NAACL 2025

Diversify-verify-adapt: Efficient and Robust Retrieval-Augmented Ambiguous Question Answering

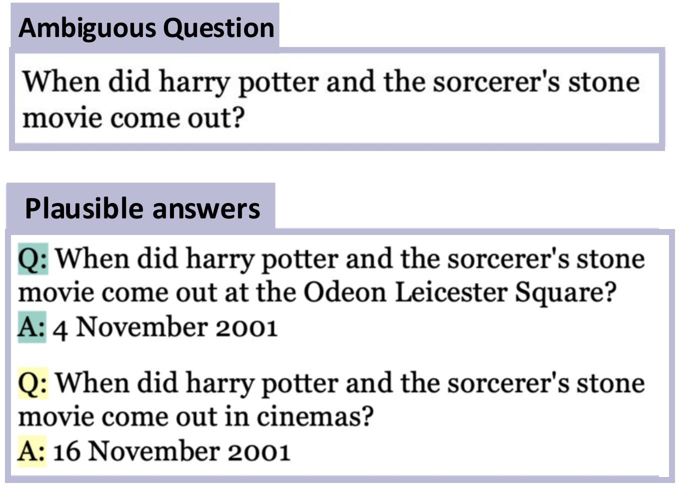
Yeonjun In, Sungchul Kim, Ryan A. Rossi, Md Mehrab Tanjim,
Tong Yu, Ritwik Sinha, Chanyoung Park

Korea Advanced Institute of Science and Technology (KAIST)
Adobe Research

Problem Formulation: Ambiguous Questions in QA Systems

In real-world QA systems, users often pose ***ambiguous questions (AQs)*** with multiple plausible interpretations

➡ RAG framework has emerged as a promising solution



Ambiguous Question

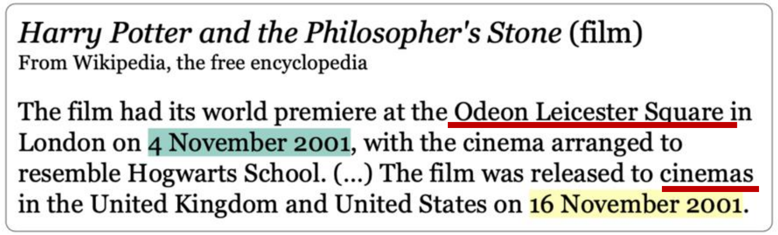
When did harry potter and the sorcerer's stone movie come out?

Plausible answers

Q: When did harry potter and the sorcerer's stone movie come out at the Odeon Leicester Square?
A: 4 November 2001

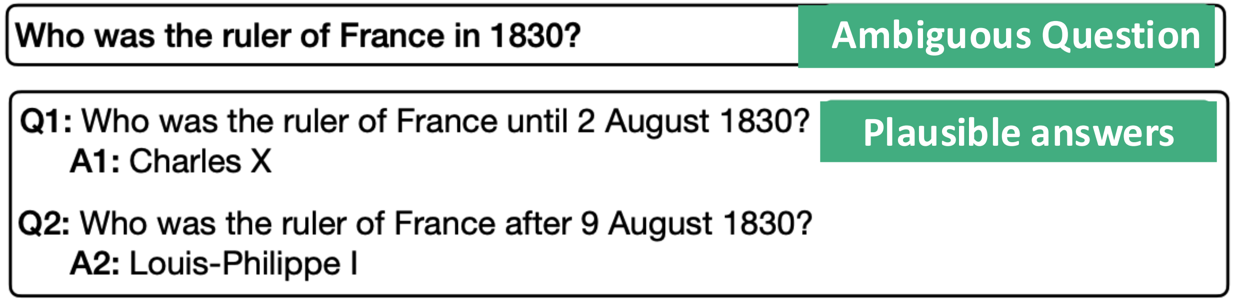
Q: When did harry potter and the sorcerer's stone movie come out in cinemas?
A: 16 November 2001

Retrieve passages !



Harry Potter and the Philosopher's Stone (film)
From Wikipedia, the free encyclopedia

The film had its world premiere at the Odeon Leicester Square in London on 4 November 2001, with the cinema arranged to resemble Hogwarts School. (...) The film was released to cinemas in the United Kingdom and United States on 16 November 2001.



Who was the ruler of France in 1830?

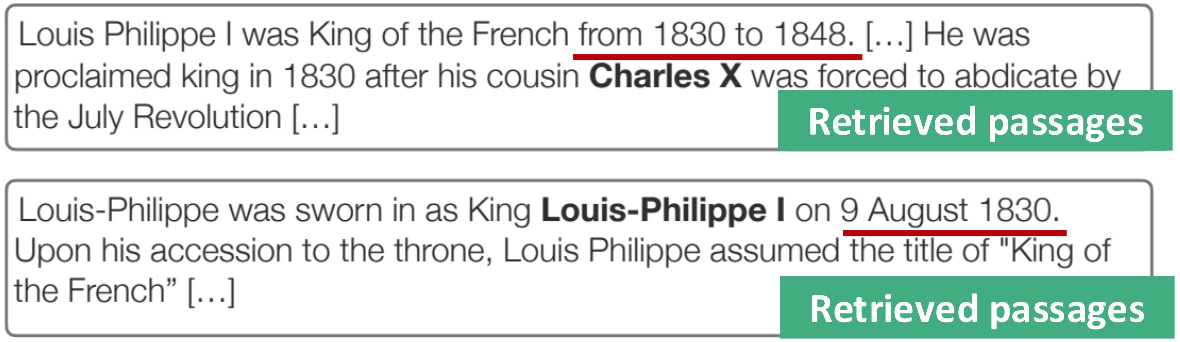
Ambiguous Question

Q1: Who was the ruler of France until 2 August 1830?
A1: Charles X

Q2: Who was the ruler of France after 9 August 1830?
A2: Louis-Philippe I

Plausible answers

Retrieve passages !



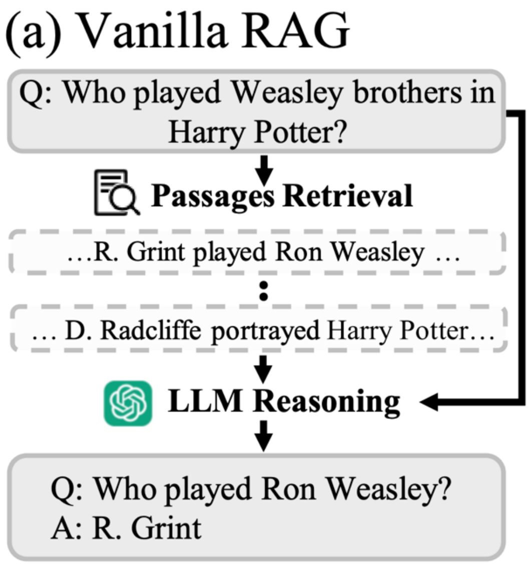
Louis Philippe I was King of the French from 1830 to 1848. […] He was proclaimed king in 1830 after his cousin **Charles X** was forced to abdicate by the July Revolution […]

Retrieved passages

Louis-Philippe was sworn in as King **Louis-Philippe I** on 9 August 1830. Upon his accession to the throne, Louis Philippe assumed the title of "King of the French" […]

Retrieved passages

Is a single retrieval process sufficient?



