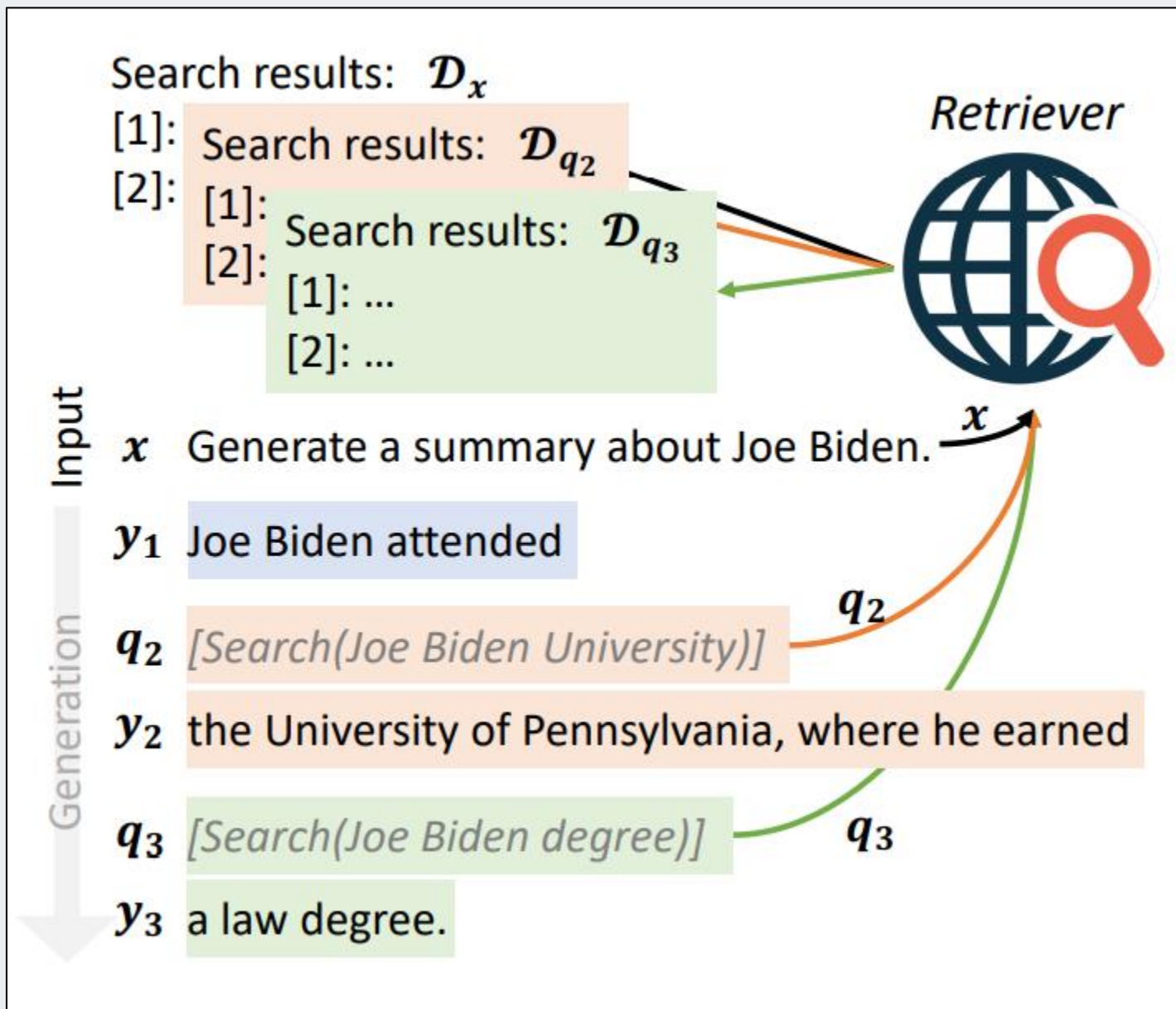
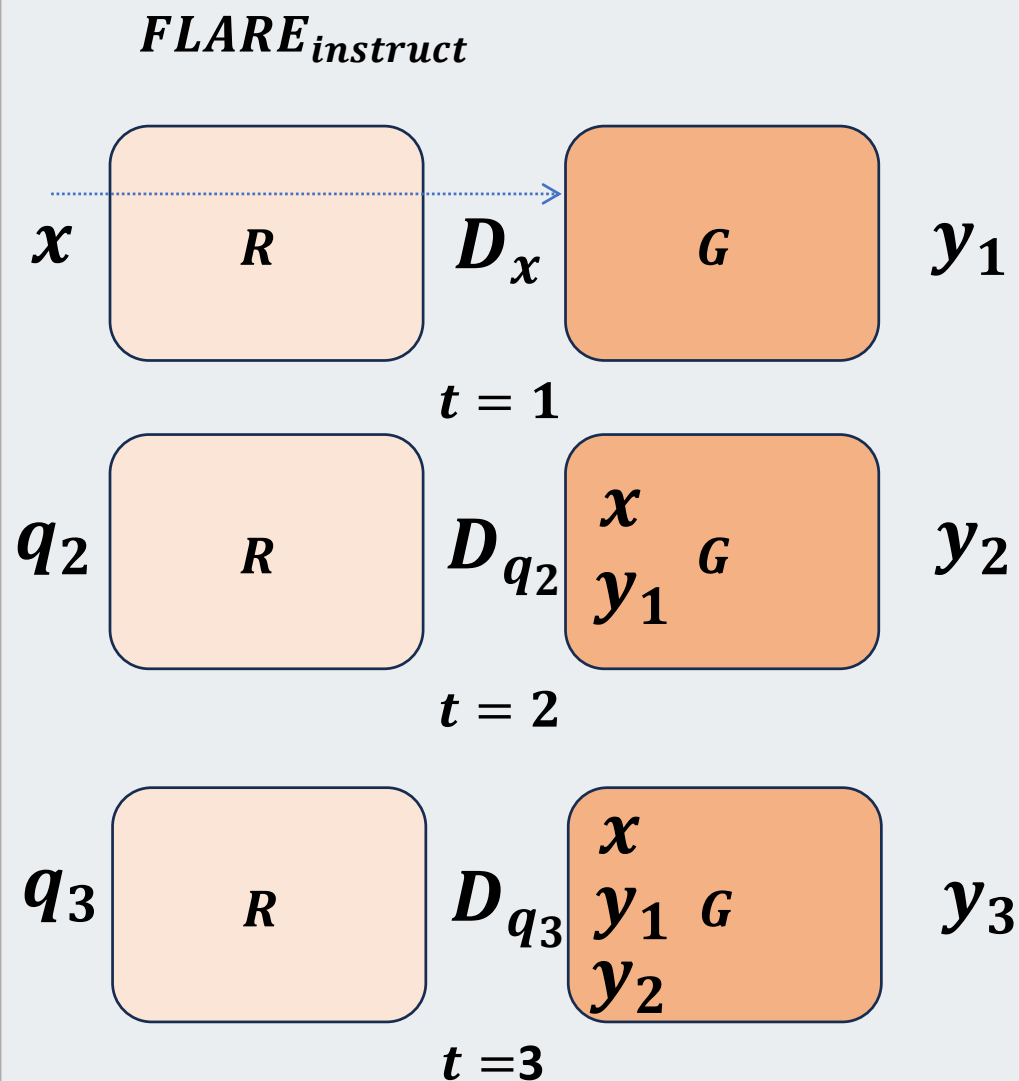
Answer (with step-by-step): The film Hypocrite was directed by Miguel Morayta. Miguel Morayta died on 19 June 2013. So the answer is 19 June 2013.

