간단 논세

RAFT: Adapting Language Model to Domain Specific RAG

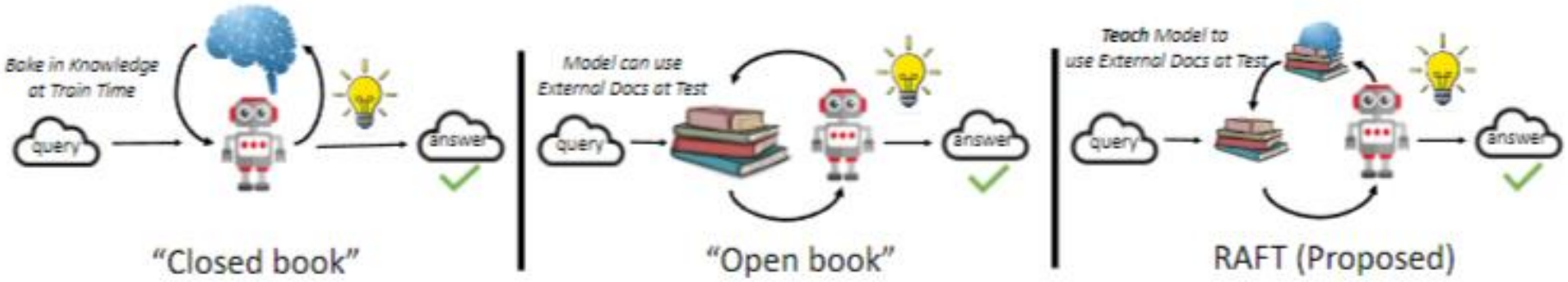
Zhang, Tianjun, et al. "Raft: Adapting language model to domain specific rag." *arXiv preprint arXiv:2403.10131* (2024).

24.11.27

유하영

Introduction

- LLM은 general knowledge reasoning tasks 에서 상당한 advanced를 이루었다.
- Specialized domain에 대한 task를 지원하고자 함. (e.g., legal or medical documents)
  일반 지식추론 < **maximize accuracy**(문서 지식기반)
- **Adapting** LLMs to the specialized domains
- How best to prepare for an Exam?



Fine-tuning

- 테스트 시간에 문서를 제대로 활용 X
- 훈련 중 검색과정의 불완전성

In-Context retrieval methods

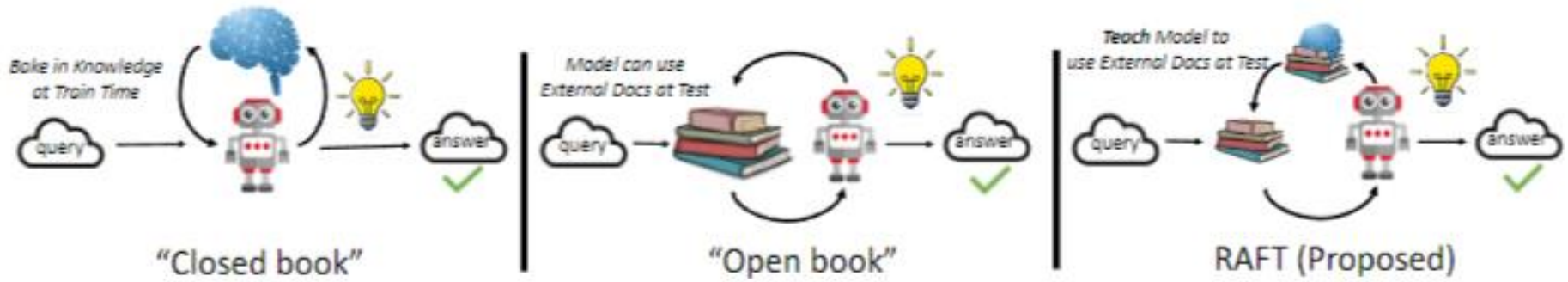
'고정된 domain에서 제공하는 학습 기회 활용 X'

- 시험 문서에 대한 조기 접근이, 제공하는 학습기회를 제대로 활용할 수 없게 됨

LLM을 특정 도메인에 맞게 training하여,
해당 도메인 내에서만 답을 찾고 생성할 수 있는 방법 제안.