(a) Vanilla RAG

Q: Who played Weasley brothers in Harry Potter?

Passages Retrieval

…R. Grint played Ron Weasley …

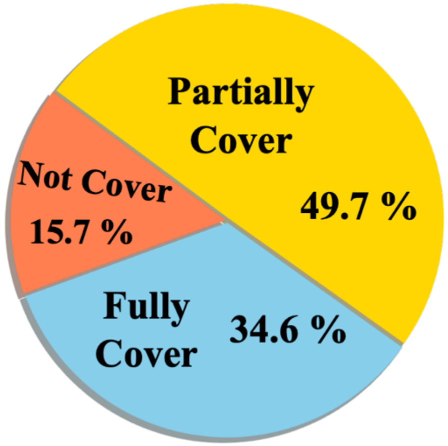
⋮

… D. Radcliffe portrayed Harry Potter…

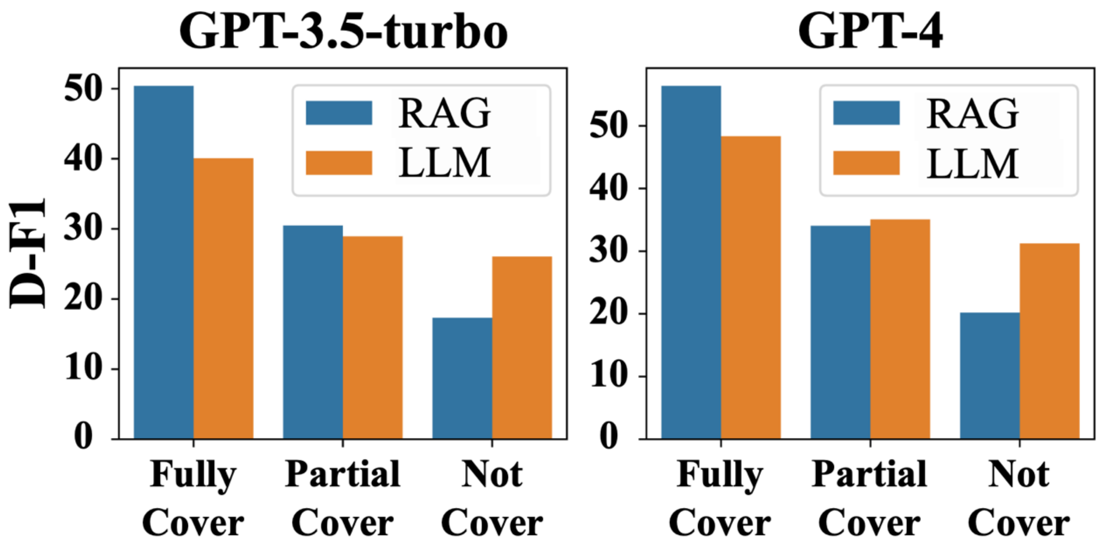
LLM Reasoning

Q: Who played Ron Weasley?
A: R. Grint

(a)



(b)



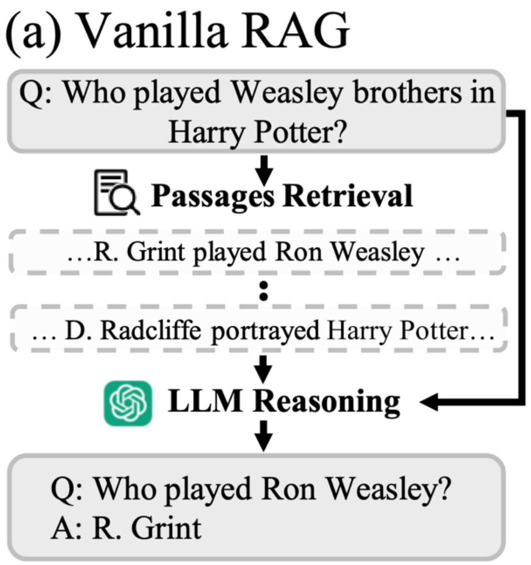
Observations

(a) The ideal retrieval case is **only 35%.**

(b) The RAG method significantly degenerates for **“Partially Cover” and “Not Cover.”**

(c) The closed-book LLM notably outperforms RAG for **“Not Cover.”**

Is a single retrieval process sufficient?



(a) Vanilla RAG

Q: Who played Weasley brothers in Harry Potter?

