xRAG: Extreme Context Compression for Retrieval-augmented Generation with One Token

SIGIR 2024

Dooyoung Kim
Natural Language Processing Lab, SKKU

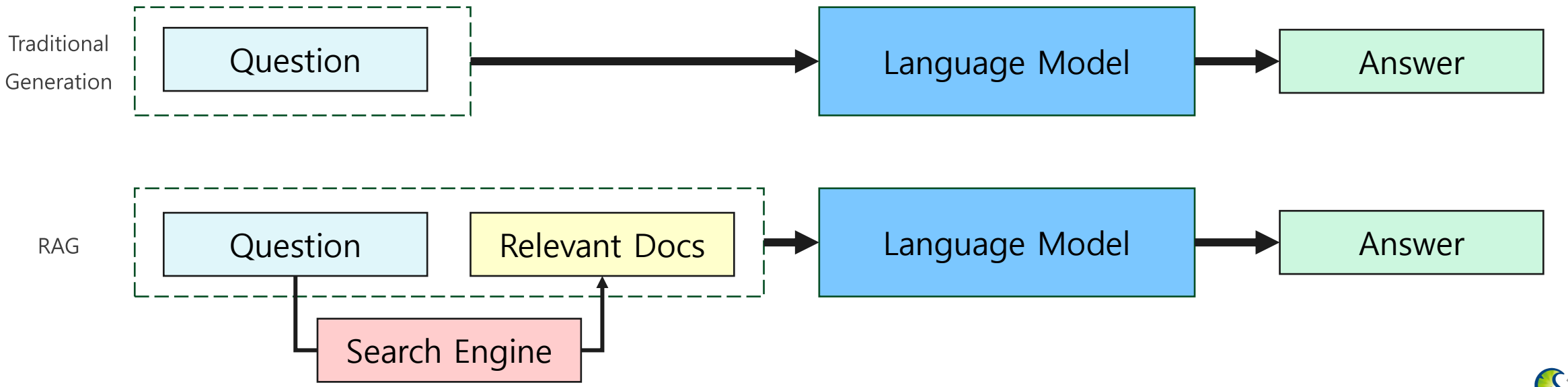
성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)

Contents

Retrieval-Augmented Generation

- **LLM의 성능 및 한계**
  - 일반적인 Task의 경우, 소수의 예시 (few-shot) 또는 예시 없이도 준수한 성능을 달성
  - Knowledge-Intensive Task에서는 모델 내부의 지식만으로 좋은 성능을 달성하기 어려움 (Hallucination 발생)