Now, combine the aforementioned two skills. First, write out the reasoning steps, then draw the conclusion, where the reasoning steps should also utilize the Search API “[Search(term)]” whenever possible.

Question: Where did Minbyauk Thihapate’s wife die?

Answer (with step-by-step & Search):

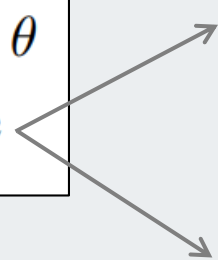
# METHOD



# METHOD

## *FLARE*<sub>direct</sub>

- The output of LMs is directly used as queries
- First, generates temporary next sentence without relevant documents
  - $\hat{s}_t = LM([x, y_{<t}])$
  - The confidence of each token in the sentence is assessed to determine whether to trigger a retrieval

$$y_t = \begin{cases} \hat{s}_t & \text{if all tokens of } \hat{s}_t \text{ have probs } \geq \theta \\ s_t = LM([\mathcal{D}_{q_t}, x, y_{<t}]) & \text{otherwise} \end{cases}$$


The diagram shows two arrows originating from the right side of the equation box. The top arrow points to the text 'Implicit query' and the bottom arrow points to the text 'Explicit query'.

**Implicit query**

**Explicit query**

# METHOD

$FLARE_{direct}$

$\beta: [0, 1]$

$\hat{S}_t$

Joe Biden attended the University of Pennsylvania,  
where he earned a law degree.

*implicit query  
by masking*

*explicit query by  
question generation*

Joe Biden attended , where he earned .

Ask a question to which the answer is "the University of Pennsylvania"

Ask a question to which the answer is "a law degree"