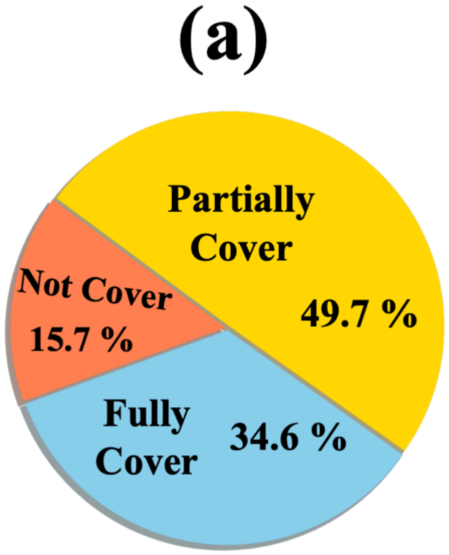
Passages Retrieval

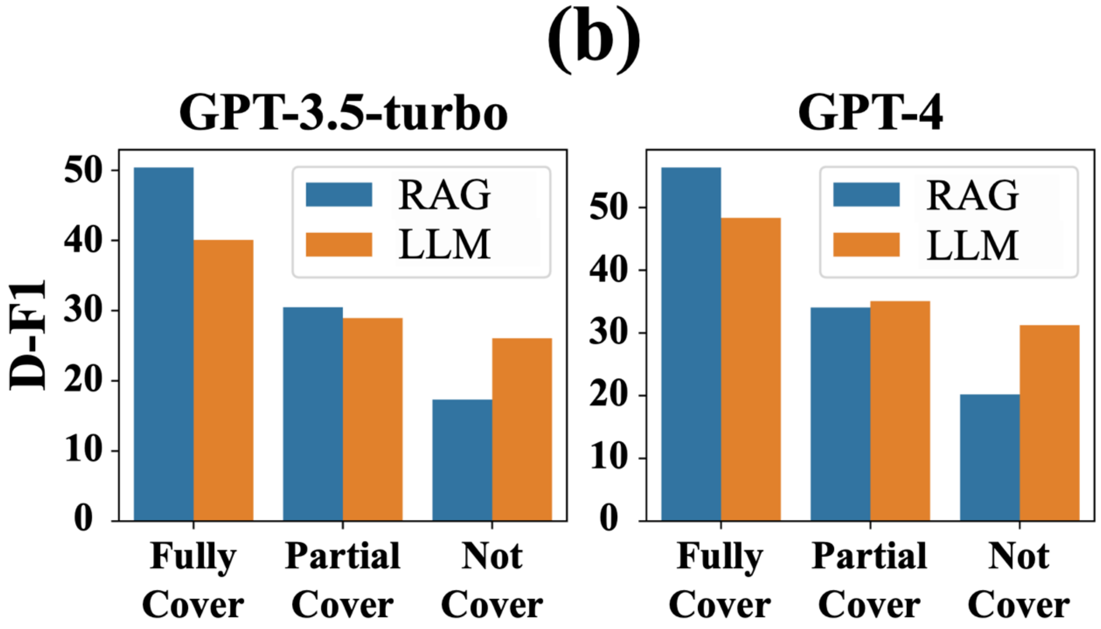
…R. Grint played Ron Weasley …

… D. Radcliffe portrayed Harry Potter…

LLM Reasoning

Q: Who played Ron Weasley?
A: R. Grint



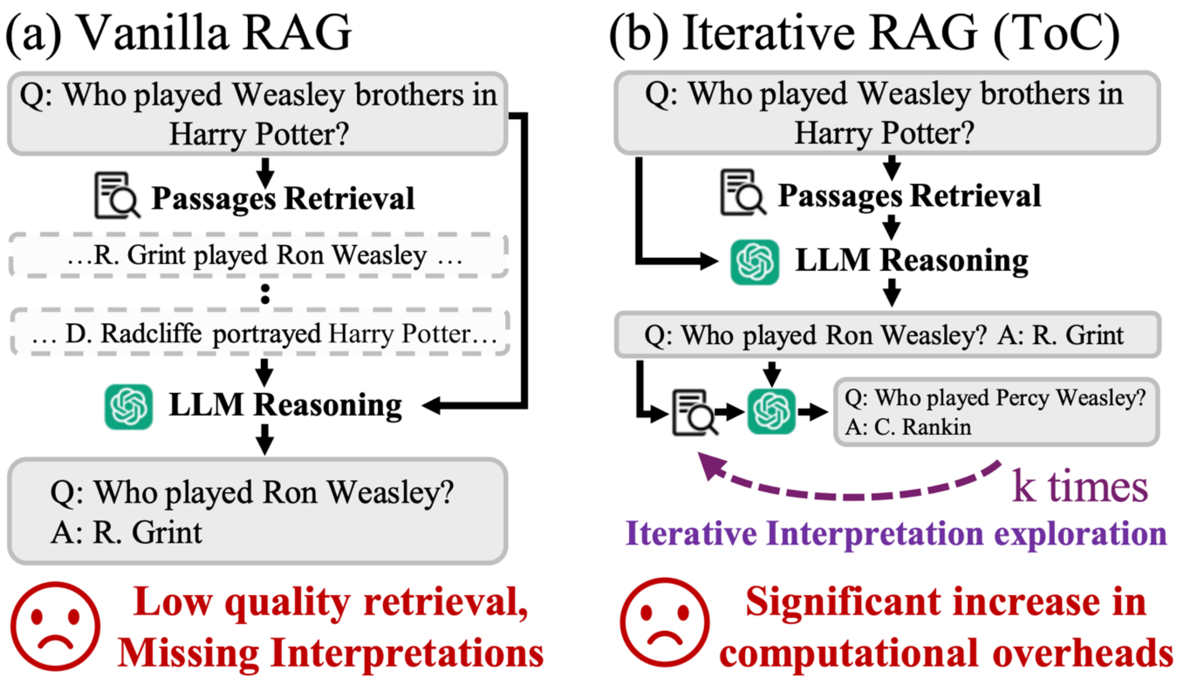


Insights

1. The quality of retrieval is crucial for the performance of the RAG framework in Ambiguous QA
2. If passages are of extremely low quality, LLM internal knowledge is more beneficial than RAG.

Inefficiency of Iterative RAG

Iterative RAG handles low quality retrieval ***with significant increase in computational overheads***



(a) Vanilla RAG

Q: Who played Weasley brothers in Harry Potter?

Passages Retrieval

…R. Grint played Ron Weasley …

:

… D. Radcliffe portrayed Harry Potter…

LLM Reasoning

Q: Who played Ron Weasley?
A: R. Grint

☹ Low quality retrieval, Missing Interpretations

(b) Iterative RAG (ToC)

Q: Who played Weasley brothers in Harry Potter?

