ReSCORE: Label-free Iterative Retriever Training for Multi-hop Question Answering with Relevance-Consistency Supervision

Dosung Lee, Wonjun Oh, Boyoung Kim, Minyoung Kim,
Joonsuk Park, Paul Hongsuck Seo

Korea University, NAVER AI Lab, NAVER Cloud, University of Richmond

{dslee1219, owj0421, bykimby, omniverse186, phseo}@korea.ac.kr, park@joonsuk.org

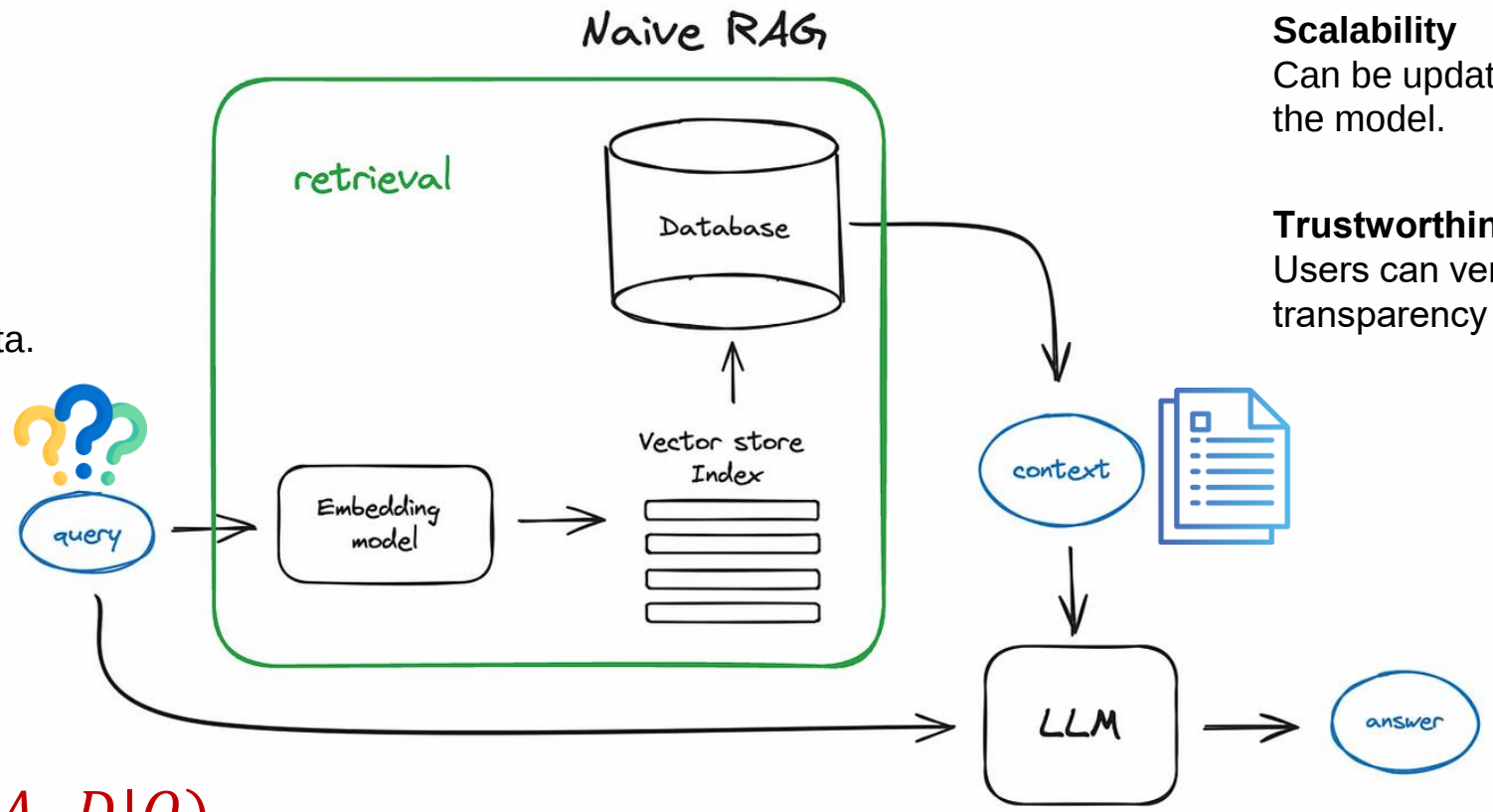
KOREA UNIVERSITY

MIIL Multimodal Interactive Intelligence Laboratory

Retrieval Augmented Generation

Improved Accuracy
Reduces hallucinations and generates factual responses.

Enhanced Knowledge Base
Up-to-date and expansive sources beyond pre-trained data.



Scalability
Can be updated without retraining the model.

Trustworthiness
Users can verify sources, increasing transparency and reliability.

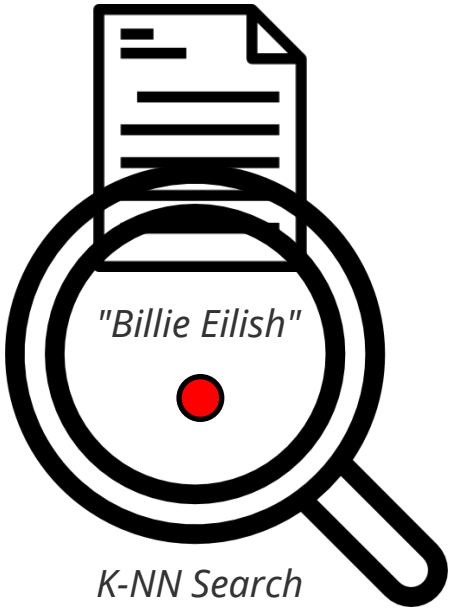
$$P_{QA}(A|Q) = \sum_{D} P(A, D|Q)$$

$$P_{RAG}(A, D|Q) = P_{generate}(A|D, Q)\, P_{retrieve}(D|Q)$$

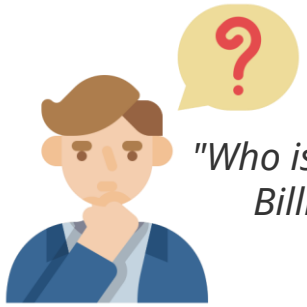
Multi-hop Question Answering & Challenges

Single-hop

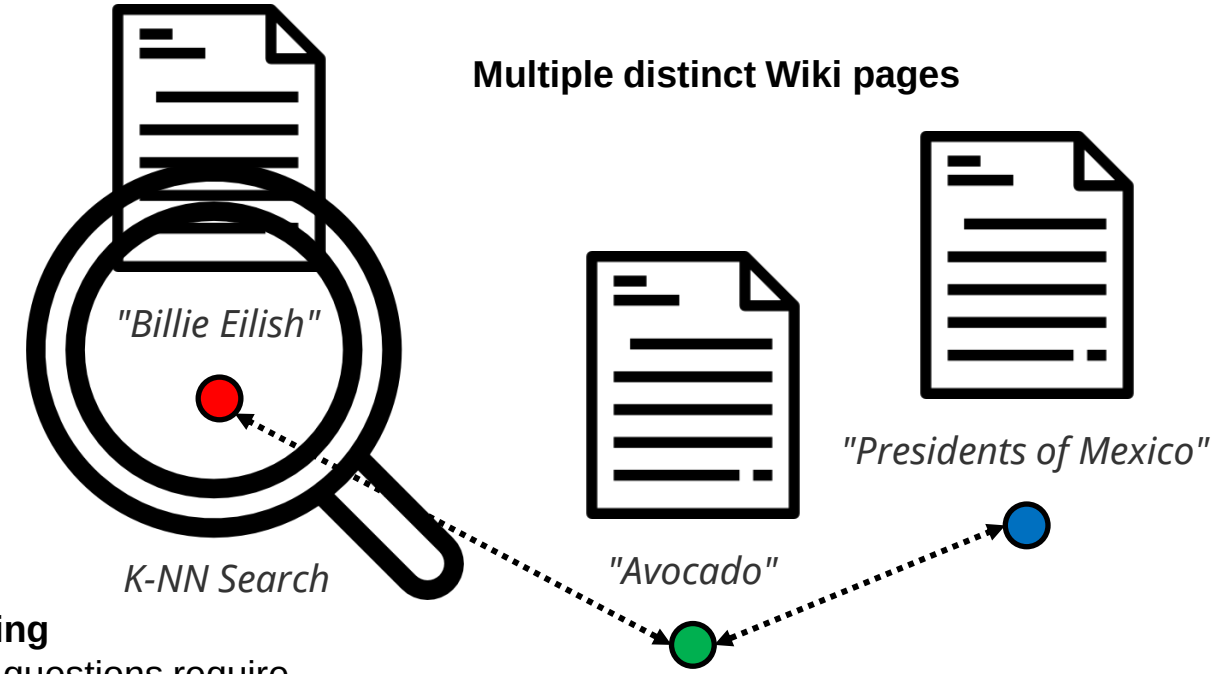
“What is Billie Eilish’s favorite food?”