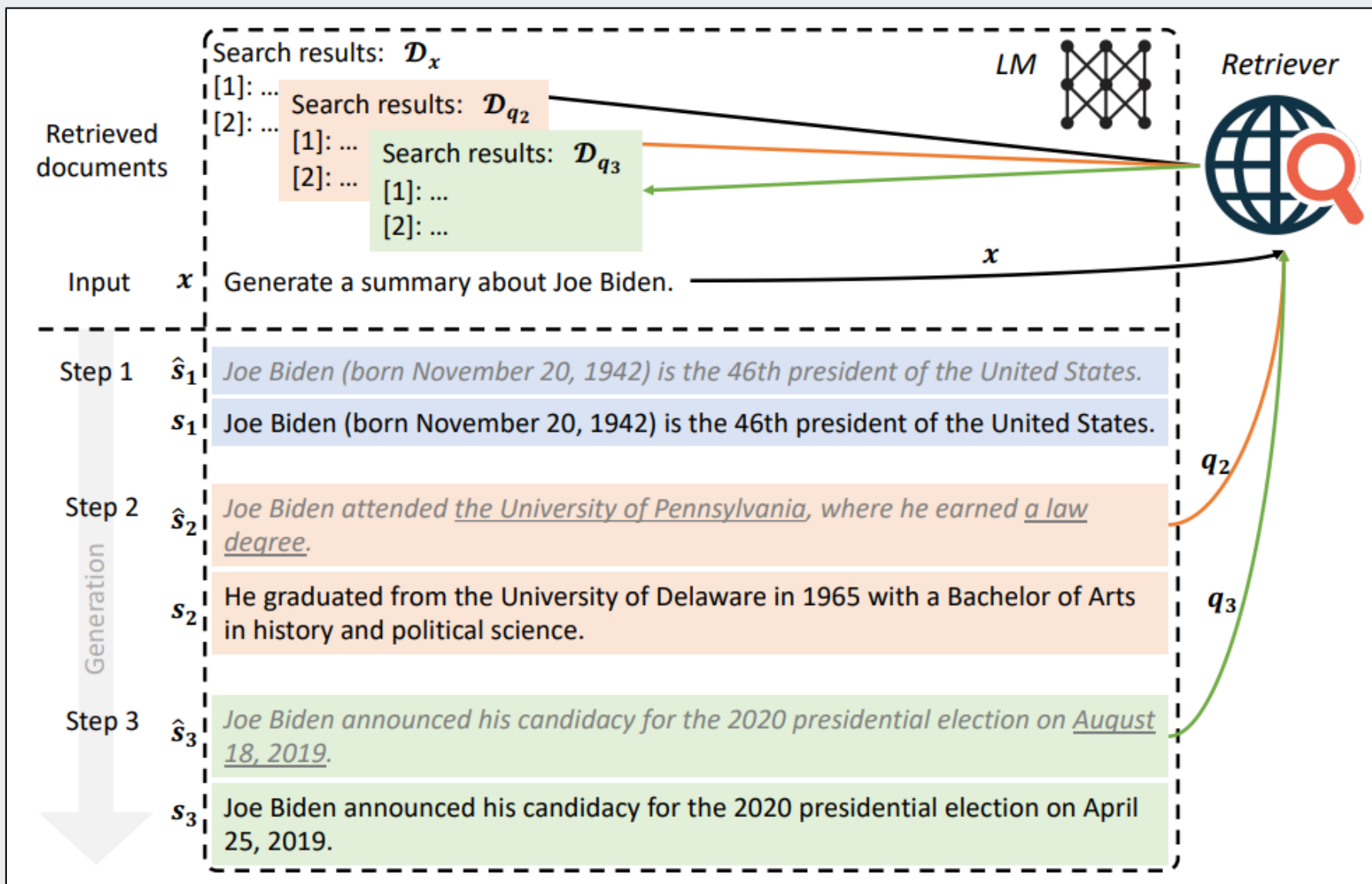


LM such as ChatGPT

What university did Joe Biden attend?  
What degree did Joe Biden earn?

# METHOD

## *FLARE*<sub>direct</sub>





# EXPERIMENTS

## Setting

- Knowledge base: Wikipedia dump
- Retriever: BM25, Search Engine (Bing)
- $FLARE_{direct}$

## Dataset

Dataset	$\theta$	$\beta$	Query formulation
2WikiMultihopQA	0.8	0.4	implicit
StrategyQA	0.4	0.4	implicit
ASQA & ASQA-hint	0.8	0.4	explicit
WikiAsp	0.8	0.4	explicit



# EXPERIMENTS

## Dataset

- 2WikiMultihopQA: 2-hop complex questions

Question: When did the director of film Hypocrite (Film) die?

Answer: The film Hypocrite was directed by Miguel Morayta. Miguel Morayta died on 19 June 2013. So the answer is 19 June 2013.

- StarategyQA: commonsense reasoning (yes/no)

Generate a yes or no answer to the following question.

Question: Do hamsters provide food for any animals?

Answer: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals. So the final answer is yes.

Dataset	$\theta$	$\beta$	Query formulation
2WikiMultihopQA	0.8	0.4	implicit
StrategyQA	0.4	0.4	implicit
ASQA & ASQA-hint	0.8	0.4	explicit
WikiAsp	0.8	0.4	explicit

# EXPERIMENTS

## Dataset

- ASQA & ASQA-hint: comprehensive answers

Dataset	$\theta$	$\beta$	Query formulation
2WikiMultihopQA	0.8	0.4	implicit
StrategyQA	0.4	0.4	implicit
ASQA & ASQA-hint	0.8	0.4	explicit
WikiAsp	0.8	0.4	explicit

Given an ambiguous question, figure out its interpretations and answer them one by one.

Question: Who played bonnie in gone with the wind?

Answer: This question is ambiguous in terms of which version or adaptation of Gone with the Wind is being referred to. In order to figure out its interpretations, we need to consider different versions or adaptations of Gone with the Wind. Gone with the Wind has two versions or adaptations: the 1939 film Gone with the Wind or the 2008 musical Gone with the Wind. Therefore, this question has 2 interpretations: (1) Who played Bonnie in the 1939 film Gone with the Wind? (2) Who played Bonnie in the 2008 musical Gone with the Wind? The answers to all interpretations are: (1) The 1939 film Gone with the Wind's character Bonnie was played by Eleanore Cammack "Cammie" King. (2) The 2008 musical Gone with the Wind's character Bonnie was played by Leilah de Meza.

Given an ambiguous question and a hint on which aspect of the question is ambiguous, figure out its interpretations and answer them one by one.

Question: Who played bonnie in gone with the wind?

Hint: This question is ambiguous in terms of which version or adaptation of Gone with the Wind is being referred to.

Answer: In order to figure out its interpretations, we need to consider different versions or adaptations of Gone with the Wind. Gone with the Wind has two versions or adaptations: the 1939 film Gone with the Wind or the 2008 musical Gone with the Wind. Therefore, this question has 2 interpretations: (1) Who played Bonnie in the 1939 film Gone with the Wind? (2) Who played Bonnie in the 2008 musical Gone with the Wind? The answers to all interpretations are: (1) The 1939 film Gone with the Wind's character Bonnie was played by Eleanore Cammack "Cammie" King. (2) The 2008 musical Gone with the Wind's character Bonnie was played by Leilah de Meza.