Passages Retrieval

LLM Reasoning

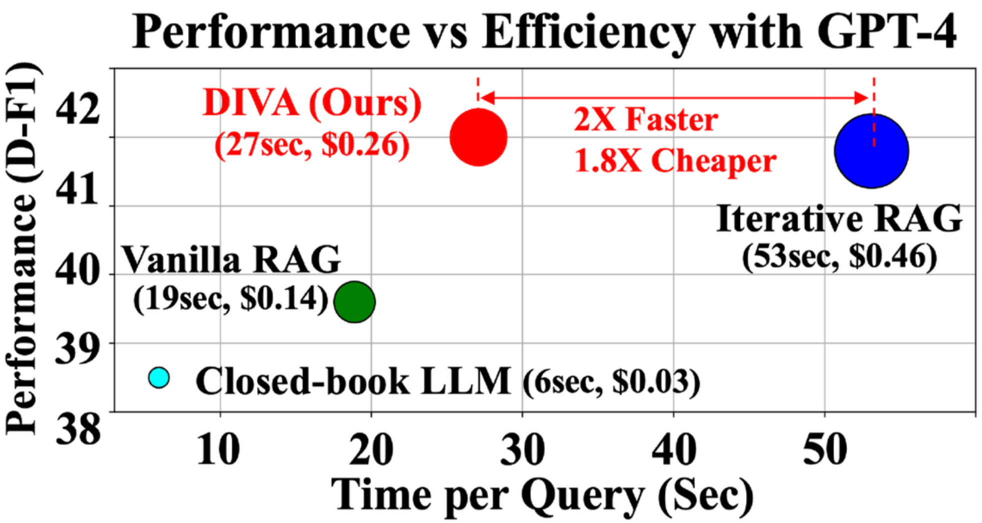
Q: Who played Ron Weasley? A: R. Grint

Q: Who played Percy Weasley?
A: C. Rankin

k times

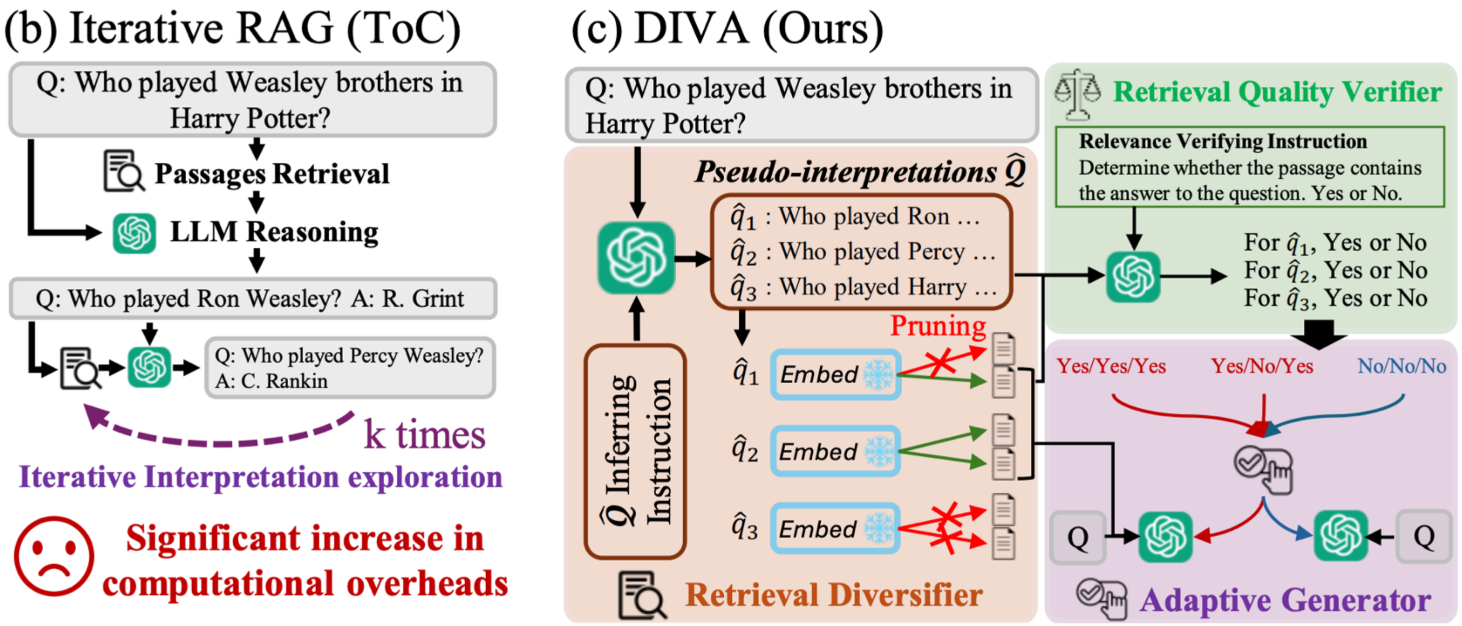
Iterative Interpretation exploration

☹ Significant increase in computational overheads

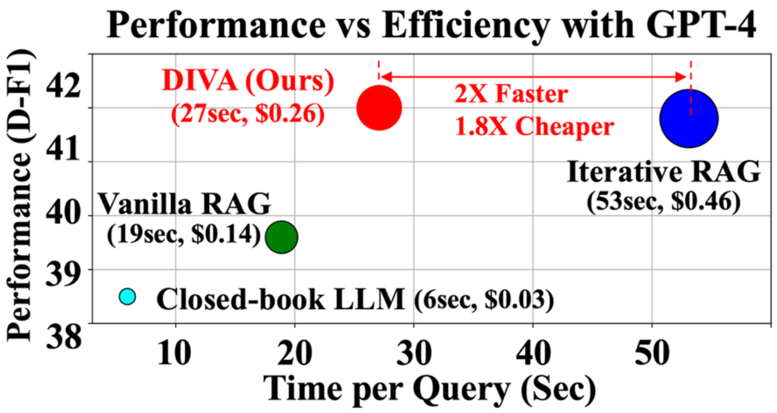


Diversify-verify-adapt (DIVA): Overview

DIVA is robust under low quality retrieval, while notably efficient (2x faster, 1.8x cheaper)

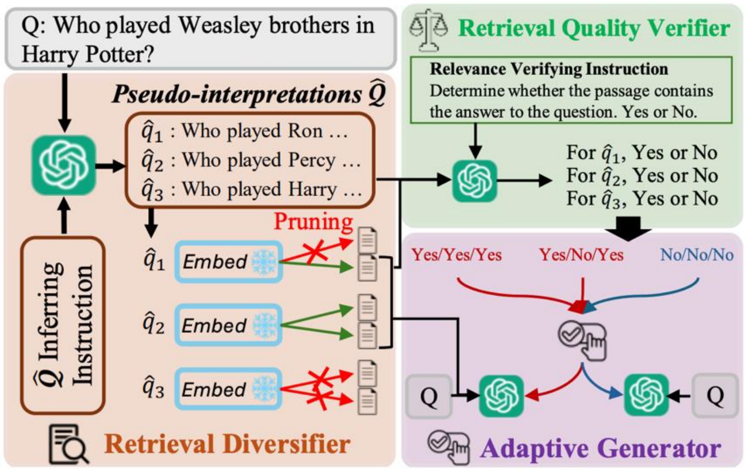


(b) Iterative RAG (ToC)
Q: Who played Weasley brothers in Harry Potter?
Passages Retrieval
LLM Reasoning
Q: Who played Ron Weasley? A: R. Grint
Q: Who played Percy Weasley? A: C. Rankin
k times
Iterative Interpretation exploration
Significant increase in computational overheads
(c) DIVA (Ours)
Q: Who played Weasley brothers in Harry Potter?
Pseudo-interpretations $\hat{Q}$
$\hat{q}_1$: Who played Ron …
$\hat{q}_2$: Who played Percy …
$\hat{q}_3$: Who played Harry …
$\hat{Q}$ Inferring Instruction
Pruning
$\hat{q}_1$ Embed
$\hat{q}_2$ Embed
$\hat{q}_3$ Embed
Retrieval Diversifier
Retrieval Quality Verifier
Relevance Verifying Instruction
Determine whether the passage contains the answer to the question. Yes or No.
For $\hat{q}_1$, Yes or No
For $\hat{q}_2$, Yes or No
For $\hat{q}_3$, Yes or No
Yes/Yes/Yes
Yes/No/Yes
No/No/No
Q
Q
Adaptive Generator



Diversify-verify-adapt (DIVA)

(c) DIVA (Ours)


Q: Who played Weasley brothers in Harry Potter?
Pseudo-interpretations $\hat{Q}$
$\hat{q}_1$: Who played Ron …
$\hat{q}_2$: Who played Percy …
$\hat{q}_3$: Who played Harry …
Pruning
$\hat{Q}$ Inferring Instruction
$\hat{q}_1$ Embed
$\hat{q}_2$ Embed
$\hat{q}_3$ Embed
Retrieval Diversifier
Retrieval Quality Verifier
Relevance Verifying Instruction
Determine whether the passage contains the answer to the question. Yes or No.
For $\hat{q}_1$, Yes or No
For $\hat{q}_2$, Yes or No
For $\hat{q}_3$, Yes or No
Yes/Yes/Yes
Yes/No/Yes
No/No/No
Q
Q
Adaptive Generator

Retrieval Diversification

efficiently retrieves passages $\mathcal{P}_i$ without any iterative process.

Stage 1

- infers pseudo-interpretations $Q_i = \{\hat{q}_{i,1}, \hat{q}_{i,2}, \dots\}$, each of which related to a true plausible answer, by mimicking human's reasoning chain.