Introduction

- LLM은 general knowledge reasoning tasks 에서 상당한 advanced를 이루었다.
- Specialized domain에 대한 task를 지원하고자 함. (e.g., legal or medical documents)
  일반 지식추론 < **maximize accuracy**(문서 지식기반)
- **Adapting** LLMs to the specialized domains
- How best to prepare for an Exam?



Fine-tuning

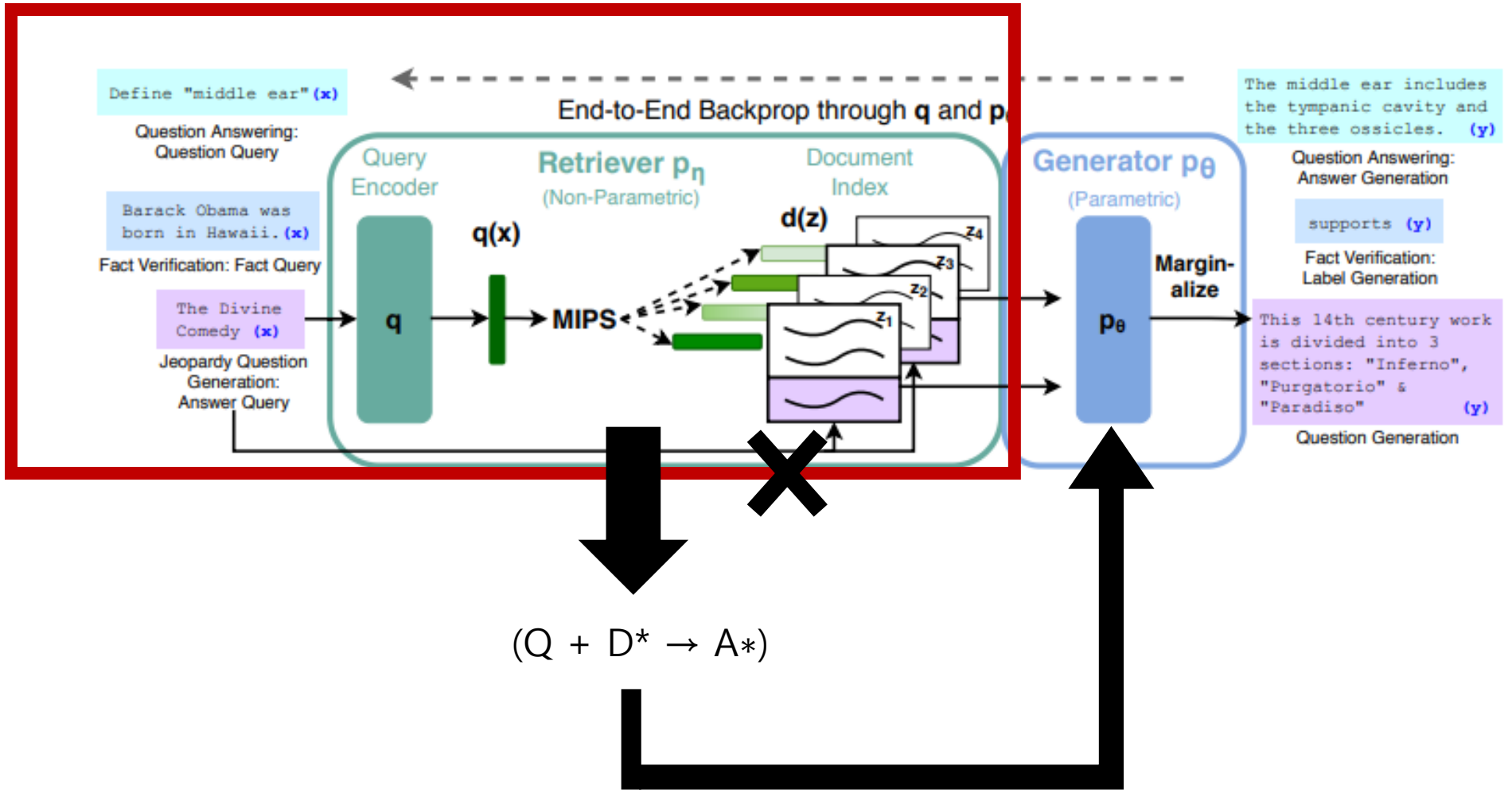
In-Context retrieval methods

Instruction fine-tuning (IFT) + retrieval augmented generation (RAG)

= Retrieval-Augmented Fine Tuning(RAFT)

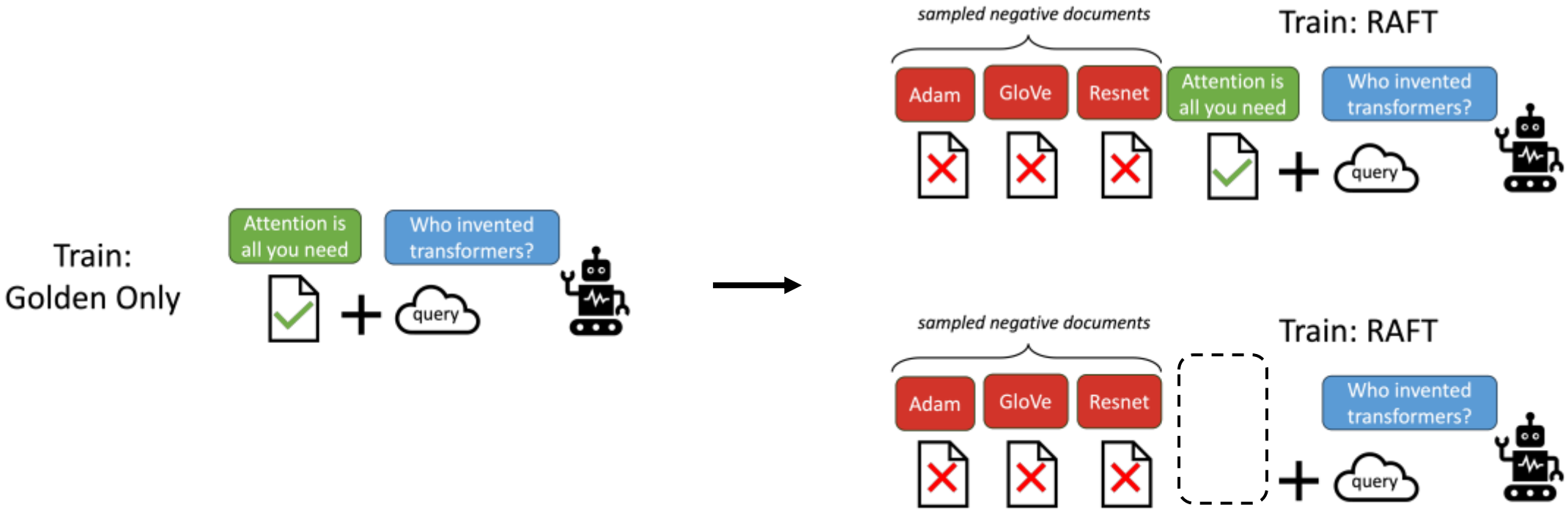
*LLM이 domain specific 지식을 배우고, RAG 성능 개선
*혼란을 주는 distractor document는 ignore, 올바른 정보만을 선별할 수 있게끔.