# EXPERIMENTS

## Dataset

- WikiAsp: open-domain summarization

Dataset	$\theta$	$\beta$	Query formulation
2WikiMultihopQA	0.8	0.4	implicit
StrategyQA	0.4	0.4	implicit
ASQA & ASQA-hint	0.8	0.4	explicit
WikiAsp	0.8	0.4	explicit

Generate a summary about Lakewood (Livingston, Alabama) including the following aspects: architecture, history with one aspect per line.

# Architecture

The house has a plan that is relatively rare in early Alabama architecture. The plan features a brick ground floor that is topped by one-and-a-half-stories of wood-frame construction. The ground floor originally contained domestic spaces, with the formal rooms on the principle floor and bedrooms on the upper floor. A central hallway is present on all levels. The facade is five bays wide, with central entrance doors on the ground and principle floors. The bays are divided by two-story Doric pilasters, with the middle third of the facade occupied by a two-tiered tetrastyle Doric portico. Two curved wrought iron staircases ascend from ground level to the front center of the upper portico, leading to the formal entrance.

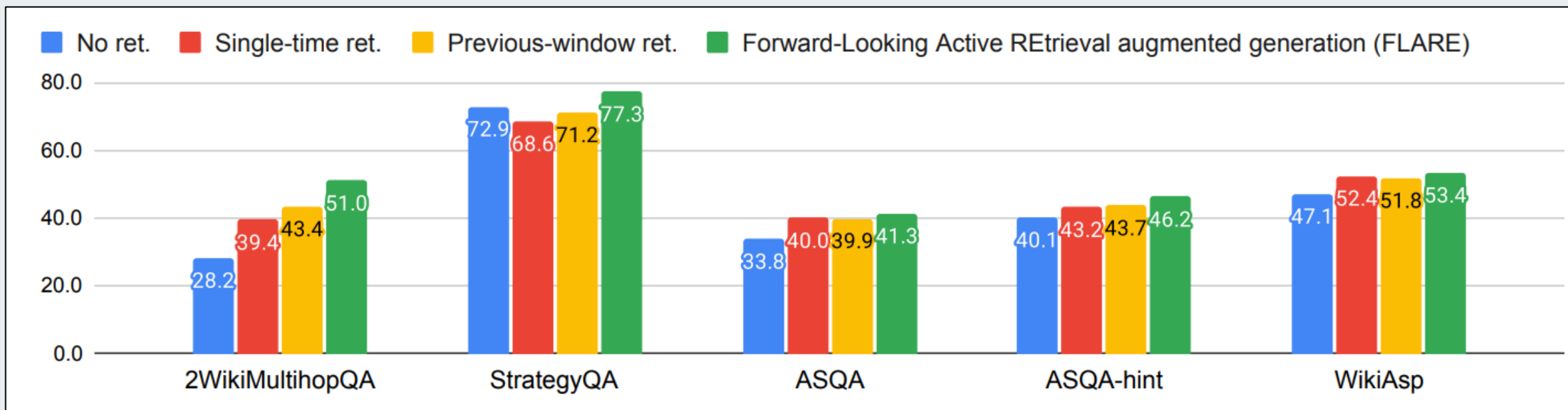
# History

Lakewood was built for Joseph lake, a native of North Carolina, by Hiram W. Bardwell, a master builder. Construction was completed in 1840. Located adjacent to the University of West Alabama, Julia Strudwick Tutwiler, a Lake relative, periodically resided in the house from 1881 to 1910 while she served as president of the university. It was then known as Livingston Normal College. The house was extensively photographed by Alex Bush for the Historic American Buildings Survey in November and December 1936. Lakewood has continued to be owned by descendants of the Lake family to the current day. The house and its surviving 10 acres (4.0 ha) of grounds were listed on the Places in Peril in 2012 due to the immediate threat of its acquisition by developers.