Q. Who played the weasley brothers in harry potter?

1. Which part is ambiguous?

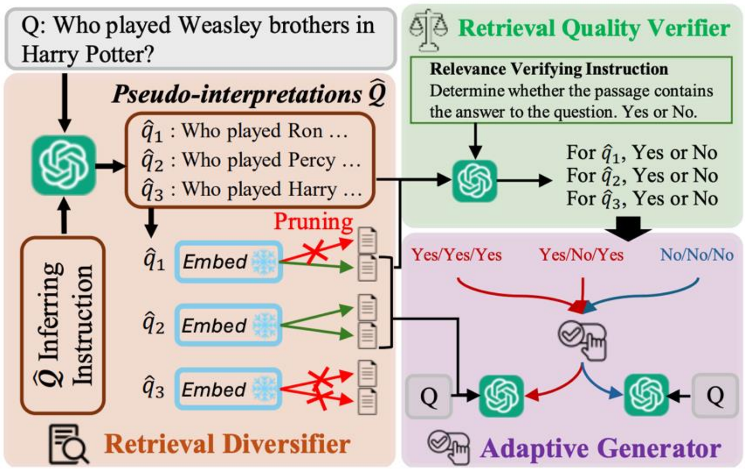
- the "Weasley brothers"

2. Why that part is ambiguous?

- It can be Ron, Percy, etc.

"Who played Ron Weasley in Harry Potter?", "Who played Percy Weasley in Harry Potter?", etc.

Diversify-verify-adapt (DIVA)



(c) DIVA (Ours)

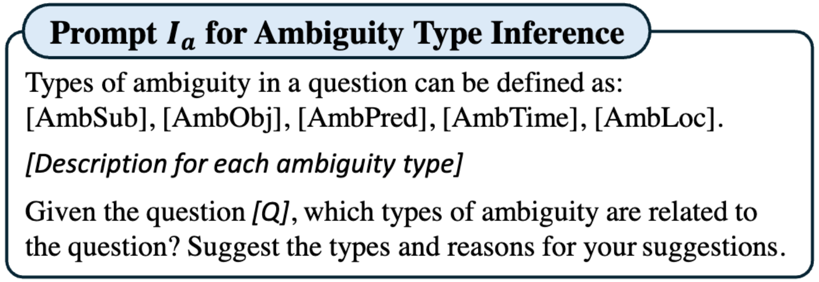
Q: Who played Weasley brothers in Harry Potter?
Pseudo-interpretations $\hat{Q}$
$\hat{q}_1$: Who played Ron …
$\hat{q}_2$: Who played Percy …
$\hat{q}_3$: Who played Harry …
Pruning
$\hat{Q}$ Inferring Instruction
$\hat{q}_1$ Embed
$\hat{q}_2$ Embed
$\hat{q}_3$ Embed
Retrieval Diversifier
Retrieval Quality Verifier
Relevance Verifying Instruction
Determine whether the passage contains the answer to the question. Yes or No.
For $\hat{q}_1$, Yes or No
For $\hat{q}_2$, Yes or No
For $\hat{q}_3$, Yes or No
Yes/Yes/Yes
Yes/No/Yes
No/No/No
Q
Q
Adaptive Generator

Retrieval Diversification

efficiently retrieves passages $\mathcal{P}_i$ without any iterative process.

Stage 1

- infers pseudo-interpretations $\mathcal{Q}_i = \{\hat{q}_{i,1}, \hat{q}_{i,2}, \ldots\}$, each of which related to a true plausible answer, by mimicking human's reasoning chain.

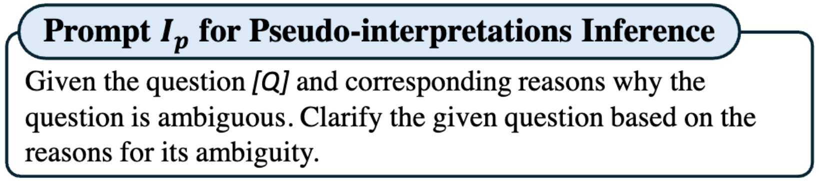


Prompt I_a for Ambiguity Type Inference

Types of ambiguity in a question can be defined as: [AmbSub], [AmbObj], [AmbPred], [AmbTime], [AmbLoc].

[Description for each ambiguity type]

Given the question *[Q]*, which types of ambiguity are related to the question? Suggest the types and reasons for your suggestions.



Prompt I_p for Pseudo-interpretations Inference

Given the question *[Q]* and corresponding reasons why the question is ambiguous. Clarify the given question based on the reasons for its ambiguity.

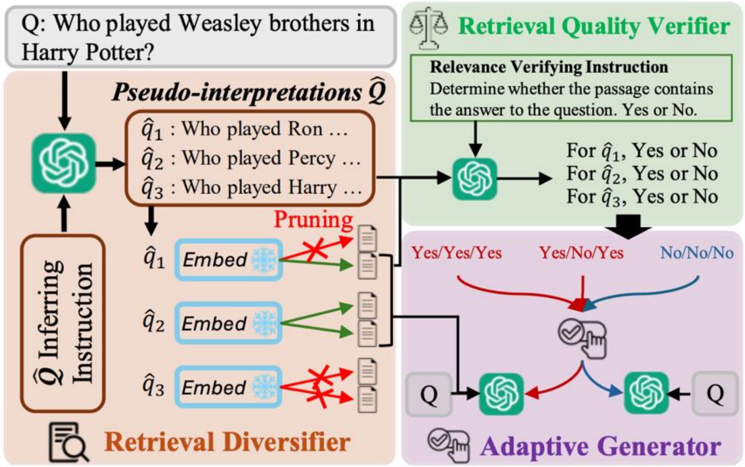
1) identify the ambiguous part of the question and the reason for the ambiguity

2) infer the pseudo-interpretations from the results.

$$\hat{\mathcal{Q}}_i \leftarrow \text{LLM}(q_i, I_\text{p}, \text{LLM}(q_i, I_\text{a})),$$

Diversify-verify-adapt (DIVA)

(c) DIVA (Ours)


Q: Who played Weasley brothers in Harry Potter?
Pseudo-interpretations $\hat{\mathcal{Q}}$
$\hat{q}_1$: Who played Ron …
$\hat{q}_2$: Who played Percy …
$\hat{q}_3$: Who played Harry …
$\hat{\mathcal{Q}}$ Inferring Instruction
Pruning
$\hat{q}_1$ Embed
$\hat{q}_2$ Embed
$\hat{q}_3$ Embed
Retrieval Diversifier
Retrieval Quality Verifier
Relevance Verifying Instruction
Determine whether the passage contains the answer to the question. Yes or No.
For $\hat{q}_1$, Yes or No
For $\hat{q}_2$, Yes or No
For $\hat{q}_3$, Yes or No
Yes/Yes/Yes
Yes/No/Yes
No/No/No
Q
Q
Adaptive Generator

Retrieval Diversification

efficiently retrieves passages $\mathcal{P}_i$ without any iterative process.

Stage 2

- retrieves a set of passages that maximally cover these interpretations

➡ There could be some noisy and irrelevant passages due to

- imperfect retriever
- noise of the inferred pseudo-interpretations

How to handle it?