Multi-hop

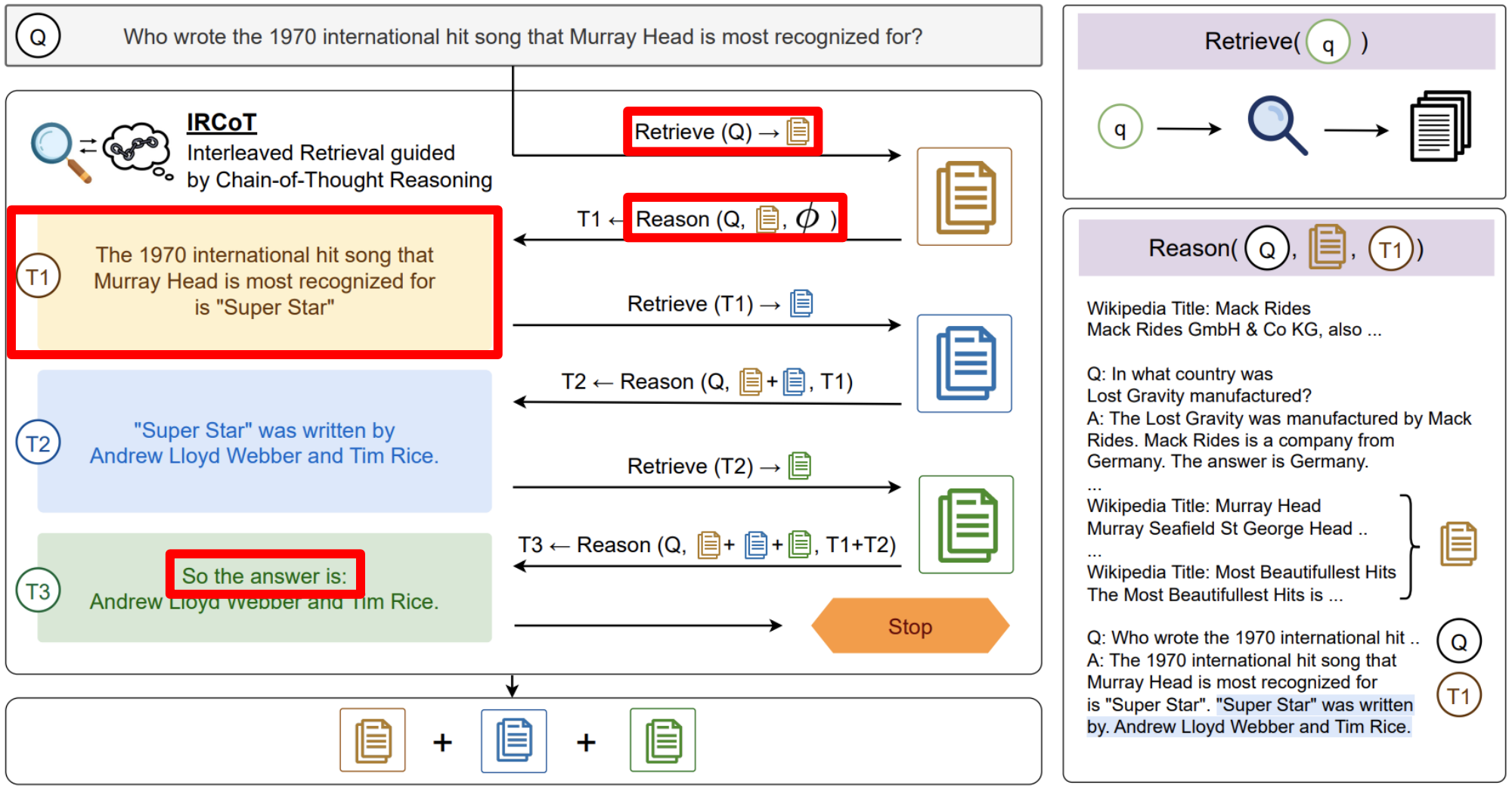


"Who is the first president of the country where did Billie Eilish’s favorite food originates from?"

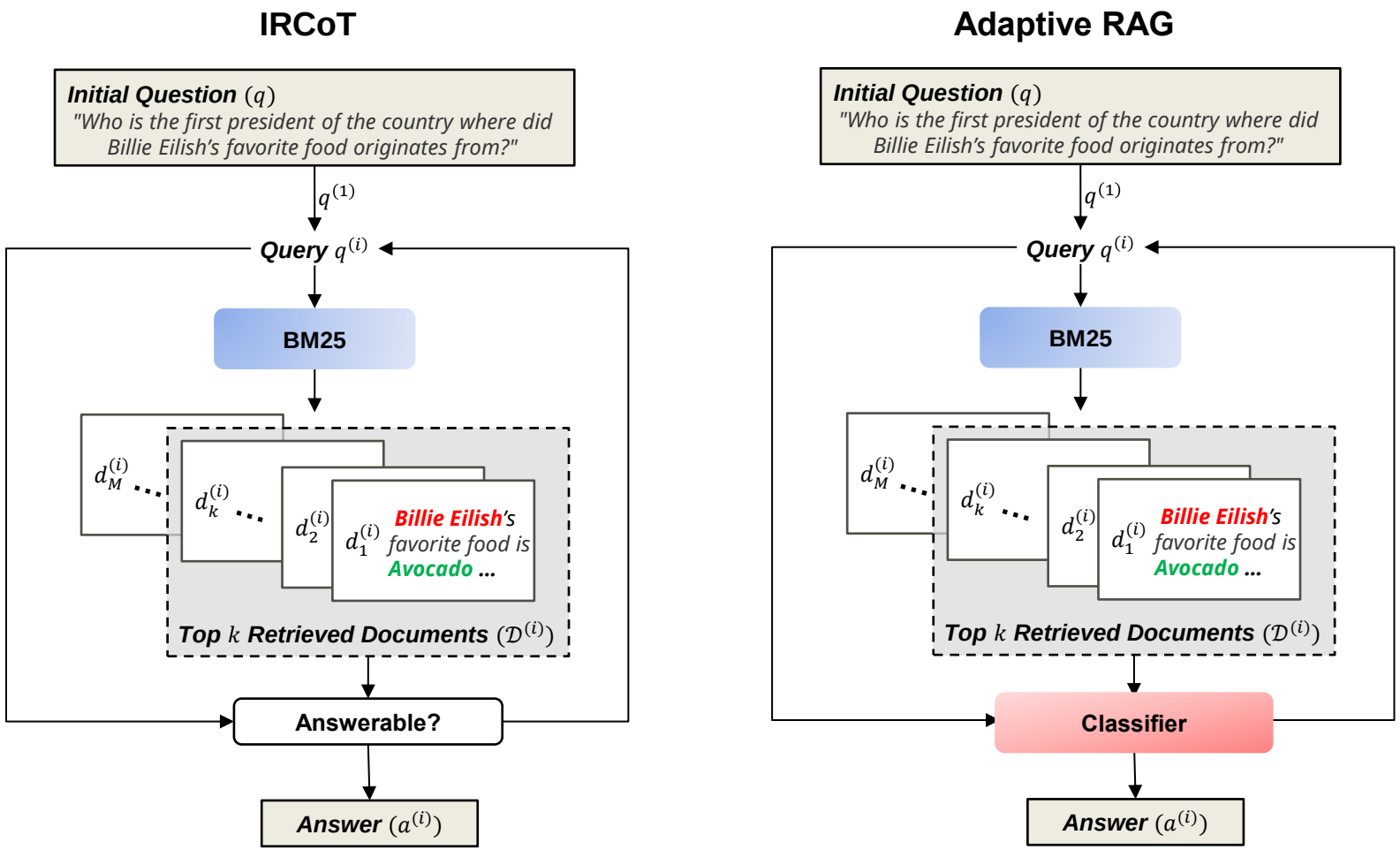


Complex Reasoning
in real world many questions require
connecting multiple facts in multiple documents