# EXPERIMENTS

## Main results

- FLARE outperforms all baselines
- Largest performance improvement: 2WikiMultihopQA > clear goal
- ASQA & WikiASP: comprehensive answer
- ASQA-hint: the hint helps LMs stay on topic



# EXPERIMENTS

## 2WikiMultihopQA

- In baselines, largest performance improvement: Question decomposition
  - Guide LMs to generate task-specific sub-question through task-specific annotation
  - FLARE surpasses this approach, demonstrating manual example is unnecessary

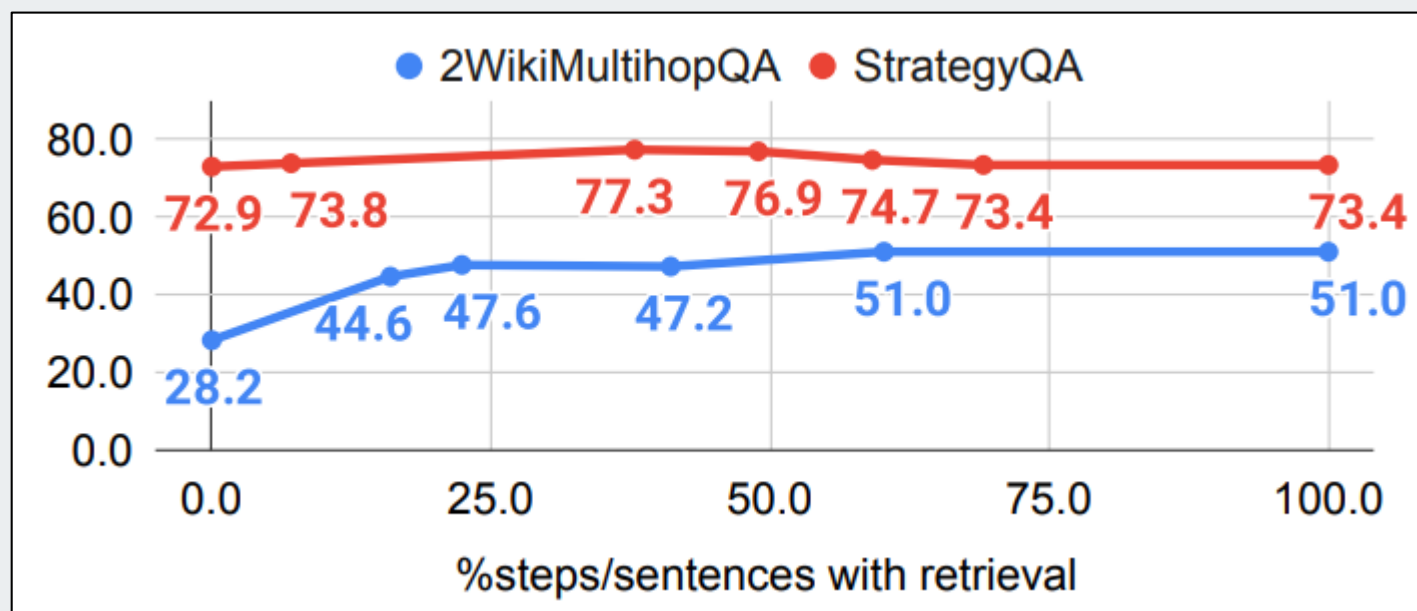
Methods	EM	F <sub>1</sub>	Prec.	Rec.
No retrieval	28.2	36.8	36.5	38.6
Single-time retrieval	39.4	48.8	48.6	51.5
<i>Multi-time retrieval</i>				
Previous-window	43.2	52.3	51.7	54.5
Previous-sentence	39.0	49.2	48.9	51.8
Question decomposition	47.8	56.4	56.1	58.6
FLARE <sub>instruct</sub> (ours)	42.4	49.8	49.1	52.5
FLARE <sub>direct</sub> (ours)	<b>51.0</b>	<b>59.7</b>	<b>59.1</b>	<b>62.6</b>