Our intuition

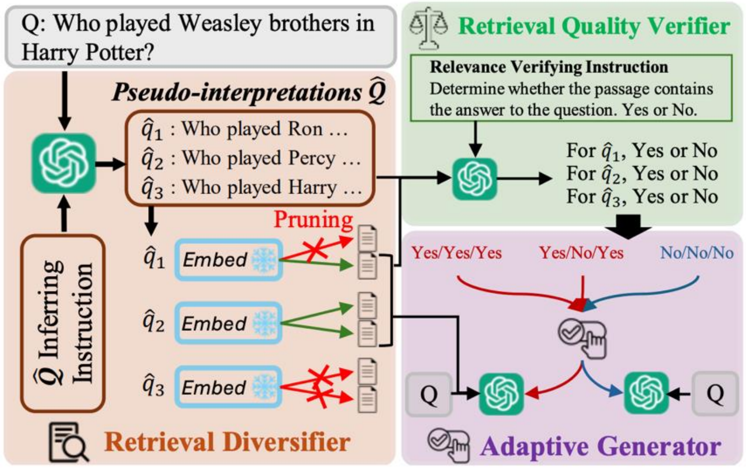
1. noisy passages caused by the imperfect retriever tend to be irrelevant to all pseudo-interpretations
2. noisy passages caused by noisy pseudo-interpretations tend to be irrelevant to most of the pseudo-interpretations.

➡ ***Solution: measuring averaged relevance***

$$\mathcal{S}(p) \leftarrow \frac{1}{|\hat{\mathcal{Q}}_i|}\sum_{j=1}^{|\hat{\mathcal{Q}}_i|}\frac{\mathsf{Enc}(\hat{q}_j)\cdot\mathsf{Enc}(p)}{||\mathsf{Enc}(\hat{q}_j)||\cdot||\mathsf{Enc}(p)||},$$

Diversify-verify-adapt (DIVA)

(c) DIVA (Ours)



Q: Who played Weasley brothers in Harry Potter?
Pseudo-interpretations $\hat{Q}$
$\hat{q}_1$: Who played Ron …
$\hat{q}_2$: Who played Percy …
$\hat{q}_3$: Who played Harry …
$\hat{Q}$ Inferring Instruction
Pruning
$\hat{q}_1$ Embed
$\hat{q}_2$ Embed
$\hat{q}_3$ Embed
Retrieval Diversifier
Retrieval Quality Verifier
Relevance Verifying Instruction
Determine whether the passage contains the answer to the question. Yes or No.
For $\hat{q}_1$, Yes or No
For $\hat{q}_2$, Yes or No
For $\hat{q}_3$, Yes or No
Yes/Yes/Yes
Yes/No/Yes
No/No/No
Q
Q
Adaptive Generator

Retrieval Quality Verifier

verifies the quality of the retrieved passages $\mathcal{P}_i$

Challenge

The retrieval quality in terms of ambiguous questions should be graded according to how many interpretations are encompassed by the retrieved passages.

Solution

Using the inferred pseudo-interpretations.

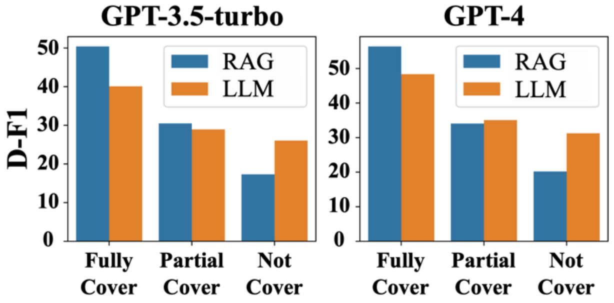
$$V_{i,1} \leftarrow \text{LLM}(\hat{q}_{i,1}, \mathcal{P}_i, I_{\text{v}})$$

$$\vdots$$

$$V_{i,|\hat{\mathcal{Q}}_i|} \leftarrow \text{LLM}(\hat{q}_{i,|\hat{\mathcal{Q}}_i|}, \mathcal{P}_i, I_{\text{v}}),$$

Adaptive Generation

utilizes the most suitable approach tailored to each retrieval quality

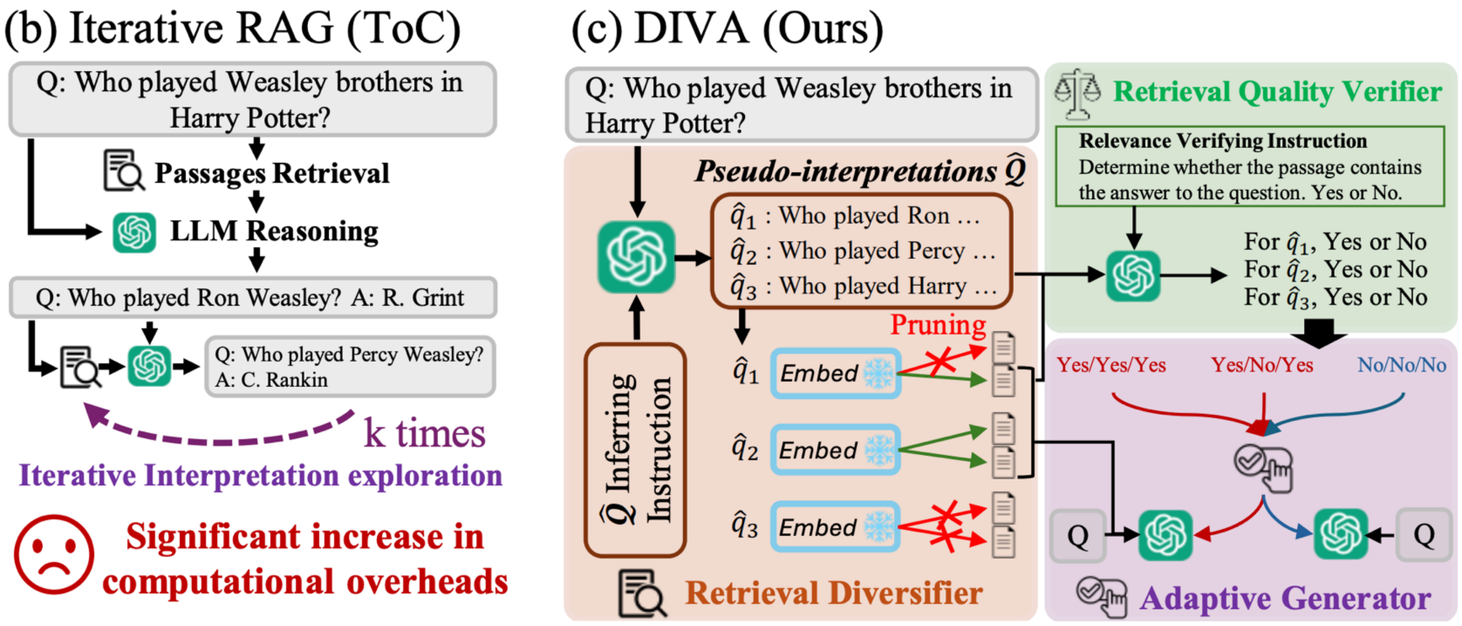


Inspired by our preliminary analyses,

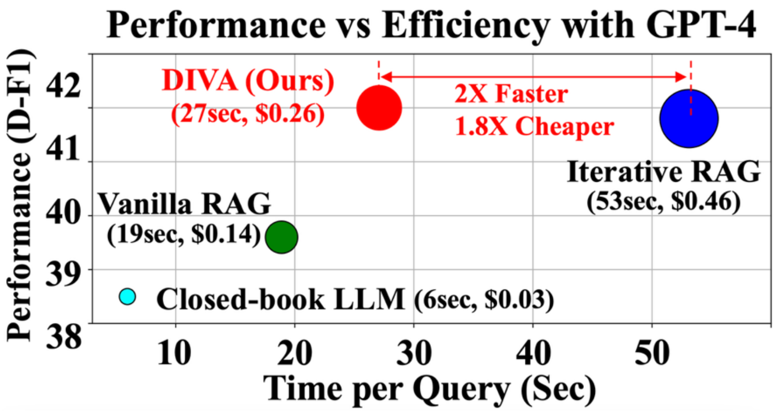
- If $\mathcal{P}_i$ do not cover any pseudo-interpretations, we only utilize the LLM's internal knowledge.
- Otherwise, we utilize the retrieved passage to generate a response.