Method

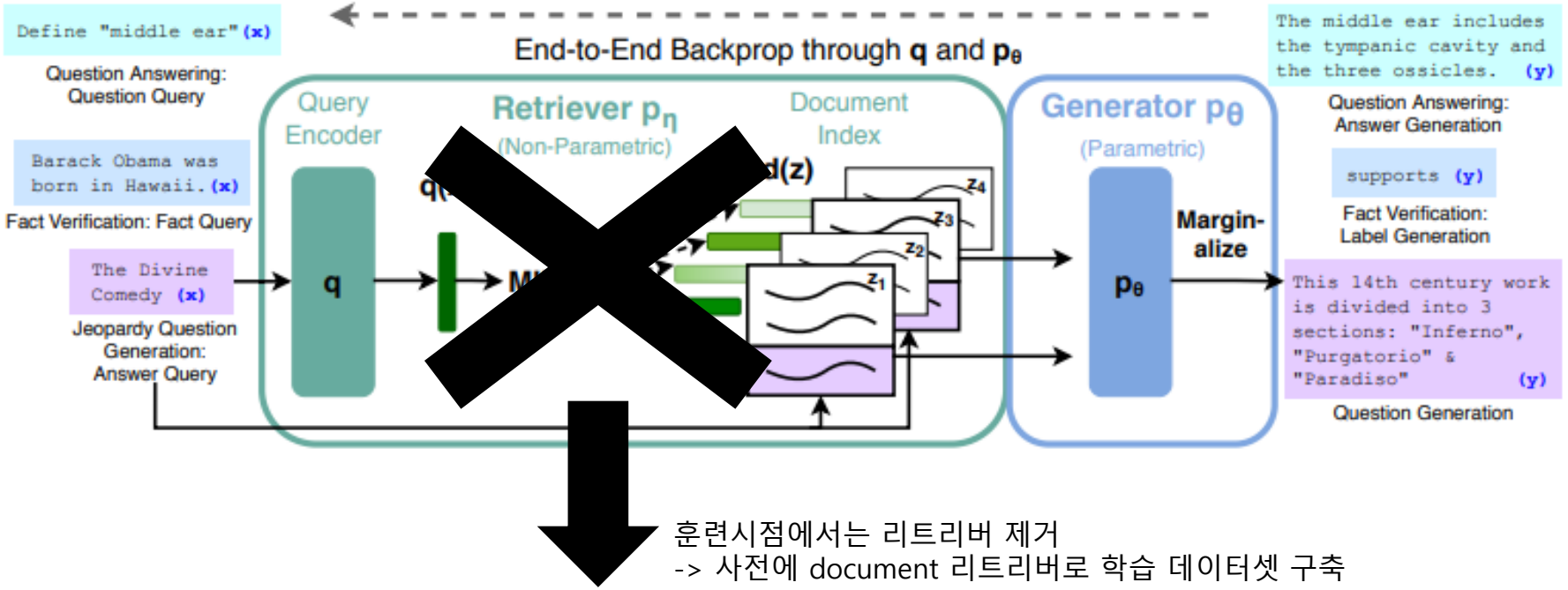

Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and p
Query Encoder
Retriever pη
(Non-Parametric)
Document Index
q
q(x)
MIPS
d(z)
z4
z3
z2
z1
Generator pθ
(Parametric)
pθ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation
(Q + D* → A*)

Method

Overview of our RAFT method



Method



훈련시점에서는 리트리버 제거
-> 사전에 document 리트리버로 학습 데이터셋 구축

(Q + D* → A*)

$$\mathbf{P}\,\%\text{ of data: } \mathbf{Q} + \mathbf{D}^* + \mathbf{D}_1 + \mathbf{D}_2 + \ldots + \mathbf{D}_k \rightarrow \mathbf{A}*$$

$$(1 - \mathbf{P})\,\%\text{ of data: } \mathbf{Q} + \mathbf{D}_1 + \mathbf{D}_2 + \ldots + \mathbf{D}_k \rightarrow \mathbf{A}*$$