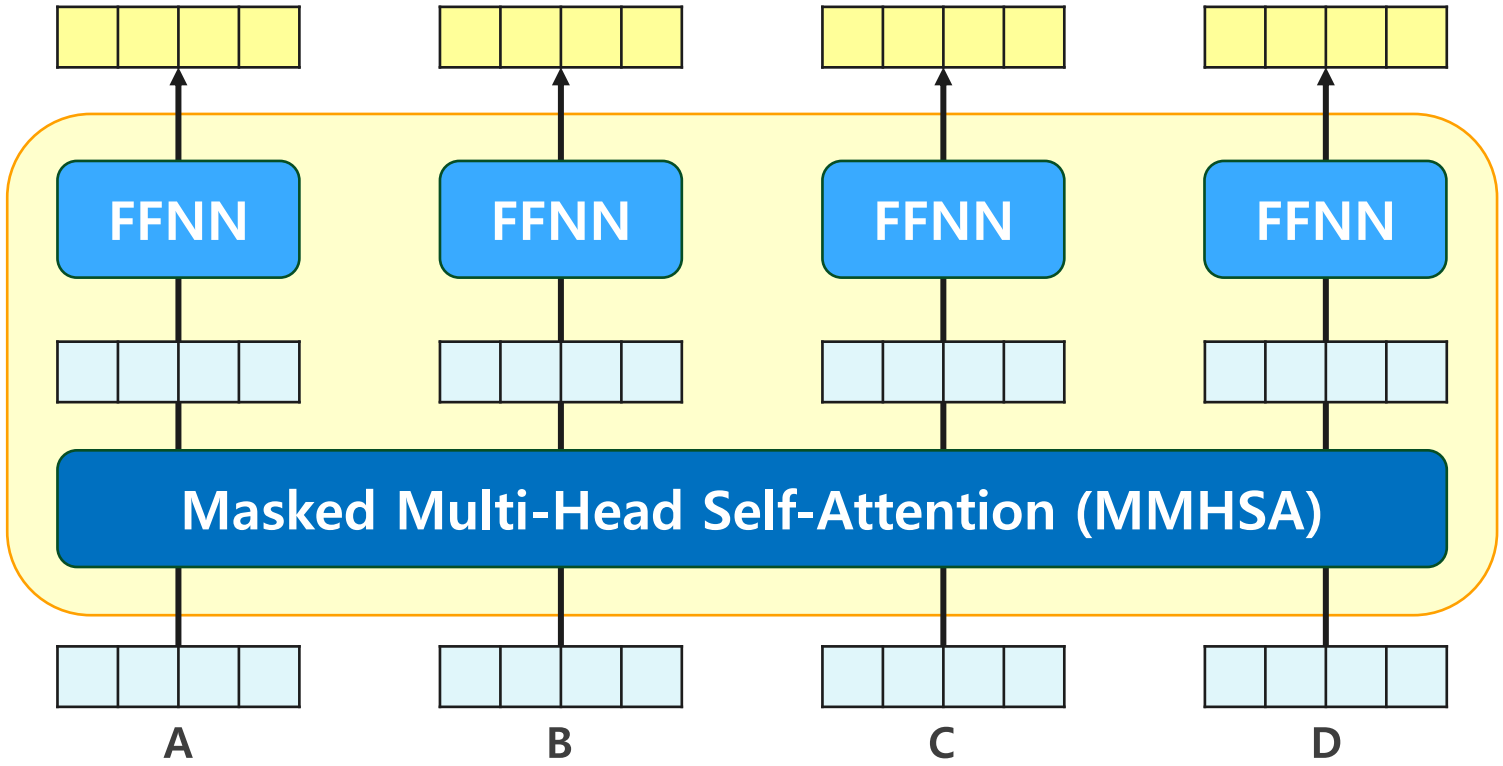
- **Retrieval-Augmented Generation (RAG)**
  - 관련 정보를 검색하여, input에 외부 지식을 추가하여 답을 생성하는 기술

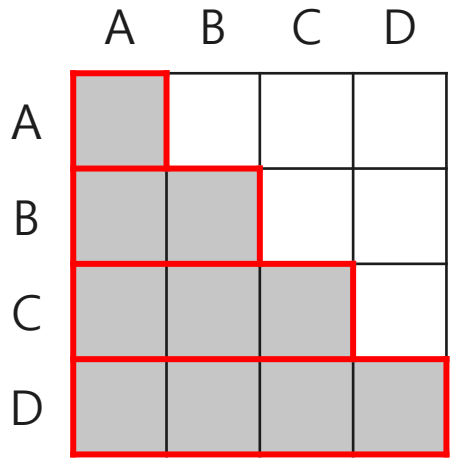


Generation Cost (Flops & Memory)

연산량 & 메모리

- 연산량: 결과를 얻기 위해 수행한 모든 연산의 횟수
- 메모리: GPU 또는 RAM에 저장된 데이터 (모델 파라미터, 연산 과정의 hidden embedding 등)





N+1번째 token 생성 시 (1-layer)