Diversify-verify-adapt (DIVA)

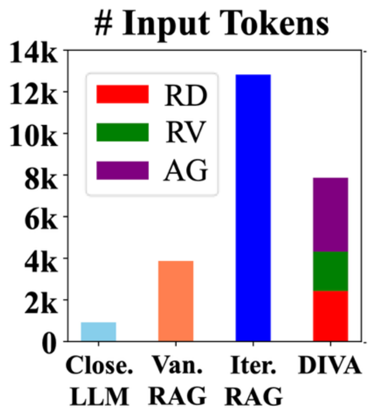
DIVA outperforms the sota baseline, Iterative RAG, in terms of both accuracy and efficiency of response generation.

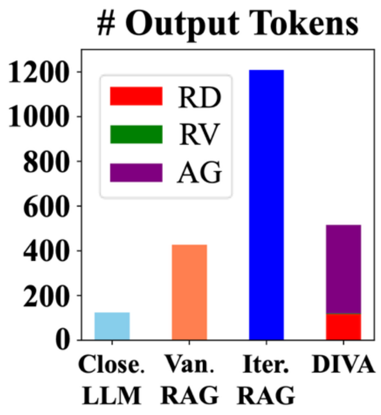


(b) Iterative RAG (ToC)
Q: Who played Weasley brothers in Harry Potter?
Passages Retrieval
LLM Reasoning
Q: Who played Ron Weasley? A: R. Grint
Q: Who played Percy Weasley? A: C. Rankin
k times
Iterative Interpretation exploration
Significant increase in computational overheads
(c) DIVA (Ours)
Q: Who played Weasley brothers in Harry Potter?
Pseudo-interpretations $\hat{Q}$
$\hat{q}_1$: Who played Ron …
$\hat{q}_2$: Who played Percy …
$\hat{q}_3$: Who played Harry …
$\hat{Q}$ Inferring Instruction
Pruning
$\hat{q}_1$ Embed
$\hat{q}_2$ Embed
$\hat{q}_3$ Embed
Retrieval Diversifier
Retrieval Quality Verifier
Relevance Verifying Instruction
Determine whether the passage contains the answer to the question. Yes or No.
For $\hat{q}_1$, Yes or No
For $\hat{q}_2$, Yes or No
For $\hat{q}_3$, Yes or No
Yes/Yes/Yes
Yes/No/Yes
No/No/No
Q
Q
Adaptive Generator

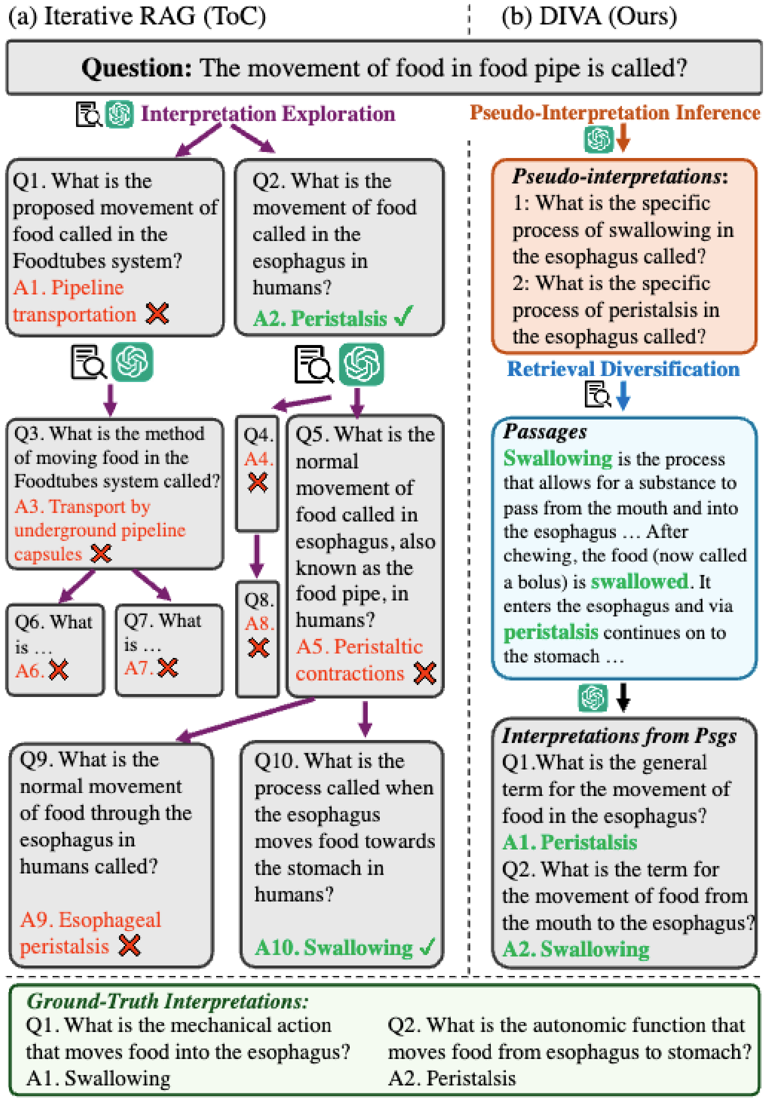


Diversify-verify-adapt (DIVA)





- **DIVA's strong efficiency is largely due to the RD method.**
- **Although the RV method introduces some additional costs, these are acceptable compared to the complexity of Iterative RAG.**



(a) Iterative RAG (ToC)

(b) DIVA (Ours)

Question: The movement of food in food pipe is called?

Interpretation Exploration

Q1. What is the proposed movement of food called in the Foodtubes system?
A1. Pipeline transportation ✗

Q2. What is the movement of food called in the esophagus in humans?
A2. Peristalsis ✓

Q3. What is the method of moving food in the Foodtubes system called?
A3. Transport by underground pipeline capsules ✗

Q4.
A4. ✗

Q5. What is the normal movement of food called in esophagus, also known as the food pipe, in humans?
A5. Peristaltic contractions ✗

Q6. What is …
A6. ✗

Q7. What is …
A7. ✗

Q8.
A8. ✗

Q9. What is the normal movement of food through the esophagus in humans called?
A9. Esophageal peristalsis ✗

Q10. What is the process called when the esophagus moves food towards the stomach in humans?
A10. Swallowing ✓

Pseudo-Interpretation Inference

***Pseudo-interpretations*:**
1: What is the specific process of swallowing in the esophagus called?
2: What is the specific process of peristalsis in the esophagus called?