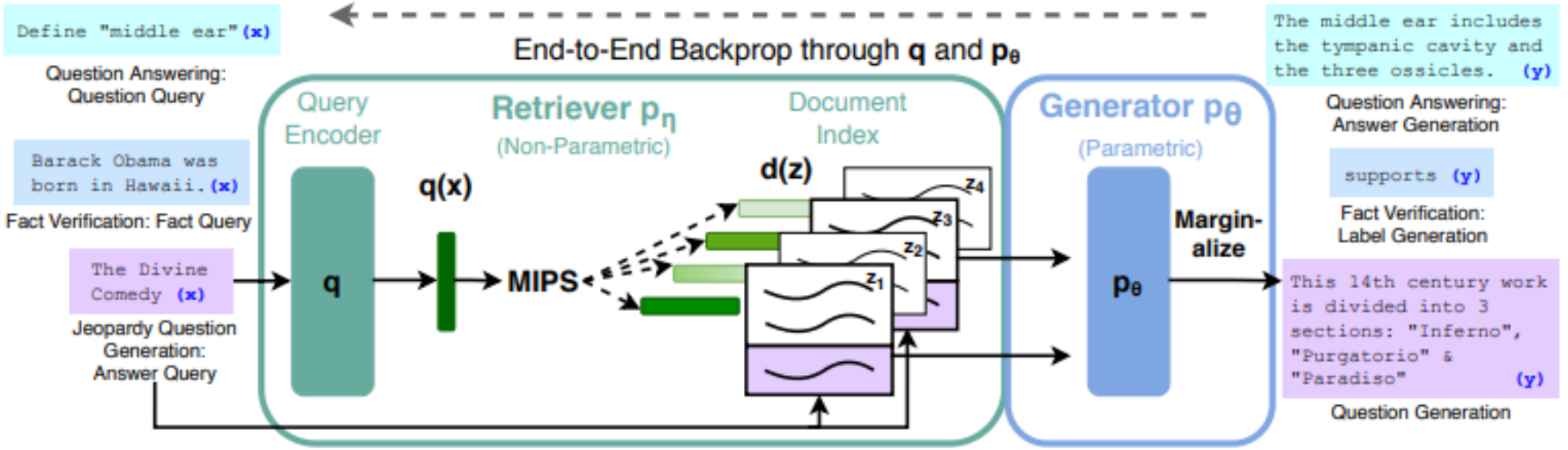
Method

RAFT Prompt

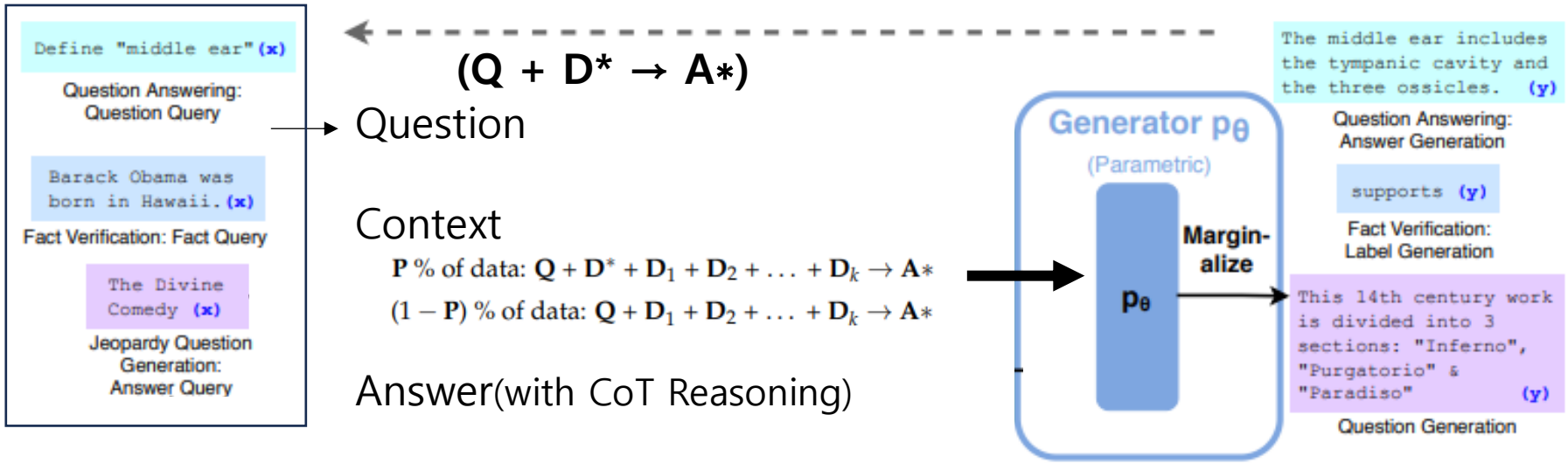
```
Question: The Oberoi family is part of a hotel company that has a head office
in what city?

context: [The Oberoi family is an Indian family that is famous for its
involvement in hotels, namely through The Oberoi Group]...[It is located in
city center of Jakarta, near Mega Kuningan, adjacent to the sister JW Marriott
Hotel. It is operated by The Ritz-Carlton Hotel Company. The complex has two
towers that comprises a hotel and the Airlangga Apartment respectively]...[The
Oberoi Group is a hotel company with its head office in Delhi.]

Instruction: Given the question, context and answer above, provide a logical
reasoning for that answer. Please use the format of: ##Reason: {reason}
##Answer: {answer}.

--------------------------------------------------------------------------------

CoT Answer: ##Reason: The document ##begin_quote## The Oberoi family is an
Indian family that is famous for its involvement in hotels, namely through The
Oberoi Group. ##end_quote## establishes that the Oberoi family is involved in
the Oberoi group, and the document ##begin_quote## The Oberoi Group is a hotel
company with its head office in Delhi. ##end_quote## establishes the head
office of The Oberoi Group. Therefore, the Oberoi family is part of a hotel
company whose head office is in Delhi. ##Answer: Delhi
```

Method

RAG



RAFT



Experiments

RAFT improves RAG performance for all specialized domains

	PubMed	HotpotQA	HuggingFace	Torch Hub	TensorFlow Hub
GPT-3.5 + RAG	71.60	**41.5**	29.08	60.21	65.59
LLaMA2-7B	56.5	0.54	0.22	0	0
LLaMA2-7B + RAG	58.8	0.03	26.43	08.60	43.06
DSF	59.7	6.38	61.06	84.94	86.56
DSF + RAG	71.6	4.41	42.59	82.80	60.29
RAFT (LLaMA2-7B)	**73.30**	35.28	**74.00**	**84.95**	**86.86**

의학 (이진 yes/no)