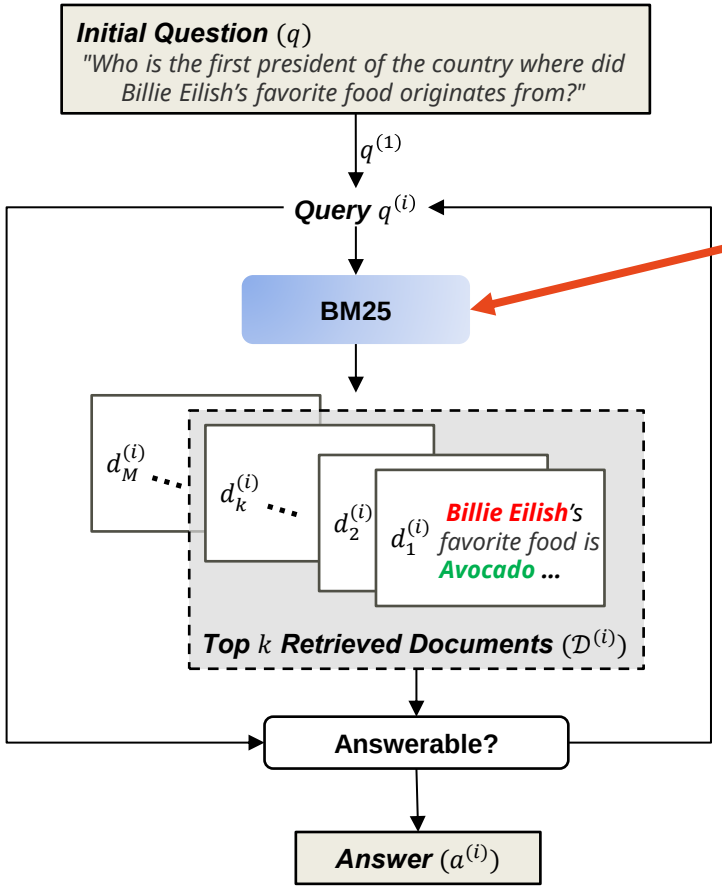
Iterative RAG Baselines for MHQA



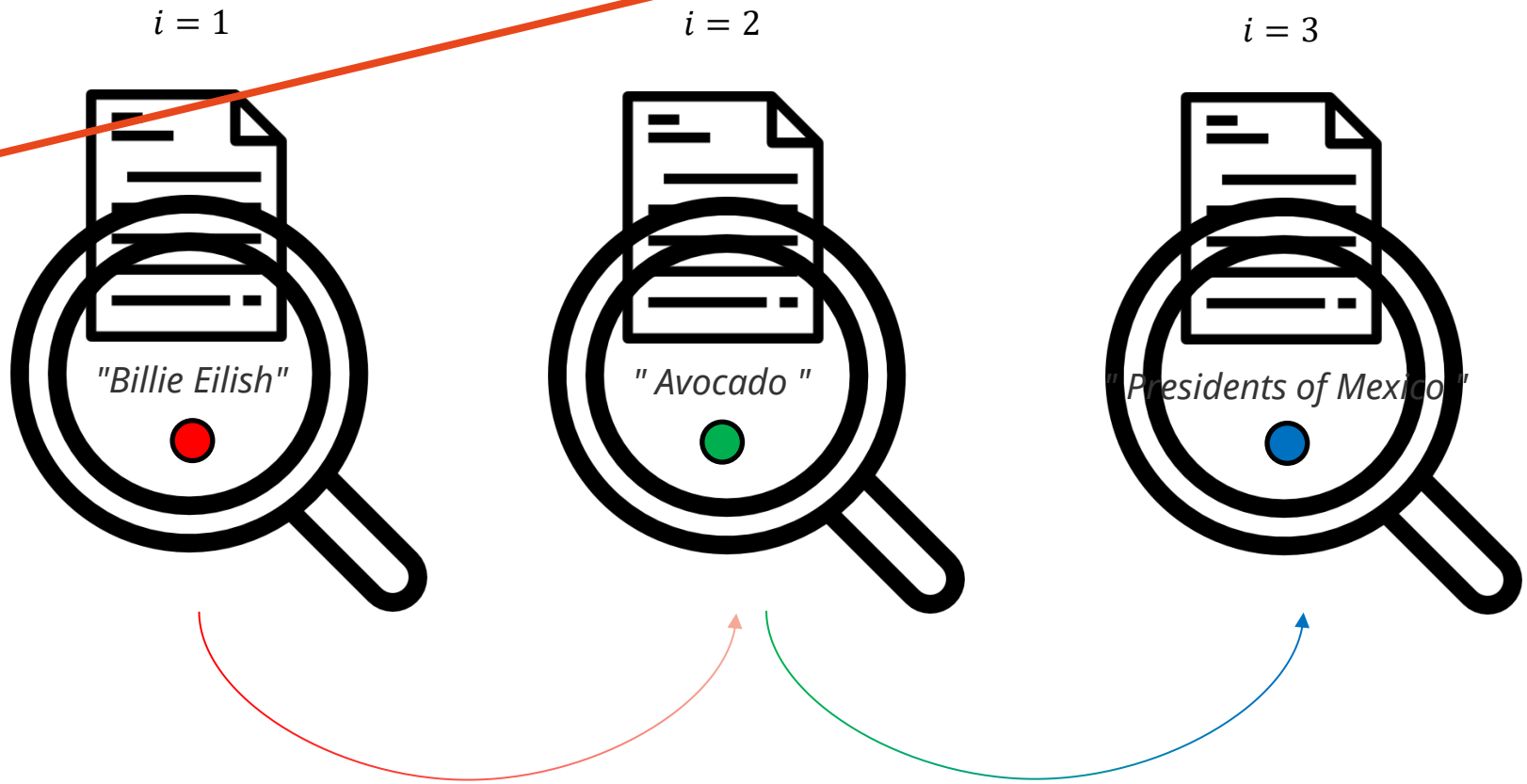
Iterative RAG Baselines for MHQA



Limitation of Iterative RAG Baselines for MHQA



Rely on word level exact matching, **without training** a dense retriever



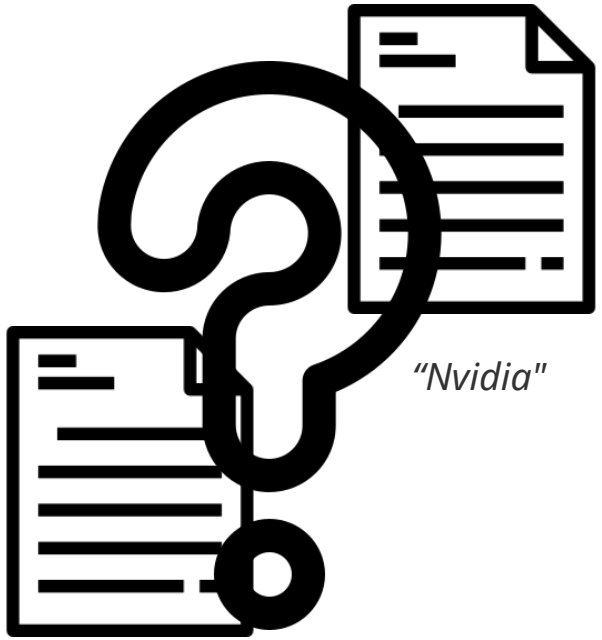
*"**Billie Eilish**'s favorite food is **Avocado**"* *"**Avocado** originated from **Mexico**"*

Challenges of Labeling MHQA & Label-free Training

"Who is the first president of the country where did Billie Eilish's favorite food originates from?"

"Billie Eilish"

"Elon Musk"



"Nvidia"

"South Korea"

" Presidents of Mexico "

" Avocado"

"Paris"

Which document is related?

Relevance-Consistency Supervision

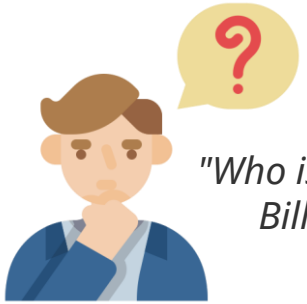
If human annotations are cost- and labor-intensive can we use LLM?

$$P_{QA}(A|Q) = \sum_{D} P(A, D|Q)$$

$$P_{RAG}(A, D|Q) = P_{generate}(A|D, Q)\ P_{retrieve}(D|Q)$$

$$P_{Retriever}(D|Q) \propto P_{LLM}(A, Q|D)$$

$$P_{LLM}(A, Q|D) = P_{Consistency}(A|D, Q)\ P_{Relevance}(Q|D)$$



"Who is the first president of the country where did Billie Eilish's favorite food originates from?"

"Presidents of Mexico"

$P_{Consistency}(A|D, Q)$

"Billie Eilish"

$P_{Relevance}(Q|D)$

How is Relevance-Consistency Label calculated exactly?

The "Predictive Confidence" of AI

🌦️ **80% chance of rain**, it's quite confident.

🌞 **10% chance of rain**, it's unlikely.

Similarly, **"Who is the CEO of Apple?“**

✅ **"Tim Cook"** (90%)

❌ **"Steve Jobs"** (10%)

This probability comes from the **logits (log probabilities)** that the model generates.

How is Relevance-Consistency Label calculated exactly?