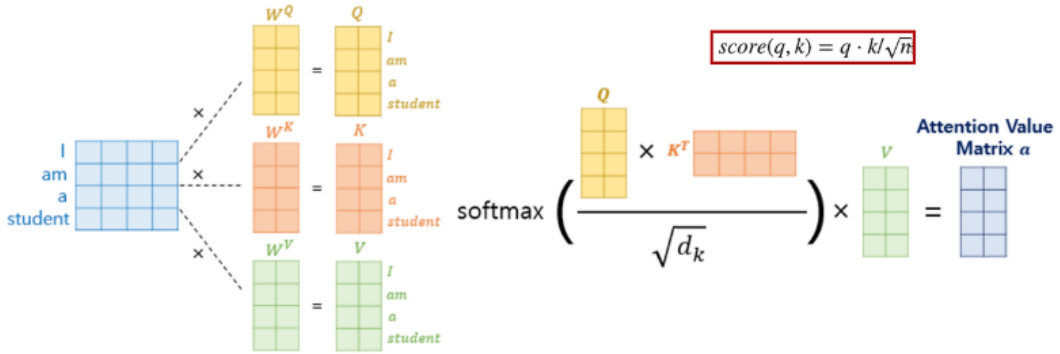
연산량: $O(N^2d + d^2)$

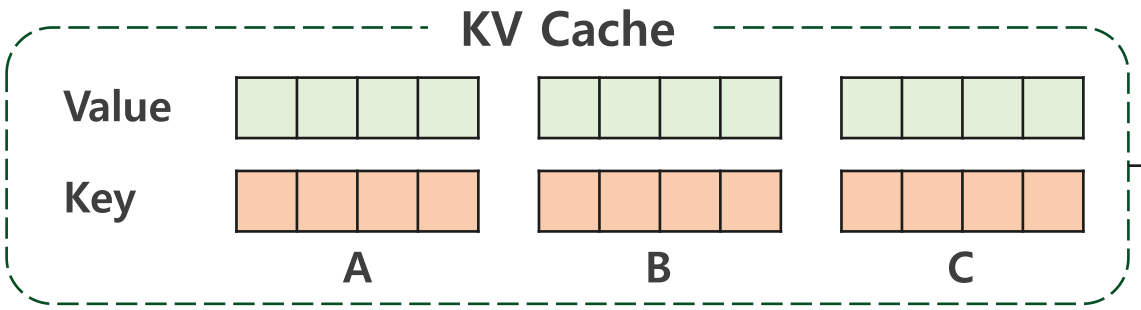
메모리: $O(N^2 + Nd)$

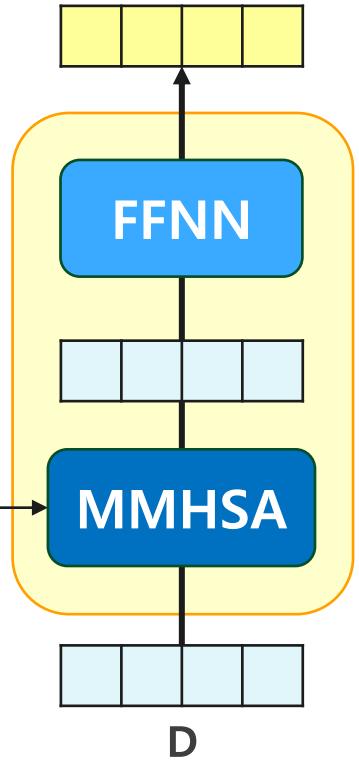
Generation Cost (Flops & Memory)

Generation Cost

- 연산량: token D에 대한 Attention (d x d + n x d) + FFNN (d x d)
- 메모리: GPU 또는 RAM에 저장된 데이터 (모델 파라미터, 연산 과정의 hidden embedding 등)



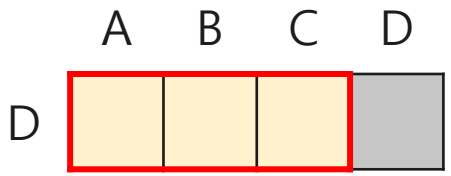




N+1번째 token 생성 시 (1-layer)