다중 문서와의 연결성 요구

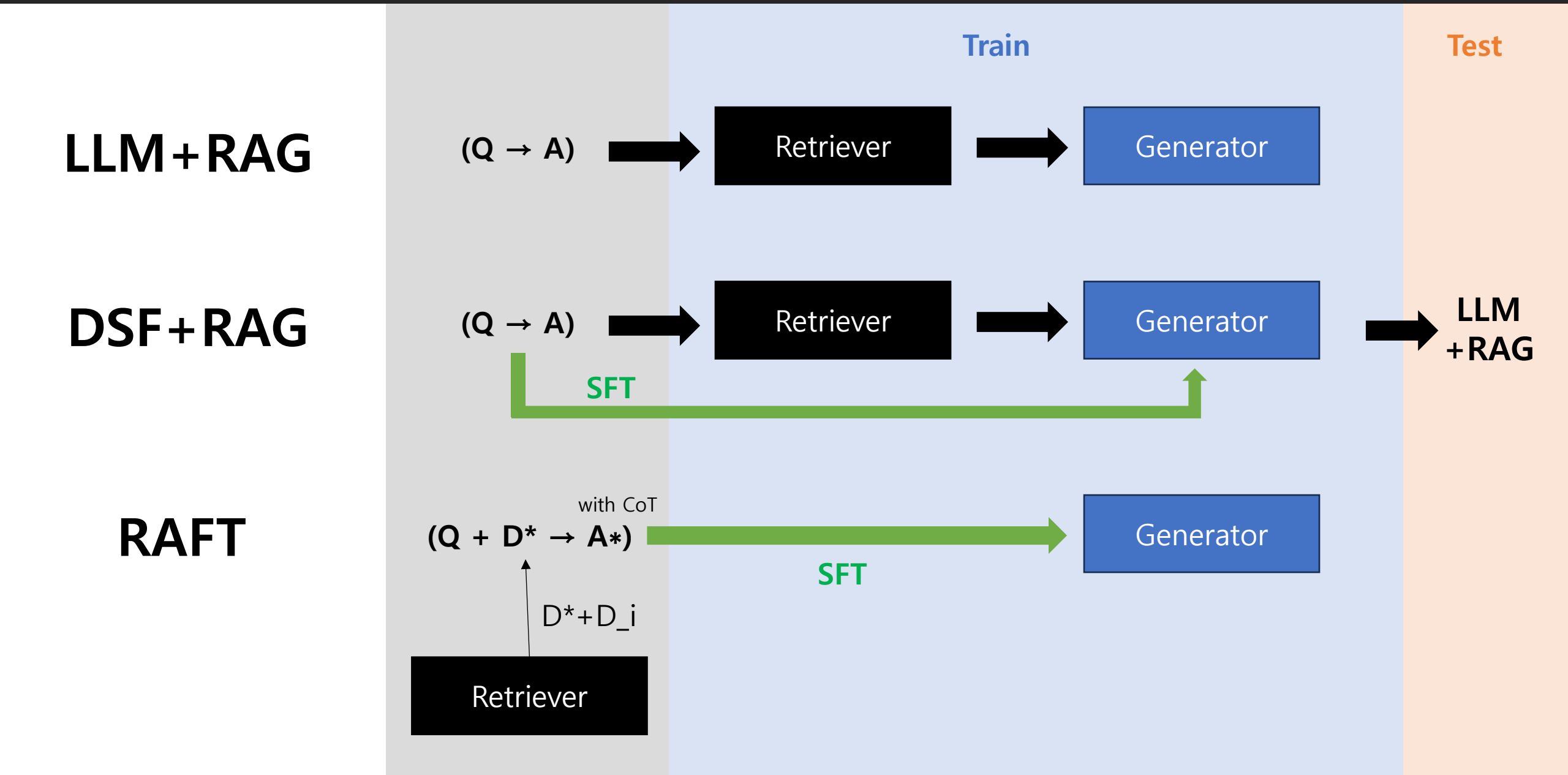
단순한 QA

Ablation on Chain-of-Thought

	PubMed	HotpotQA	HuggingFace	Torch Hub	TensorFlow
RAFT w.o CoT	68.30	25.62	59.07	**86.56**	83.21
RAFT	**73.30**	**35.28**	**74.00**	84.95	**86.86**