Checking Probability of the Exact GT Question & Answer

Steve Jobs co-founded Apple in 1976. **Who co-founded Apple? Steve Jobs.**

Steve Jobs co-founded Apple in 1976.	**Who**	**co-founded**	**Apple?**	**Steve**	**Jobs.**
Steve Jobs co-founded Apple in 1976.	-0.2	-0.3	-0.1	-0.15	-0.1

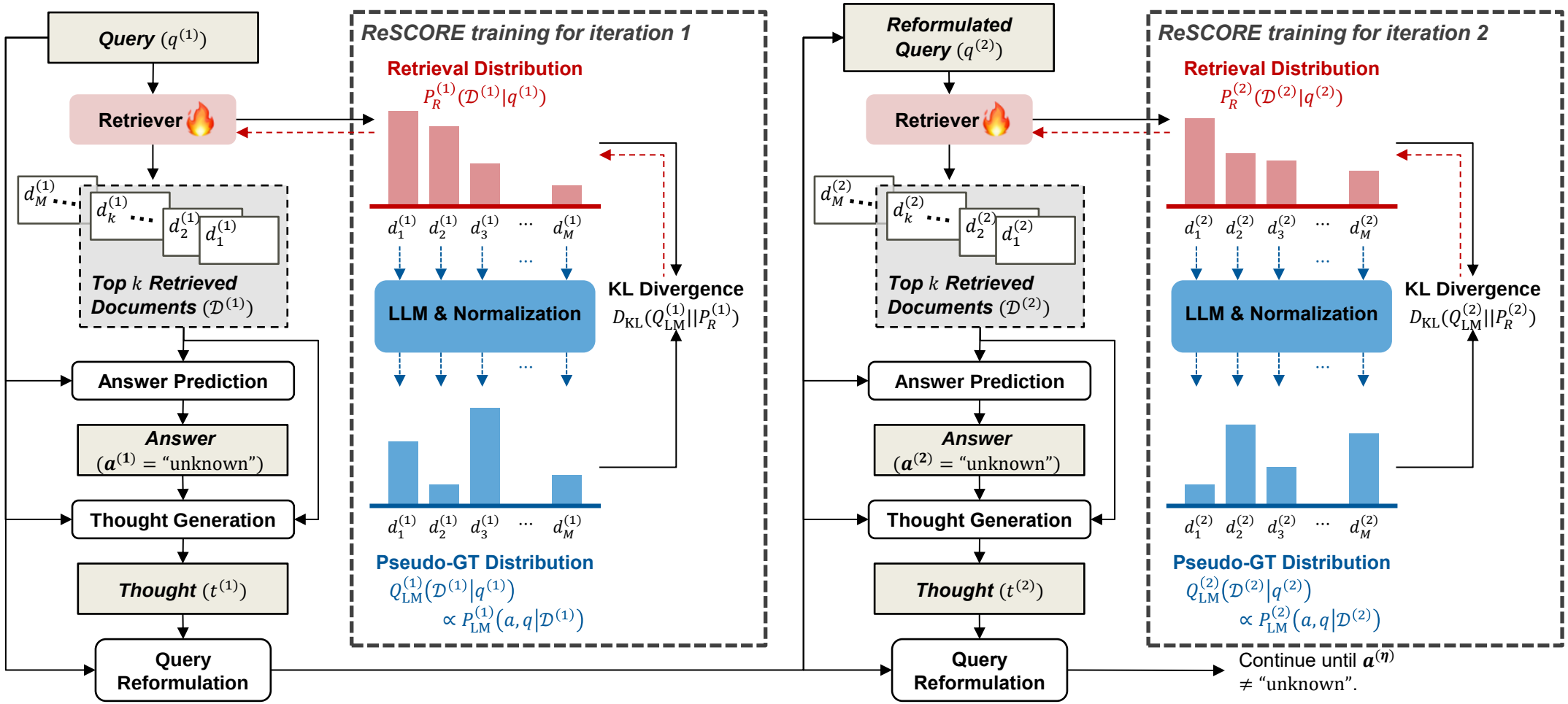
Ignored ?, -, . for convenience

Sum of the log probs: -0.2 + -0.3 + -0.1 + -0.15 + -0.1 = -0.85

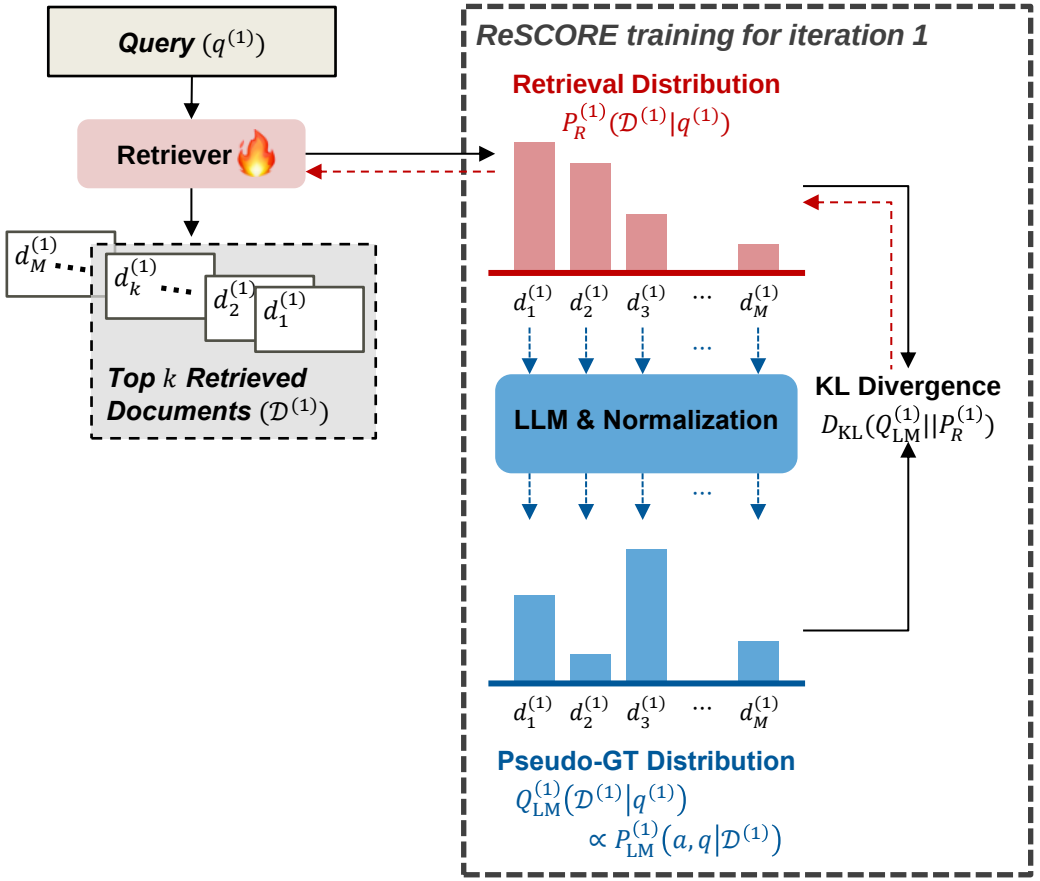
1. Make a **prompt** asking to **generate a question and answer given context.**
2. Give the model the **context prompt + GT question + GT answer.**
3. Extract the probability (log probability) of predicting the tokens at the **GT question + GT answer** positions (which is fixed.)

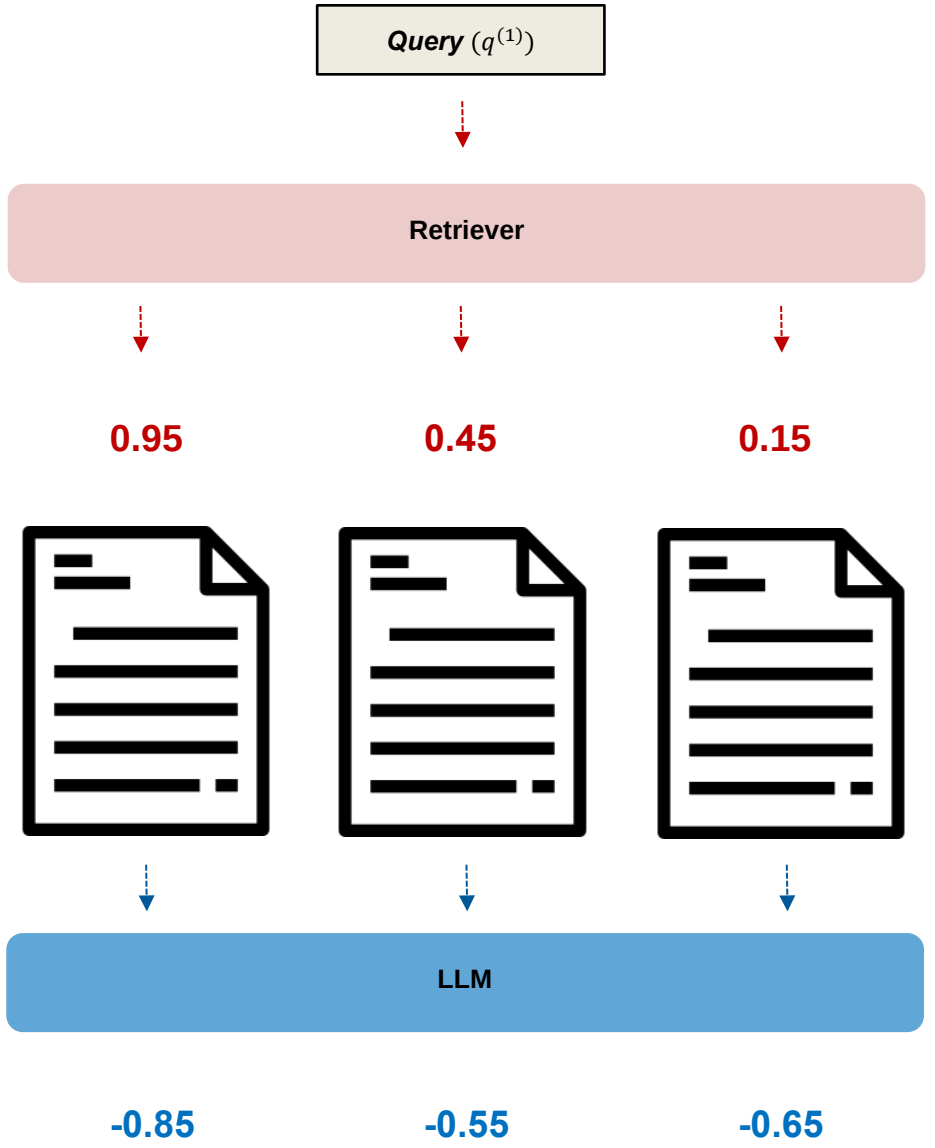
The higher the log probability, the more confident the model is in generating the exact Q&A.