연산량: $O(Nd + d^2)$

메모리: $O(Nd)$



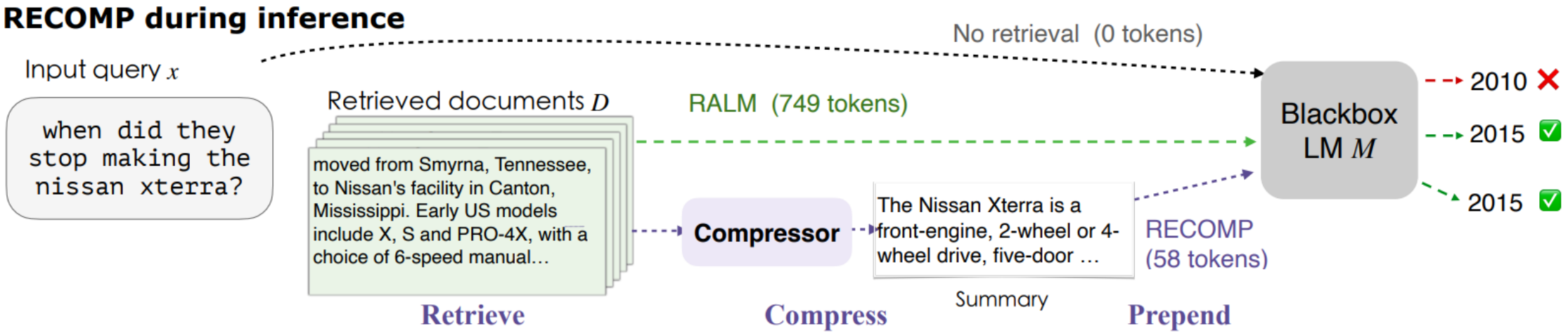
저장된 KV Cache를 활용해 계산

Context Compression in LLM

Context Compression

입력으로 사용되는 Token의 수를 줄이는 연구

- RAG 시스템은 모델의 정확도를 개선하지만, 많은 양의 Document token으로 인해서 많은 연산 필요
- 문서 내 중요한 토큰을 선별하여 최소한의 토큰을 사용하는 RAG 시스템