GPT-4-1106 CoT Prompt

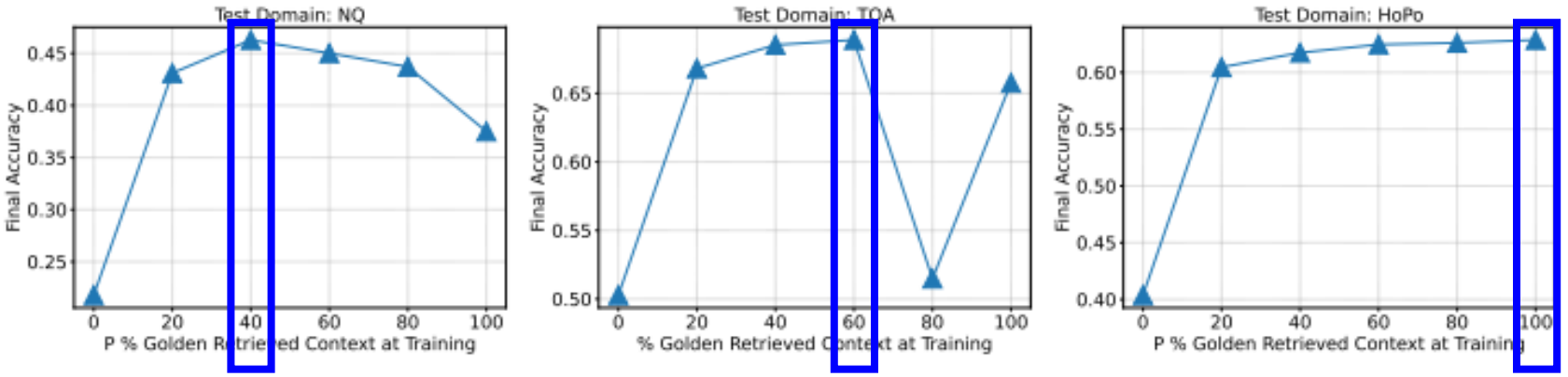
Experiments



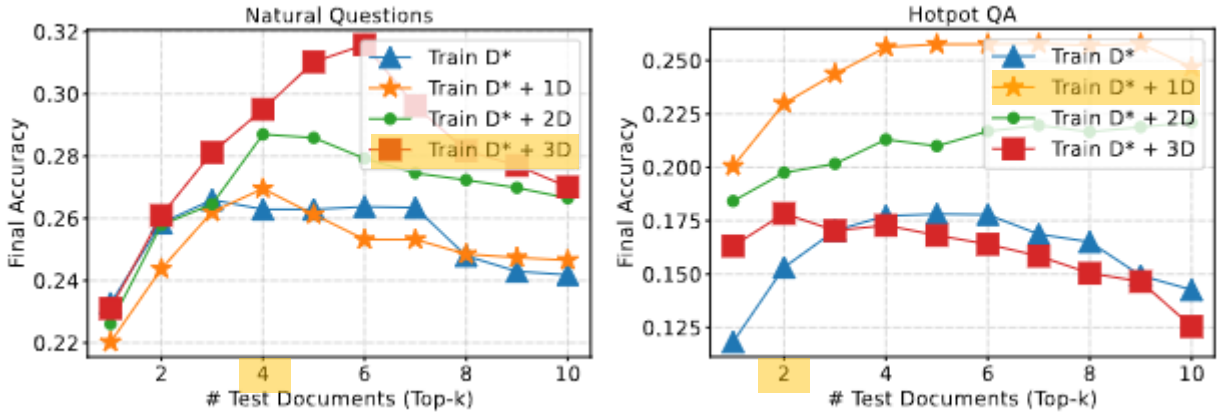
Experiments

How many golden documents to involve?

최적의 P 비율은 도메인마다 다르다.



How robust RAFT is to varying number of test-time document?



- 데이터셋 특성에 따라 다르다.
- irrelevant document에 대한 중요성 시사.

Discussion

도메인 특화 RAG 성능 개선을 위한 새로운 RAG 활용방안 제안

- distractor을 포함한 학습, golden document를 제외하는 방식, CoT로 답변을 생성하는 방식 등

[한계]
- 특정 domain에서의 낮은 성능
- 성능 향상의 대부분은 CoT를 추가했을 때 이루어진 것

[의의]
- CoT와 Distractor의 결합 가능성 증명
- 특정 데이터셋에 대한 성능 향상 달성