ReSCORE Training

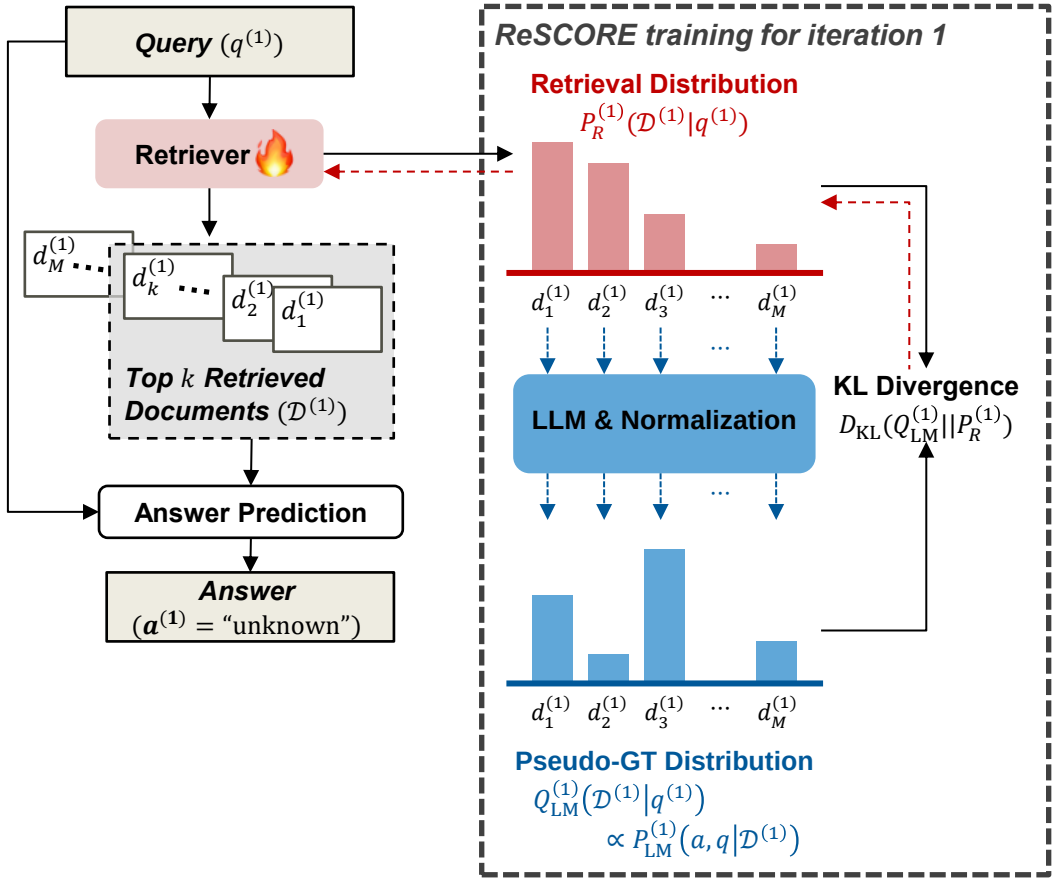


ReSCORE Training Retrieval Step





ReSCORE Training Answer Generation Step



Like our baseline

T3 So the answer is: Andrew Lloyd Webber and Tim Rice. → Stop

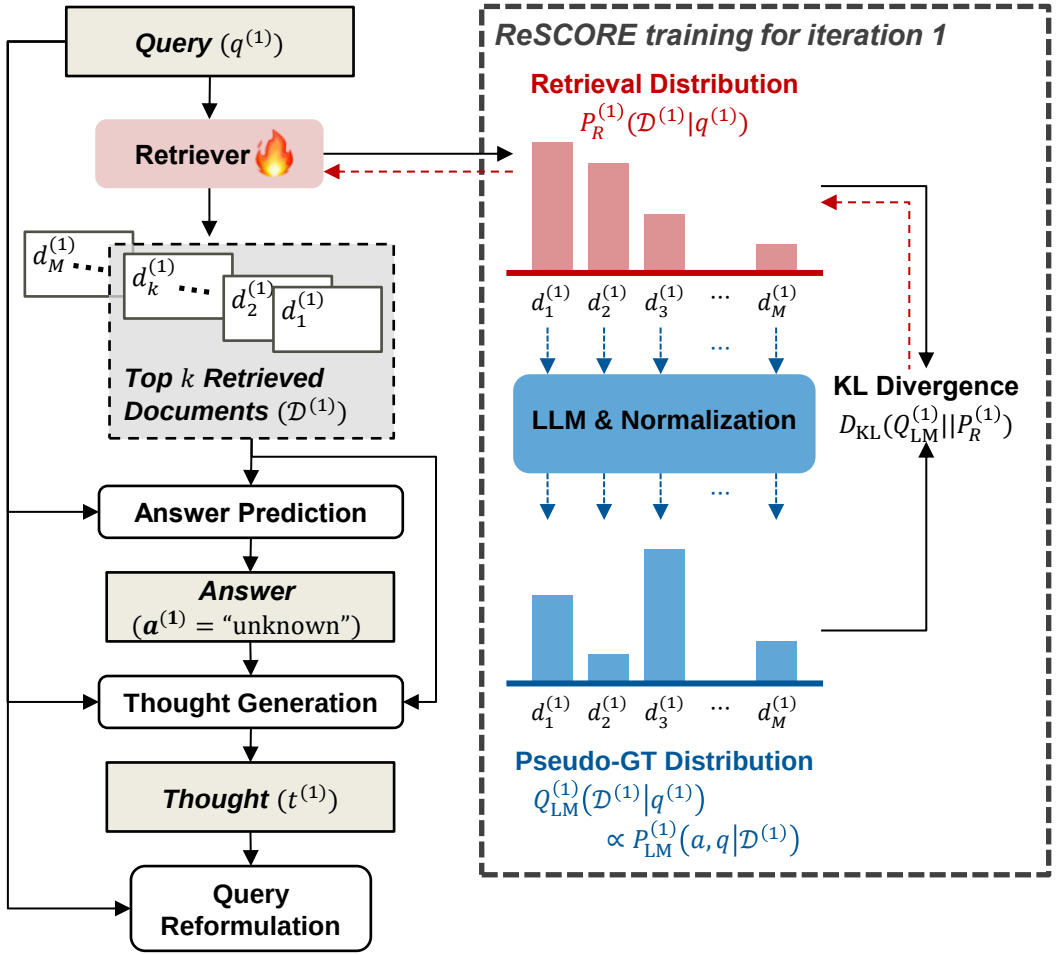
Rely on LLM to stop

```
Instructions:
 - Carefully read the documents and hints.
 - If you know the answer to the question confidently, generate an answer,
 using documents and hints provided.
 - If you don't know, generate "Unknown".
```

(a) If $a^{(i)}$ is "unknown", continue iterations.

(b) If $a^{(i)}$ is **not** "unknown", return the answer.