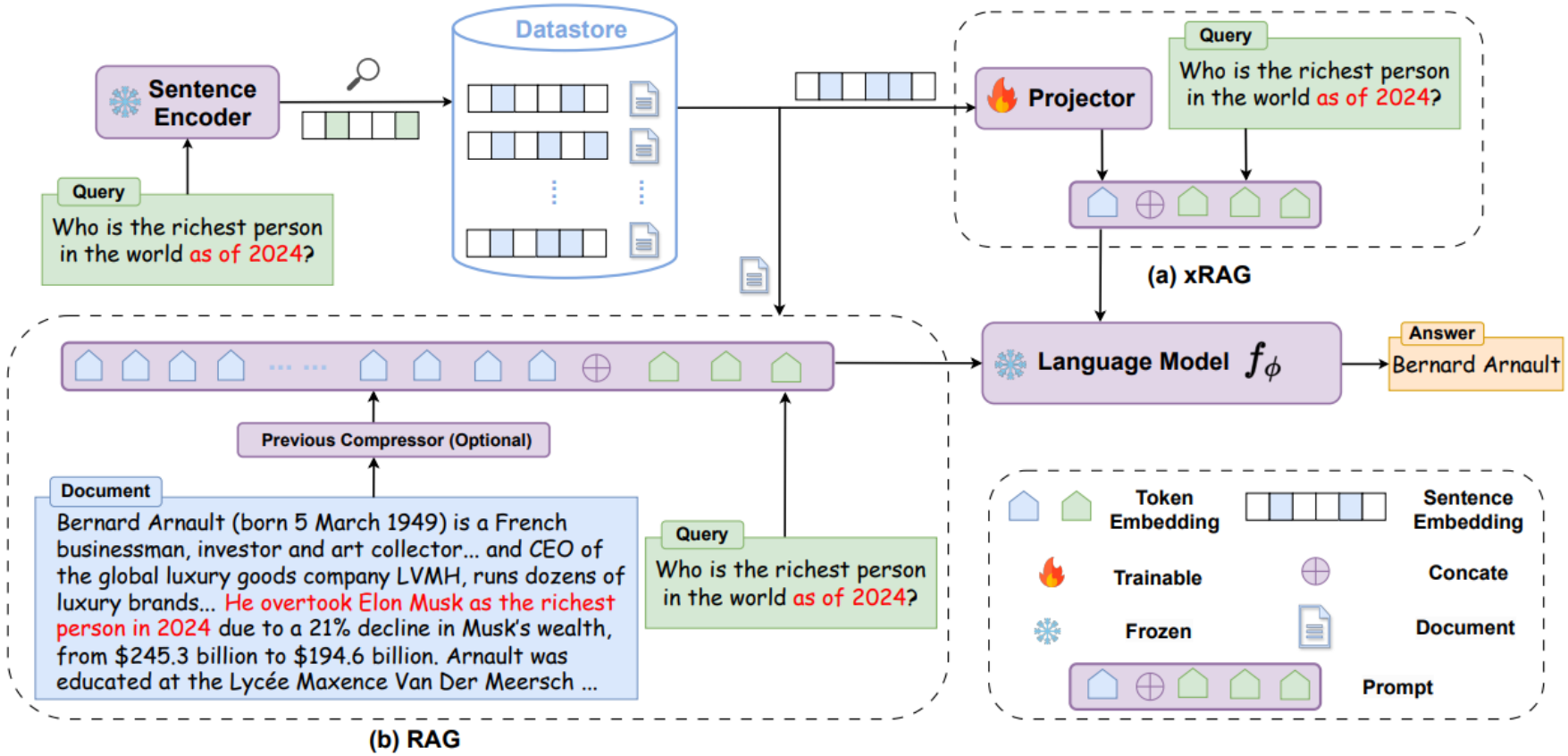


Xu, Fangyuan, Weijia Shi, and Eunsol Choi. "Recomp: Improving retrieval-augmented lms with compression and selective augmentation." *arXiv preprint arXiv:2310.04408* (2023).