Retrieval Diversification

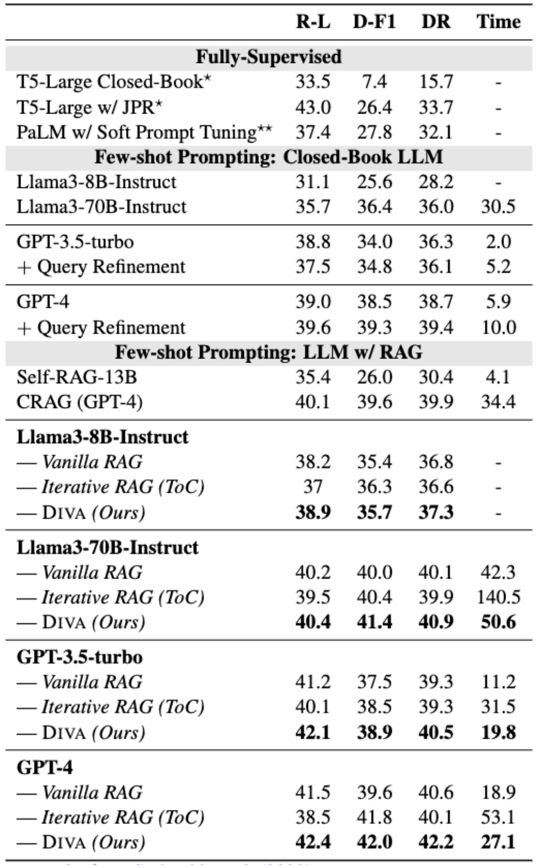
Passages
Swallowing is the process that allows for a substance to pass from the mouth and into the esophagus … After chewing, the food (now called a bolus) is **swallowed**. It enters the esophagus and via **peristalsis** continues on to the stomach …

Interpretations from Psgs
Q1.What is the general term for the movement of food in the esophagus?
A1. Peristalsis
Q2. What is the term for the movement of food from the mouth to the esophagus?
A2. Swallowing

Ground-Truth Interpretations:
Q1. What is the mechanical action that moves food into the esophagus?
A1. Swallowing
Q2. What is the autonomic function that moves food from esophagus to stomach?
A2. Peristalsis

Figure 8: Case study with GPT-4.

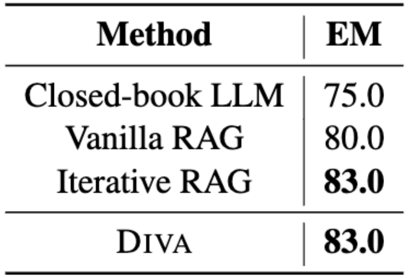
Diversify-verify-adapt (DIVA)



	R-L	D-F1	DR	Time
Fully-Supervised				
T5-Large Closed-Book*	33.5	7.4	15.7	-
T5-Large w/ JPR*	43.0	26.4	33.7	-
PaLM w/ Soft Prompt Tuning**	37.4	27.8	32.1	-
Few-shot Prompting: Closed-Book LLM				
Llama3-8B-Instruct	31.1	25.6	28.2	-
Llama3-70B-Instruct	35.7	36.4	36.0	30.5
GPT-3.5-turbo	38.8	34.0	36.3	2.0
+ Query Refinement	37.5	34.8	36.1	5.2
GPT-4	39.0	38.5	38.7	5.9
+ Query Refinement	39.6	39.3	39.4	10.0
Few-shot Prompting: LLM w/ RAG				
Self-RAG-13B	35.4	26.0	30.4	4.1
CRAG (GPT-4)	40.1	39.6	39.9	34.4
Llama3-8B-Instruct				
— *Vanilla RAG*	38.2	35.4	36.8	-
— *Iterative RAG (ToC)*	37	36.3	36.6	-
— DIVA *(Ours)*	**38.9**	**35.7**	**37.3**	-
Llama3-70B-Instruct				
— *Vanilla RAG*	40.2	40.0	40.1	42.3
— *Iterative RAG (ToC)*	39.5	40.4	39.9	140.5
— DIVA *(Ours)*	**40.4**	**41.4**	**40.9**	**50.6**
GPT-3.5-turbo				
— *Vanilla RAG*	41.2	37.5	39.3	11.2
— *Iterative RAG (ToC)*	40.1	38.5	39.3	31.5
— DIVA *(Ours)*	**42.1**	**38.9**	**40.5**	**19.8**
GPT-4				
— *Vanilla RAG*	41.5	39.6	40.6	18.9
— *Iterative RAG (ToC)*	38.5	41.8	40.1	53.1
— DIVA *(Ours)*	**42.4**	**42.0**	**42.2**	**27.1**

- **DIVA demonstrates good adaptability in switching out the underlying LLM backbones.**
  - **Llama-3-8B-Instruct, Llama-3-70B-Instruct, GPT-3.5-turbo, GPT-4**

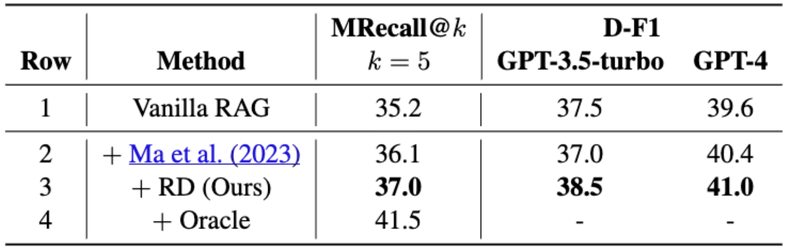
- **DIVA exhibits strong performance on unambiguous questions as well**



Method	EM
Closed-book LLM	75.0
Vanilla RAG	80.0
Iterative RAG	**83.0**
DIVA	**83.0**

Table 5: QA performance on unambiguous questions.

- **Retrieval Diversification (RD) method improves retrieval accuracy, resulting in improvement of QA performance.**



		MRecall@k	**D-F1**	
Row	**Method**	$k = 5$	**GPT-3.5-turbo**	**GPT-4**
1	Vanilla RAG	35.2	37.5	39.6
2	+ Ma et al. (2023)	36.1	37.0	40.4
3	+ RD (Ours)	**37.0**	**38.5**	**41.0**
4	+ Oracle	41.5	-	-

Conclusion

- In this study, we examined the shortcomings of the current RAG-based method in dealing with ambiguous questions, specifically its low-quality retrieval and inefficiency.
- Our proposed framework, DIVA, effectively diversifies the retrieved passages to capture various interpretations, verifies their quality, and adapts the most appropriate approach based on that quality.
- This strategy improves QA performance while minimizing inefficiency.

NAACL 2025

Thank you for listening!

Please refer to our paper

“Diversify-verify-adapt: Efficient and Robust Retrieval-Augmented Ambiguous Question Answering"

Contact: yeonjun.in@kaist.ac.kr
Personal homepage: yeonjun-in.notion.site

Adobe KAIST DSAIL Data Science & Artificial Intelligence