ReSCORE Training Query Reformulation Step

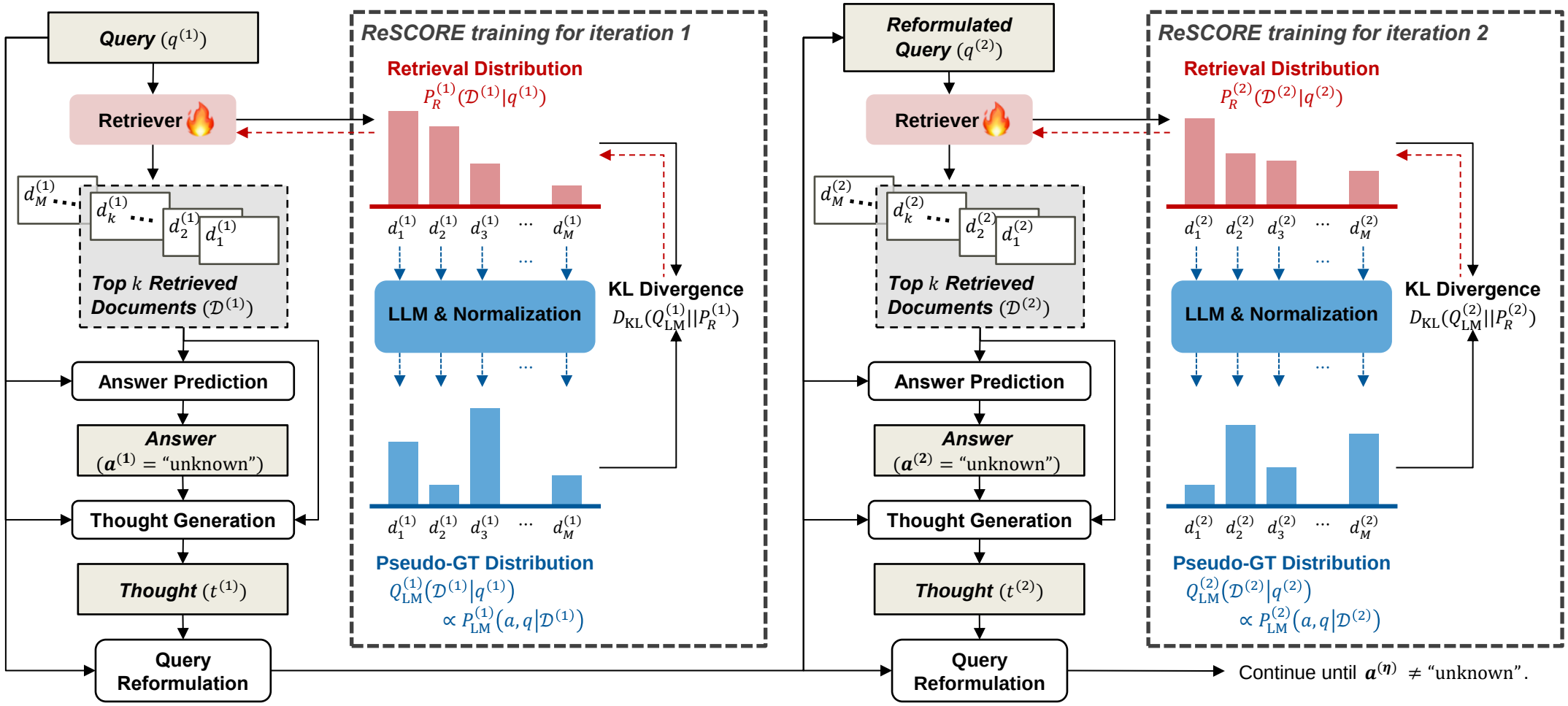


"Who is the first president of the country where did Billie Eilish's favorite food originates from?"

Query Reformulation

None	*" Who is the first president of the country where did Billie Eilish's favorite food originates from?"*
	OR
LLM-rewrite	*"Who is the first president of the country where did **Avocado** originates from?"*
	OR
Thought-concat	*"Who is the first president of the country where did Billie Eilish's favorite food originates from? **Billie Eilish**'s favorite food is **Avocado.**"*

ReSCORE Training



Comparison to SOTA MHQA Baselines

Model	MuSiQue EM	MuSiQue F1	HotpotQA EM	HotpotQA F1	2WikiMHQA EM	2WikiMHQA F1
ReAcT (GPT-3.5+BM25)†	10.2	19.7	36.0	46.9	28.0	37.3
FLARE (GPT-3.5+BM25)†	11.2	18.7	36.4	47.8	31.8	42.8
Self-RAG (GPT-3.5+BM25)†	10.6	19.2	33.8	44.4	24.4	30.8
Adaptive-Note (GPT-3.5+BM25)†	13.2	24.2	45.6	58.4	43.2	54.2
IRCoT (Flan-T5-XL+BM25)‡	22.0	31.8	44.42	56.2	49.7	54.9
Adaptive-RAG (Flan-T5-XL+BM25)‡‡	**23.6**	31.8	42.0	53.8	40.6	49.8
Our Baseline (Llama-3.1-8B+BM25)	15.2	23.6	42.2	55.7	44.6	52.2
Our Baseline (Llama-3.1-8B+Contriever)	15.2	23.8	39.4	52.3	32.8	41.6
IQATR (Llama-3.1-8B+Contriever trained w/ ReSCORE)	23.4	**32.7**	**47.2**	**59.3**	**50.0**	**59.7**

	MuSiQue			HotpotQA			2WikiMHQA		
	cEM	EM	F1	cEM	EM	F1	cEM	EM	F1
Self-Ask	-	13.8	27.0	-	-	-	-	30.0	36.1
Self-Ask + SE	-	15.2	27.2	-	-	-	-	40.1	52.6
SearChain + ColBERT	17.1	-	-	56.9	-	-	46.3	-	-
IQATR (Ours)	30.4	23.4	32.7	59.6	47.2	59.3	57.0	50.0	59.7

Comparison to SOTA iterative baselines

ReSCORE on Iterative baselines other than IRCoT & Adaptive RAG

Model	QA EM	QA F1	MHR$_i$@8 $i=1$	MHR$_i$@8 $i=2$	MHR$_i$@8 $i=\eta_n$
MuSiQue					
Self-RAG*	1.2	8.2	25.8	25.8	25.8
+ReSCORE	2.8	10.8	24.9	31.6	31.6
FLARE	7.3	13.3	31.0	37.1	37.1
+ReSCORE	8.2	15.3	30.9	40.1	43.3
Adaptive-Note	9.6	17.7	44.9	50.2	50.2
+ReSCORE	11.2	20.5	45.1	49.8	55.3
Our Baseline	15.2	23.8	44.9	51.6	51.6
+ReSCORE	23.4	32.7	46.8	63.0	65.2
HotpotQA					
Self-RAG*	5.6	17.9	36.1	36.5	36.5
+ReSCORE	8.7	19.2	33.8	37.2	37.2
FLARE	27.5	38.9	37.2	48.4	48.4
+ReSCORE	31.4	42.5	39.2	48.5	51.7
Adaptive-Note	42.0	55.3	44.8	49.8	50.1
+ReSCORE	43.8	58.0	47.3	63.3	77.2
Our Baseline	39.4	52.3	44.8	47.5	47.5
+ReSCORE	47.2	59.3	46.6	69.3	72.4
2WikiMHQA					
Self-RAG*	3.0	19.1	26.3	27.1	27.1
+ReSCORE	5.6	21.2	25.9	28.4	32.8
FLARE	23.2	35.0	32.5	42.9	42.9
+ReSCORE	26.5	38.0	33.2	45.6	45.6
Adaptive-Note	35.7	46.1	45.7	59.2	59.2
+ReSCORE	37.4	49.3	49.8	63.2	67.5
Our Baseline	32.8	41.6	45.7	56.9	56.9
+ReSCORE	50.0	59.7	51.2	81.2	88.0

Pseudo-GT Supervision Ablation

Method	MuSiQue			HotpotQA			2WikiMHQA		
	R@8	EM	F1	R@8	EM	F1	R@8	EM	F1
None	47.1	15.2	23.8	61.7	39.4	52.3	58.9	32.8	41.6
$P_{LM}(a\|d,q)$	41.4	5.8	12.3	42.8	19.2	26.4	41.9	18.8	26.5
$P_{LM}(q\|d)$	47.9	15.9	25.9	65.9	42.0	53.9	63.2	39.2	47.9
$P_{LM}(q,a\|d)$	55.7	16.4	26.3	68.3	43.6	56.4	67.1	41.4	51.7