# EXPERIMENTS

## Ablation study

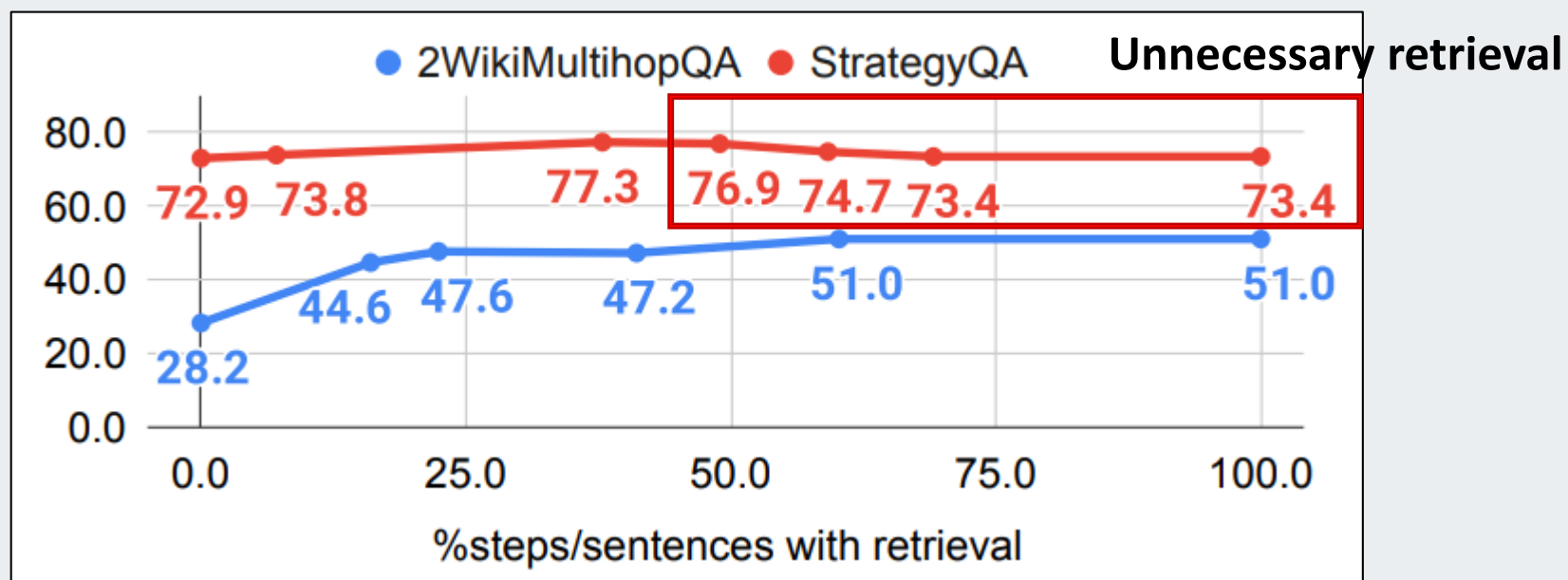
- Confidence of threshold
  - $\theta = 0$ : no search
  - $\theta = 1$ : retrieve every time



# EXPERIMENTS

## Ablation study

- Confidence of threshold
  - $\theta = 0$ : no search
  - $\theta = 1$ : retrieve every time



# EXPERIMENTS

## Ablation study

- Confidence of beta
  - $\beta = 0$ : no masking
  - Error tokens with low confidence interfere with retrieval

$\beta$	EM	F <sub>1</sub>	Prec.	Rec.
0.0	0.488	0.576	0.571	0.605
0.2	0.498	0.588	0.582	0.616
0.4	0.510	0.597	0.591	0.627
0.6	0.506	0.593	0.586	0.622



# EXPERIMENTS

## Ablation study

- Comparison between implicit and explicit query formulation

Dataset	$\theta$	$\beta$	Query formulation
2WikiMultihopQA	0.8	0.4	implicit
StrategyQA	0.4	0.4	implicit
ASQA & ASQA-hint	0.8	0.4	explicit
WikiAsp	0.8	0.4	explicit

	ASQA-hint				WikiAsp		
	EM	D-F <sub>1</sub>	R-L	DR	UniEval	E-F <sub>1</sub>	R-L
Implicit	45.7	36.9	37.7	37.3	53.4	18.8	27.7
Explicit	46.2	36.7	37.7	37.2	53.4	18.9	27.6

# | LIMITATIONS

## Wizard of Wikipedia & ELI5

- QA tasks that generate overly simple answers or require in-depth responses did not provide significant performance improvements

### <Wizard of Wikipedia>

**Topic:** Shower (*unseen*)

Human: So I've been looking for a better showerhead with more water pressure for my shower.

Model: I love pressure-versed water at nearly constant temperature to a shower

Human: That sounds pretty great! Love those rainshower showers!

Model: the ancient greeks were the first people to have showers.

Human: Well I knew they were good for something.

Model: the original showers were neither indoor structures nor man made but were common.

Human: Ah that sounds intimidating!

Model: it is required manual transport of both fresh and waste water.