$$P_{QA}(A|Q) = \sum_{D} P(A, D|Q)$$

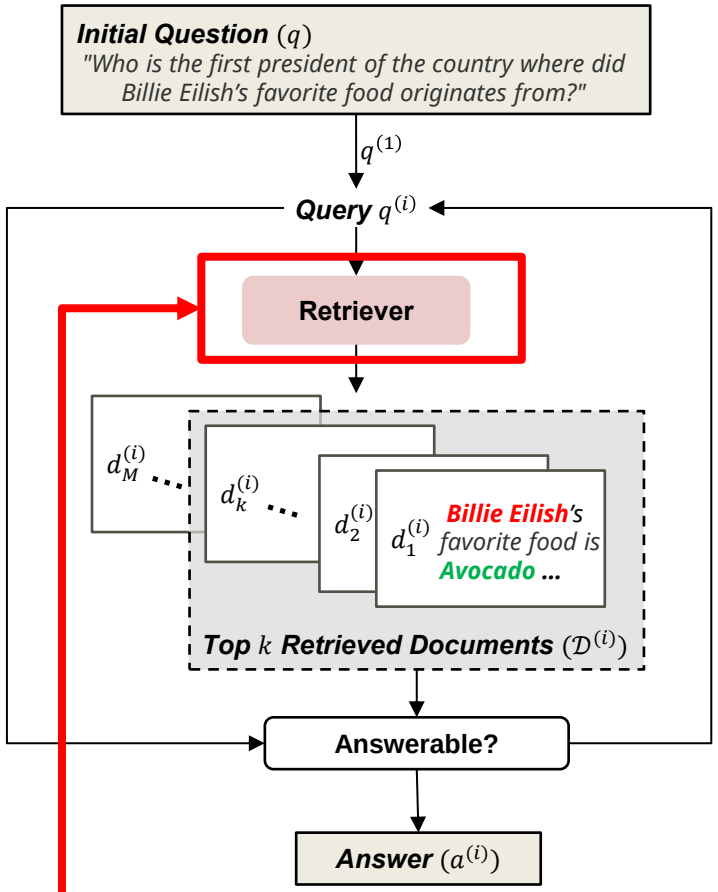
$$P_{RAG}(A, D|Q) = P_{generate}(A|D, Q)\, P_{retrieve}(D|Q)$$

$$P_{Retriever}(D|Q) \propto P_{LLM}(A, Q|D)$$

$$P_{LLM}(A, Q|D) = P_{Consistency}(A|D, Q)\, P_{Relevance}(Q|D)$$

Pseudo-GT Label	R@2	R@4	R@8	R@16
MuSiQue				
None	32.7	40.1	47.1	53.6
$P_{LM}(q \mid d)$	34.6	41.1	47.9	54.2
$P_{LM}(a \mid q, d)$	28.9	35.1	41.4	47.8
$P_{LM}(q, a \mid d)$	**42.7**	**50.3**	**55.7**	**60.4**
HotpotQA				
None	49.4	56.5	61.7	66.3
$P_{LM}(q \mid d)$	55.2	62.4	65.9	69.1
$P_{LM}(a \mid q, d)$	27.5	34.4	42.8	52.5
$P_{LM}(q, a \mid d)$	**58.1**	**64.6**	**68.3**	**70.7**
2WikiMHQA				
None	46.4	54.3	58.9	63.4
$P_{LM}(q \mid d)$	50.8	59.1	63.2	66.1
$P_{LM}(a \mid q, d)$	26.1	33.3	41.9	51.2
$P_{LM}(q, a \mid d)$	**53.7**	**63.0**	**67.1**	**68.7**

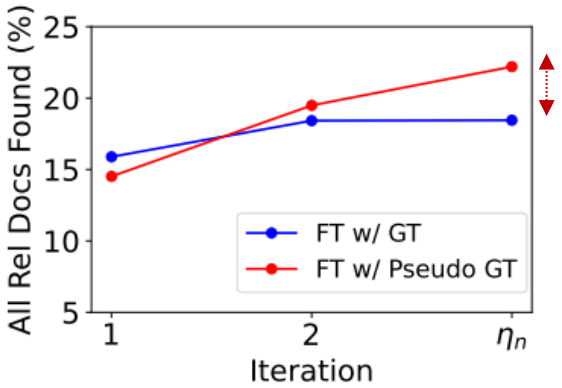
Comparison with GT



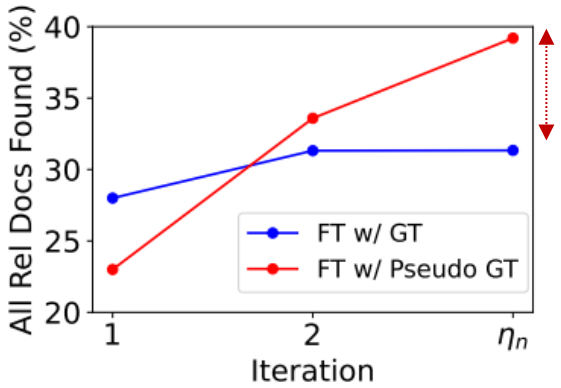
Replaced the Retriever with GT-trained

Label	QA EM	QA F1	MHR$_i$@8 $i=1$	MHR$_i$@8 $i=2$	MHR$_i$@8 $i=\eta_n$
MuSiQue					
None	15.2	23.8	44.9	51.6	51.6
GT	15.8	24.9	46.7	54.8	54.8
Pseudo-GT	23.4	32.7	46.8	63.0	65.2
HotpotQA					
None	39.4	52.3	44.8	47.5	47.5
GT	45.2	55.8	48.7	52.7	52.7
Pseudo-GT	47.2	59.3	46.6	69.3	72.4
2WikiMHQA					
None	32.8	41.6	45.7	56.9	56.9
GT	37.1	46.2	48.5	61.7	61.7
Pseudo-GT	50.0	59.7	51.2	81.2	88.0

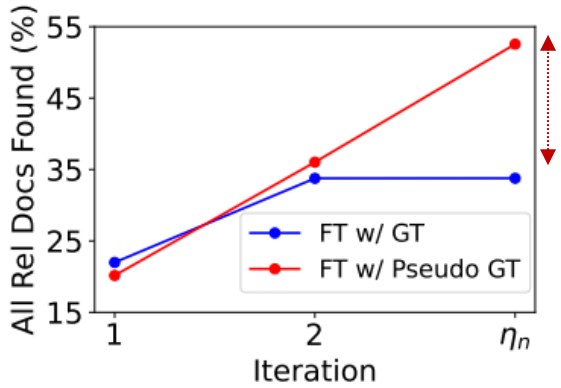
Comparison of GT and Pseudo-GT Labels



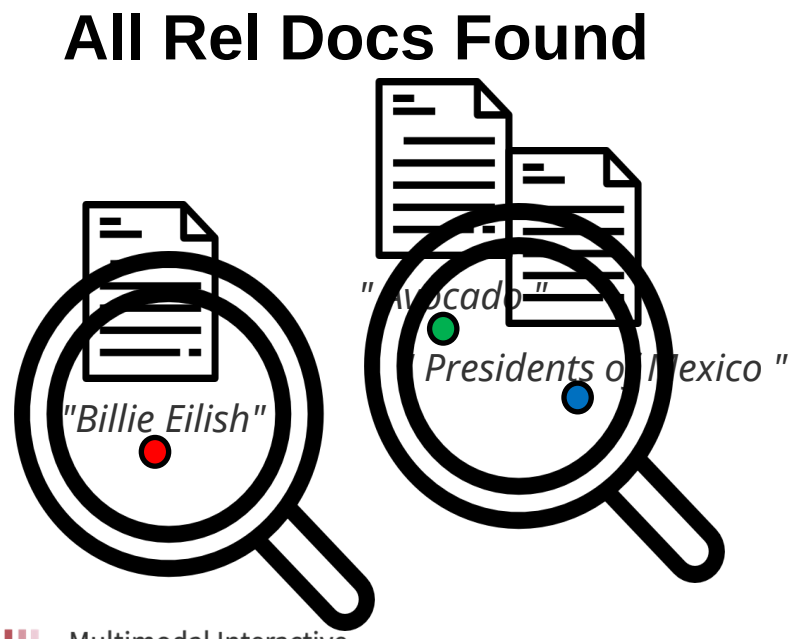
(a) MuSiQue Dataset

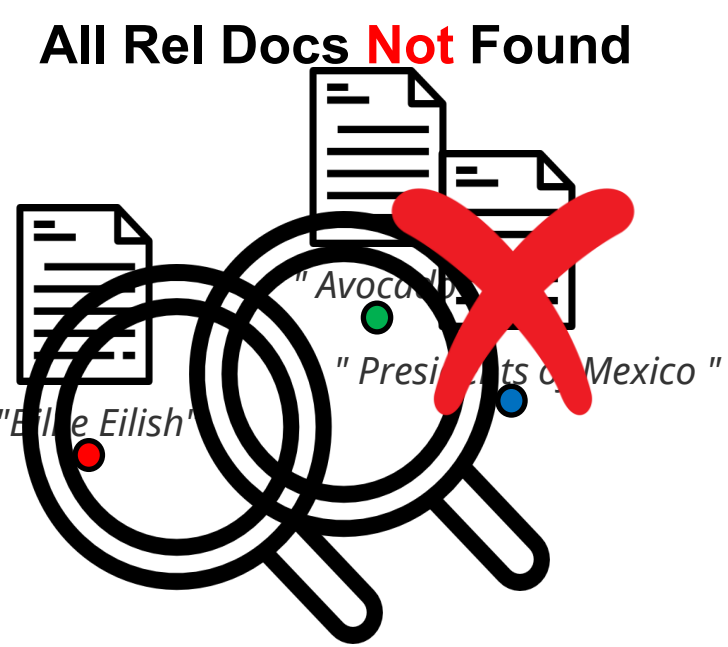


(b) HotpotQA Dataset

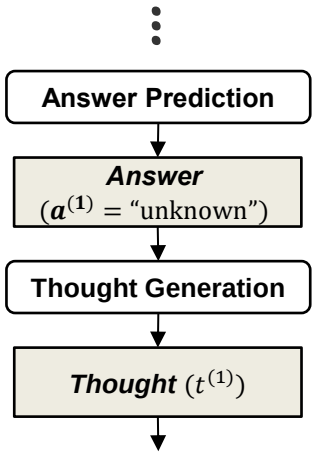


(c) 2WikiMHQA Dataset





Query Reformulation Ablation



"Who is the first president of the country where did Billie Eilish's favorite food originates from?"

*"**Billie Eilish**'s favorite food is **Avocado**"*

Query Reformulation

None

" Who is the first president of the country where did Billie Eilish's favorite food originates from?"

OR

LLM-rewrite

*"Who is the first president of the country where did **Avocado** originates from?"*

OR

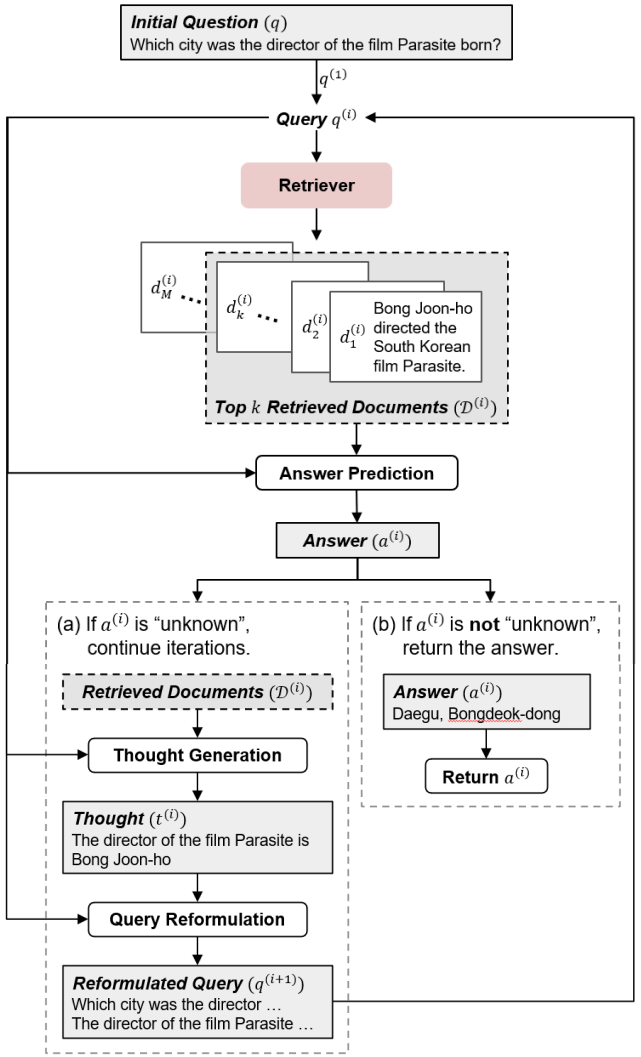
Thought-concat

*"Who is the first president of the country where did Billie Eilish's favorite food originates from? **Billie Eilish**'s favorite food is **Avocado.**"*

Reformulation Method	QA EM	QA F1	MHR$_i$@8 $i=1$	MHR$_i$@8 $i=2$	MHR$_i$@8 $i=\eta_n$
MuSiQue					
None	10.8	17.8	44.7	45.4	47.4
LLM-rewrite	21.2	30.5	45.1	56.7	63.7
Thought-concat	23.4	32.7	46.8	63.0	65.2
HotpotQA					
None	29.4	41.1	42.8	43.6	43.8
LLM-rewrite	44.2	57.4	41.9	54.8	64.7
Thought-concat	47.2	59.3	46.6	69.3	72.4
2WikiMHQA					
None	35.6	44.7	48.6	49.7	49.8
LLM-rewrite	51.7	60.1	50.0	86.0	89.5
Thought-concat	50.0	59.7	51.2	81.2	88.0

Summary: Label-free Iterative Retriever Training for Multi-hop QA

Research Objective



Task

- **Complex questions** that need to be answered by logically-connecting relevant information from multiple documents.

Prior Works

- **Rely on BM25**, as it is cost- and labor-intensive to prepare documents labeled with their relevance to respective queries across iterations.

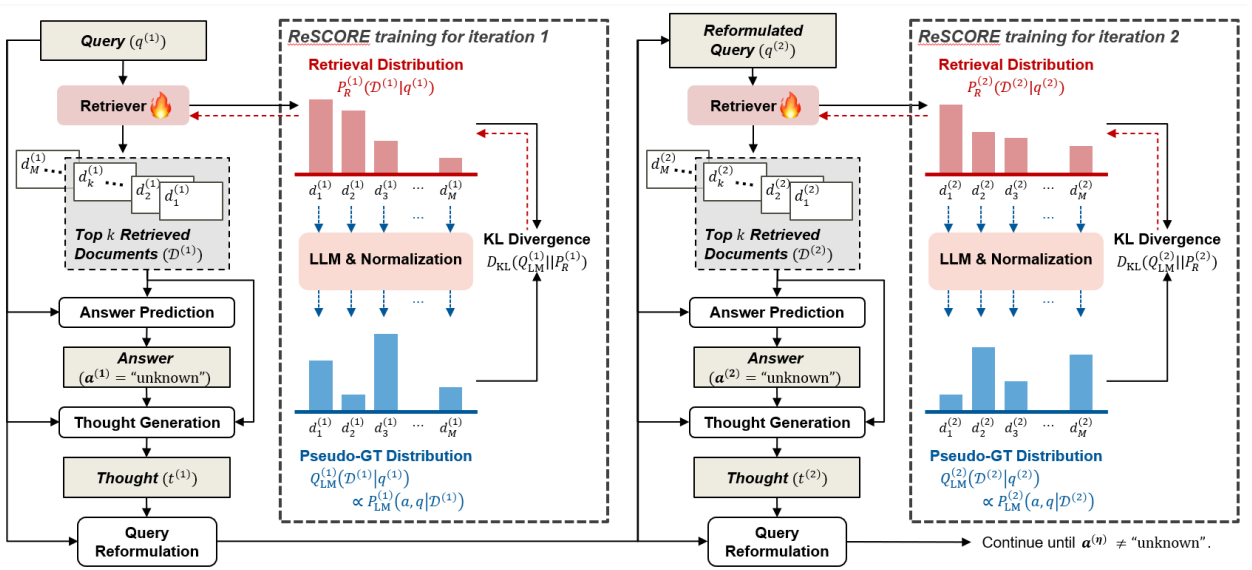
$$Q_{\mathrm{LM}}^{(i)}(d_j^{(i)} \mid q) \propto P_{\mathrm{LM}}^{(i)}(a, q \mid d_j^{(i)}) \quad (1)$$

$$= P_{\mathrm{LM}}^{(i)}(q \mid d_j^{(i)}) \cdot P_{\mathrm{LM}}^{(i)}(a \mid q, d_j^{(i)}) \quad (2)$$

Goal

- ReSCORE leverages **LLM** probability as **pseudo-ground truth label** to train the retriever

Our Method



Results

Model	MuSiQue EM	MuSiQue F1	HotpotQA EM	HotpotQA F1	2WikiMHQA EM	2WikiMHQA F1
ReAcT (GPT-3.5+BM25)†	10.2	19.7	36.0	46.9	28.0	37.3
FLARE (GPT-3.5+BM25)†	11.2	18.7	36.4	47.8	31.8	42.8
Self-RAG (GPT-3.5+BM25)†	10.6	19.2	33.8	44.4	24.4	30.8
Adaptive-Note (GPT-3.5+BM25)†	13.2	24.2	45.6	58.4	43.2	54.2
IRCoT (Flan-T5-XL+BM25)‡	22.0	31.8	44.42	56.2	49.7	54.9
Adaptive-RAG (Flan-T5-XL+BM25)‡‡	**23.6**	31.8	42.0	53.8	40.6	49.8
Our Baseline (Llama-3.1-8B+BM25)	15.2	23.6	42.2	55.7	44.6	52.2
Our Baseline (Llama-3.1-8B+Contriever)	15.2	23.8	39.4	52.3	32.8	41.6
IQATR (Llama-3.1-8B+Contriever trained w/ ReSCORE)	23.4	**32.7**	**47.2**	**59.3**	**50.0**	**59.7**