xRAG

xRAG

- 문서를 하나의 Dense Vector로 압축



xRAG Structure

Structure

Language Model $\mathcal{F}_\phi$

- 응답을 생성하는 언어모델

Sentence Encoder SE_θ

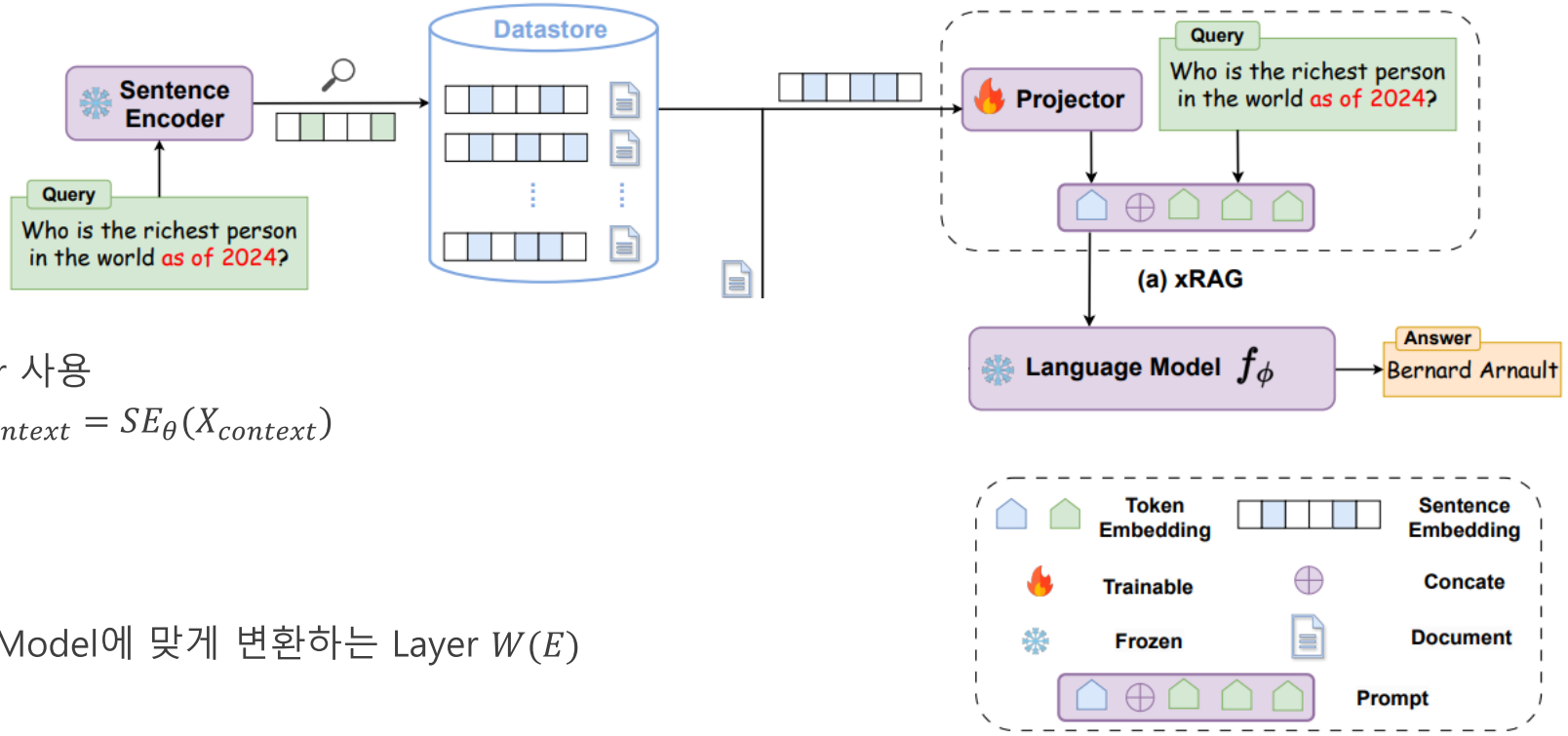
- Opensource Sentence Encoder 사용
- 문서를 하나의 벡터로 표현 $E_{context} = SE_\theta(X_{context})$

Projector W

- Two-layer MLP
- 문서 벡터를 Target Language Model에 맞게 변환하는 Layer $W(E)$
- 유일하게 학습되는 모듈

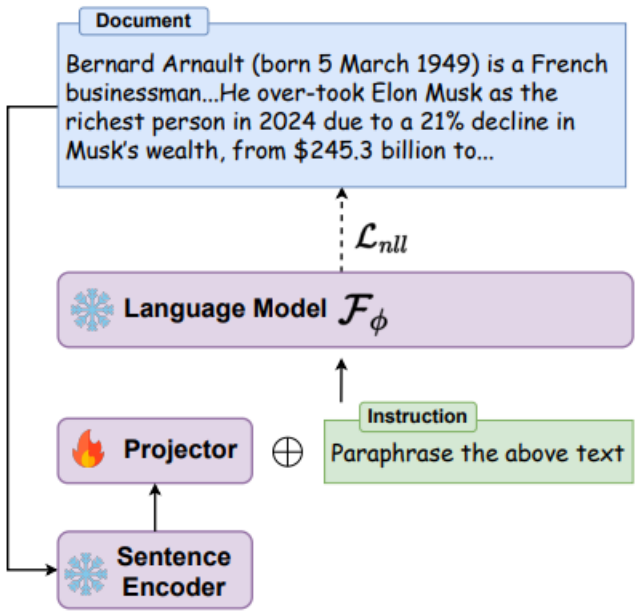
Retrieval System (Optional)

- Question에 대해서 문서를 반환하는 시스템
- 실험 환경에 따라 관련된 문서를 명시적으로 제공하는 경우도 존재

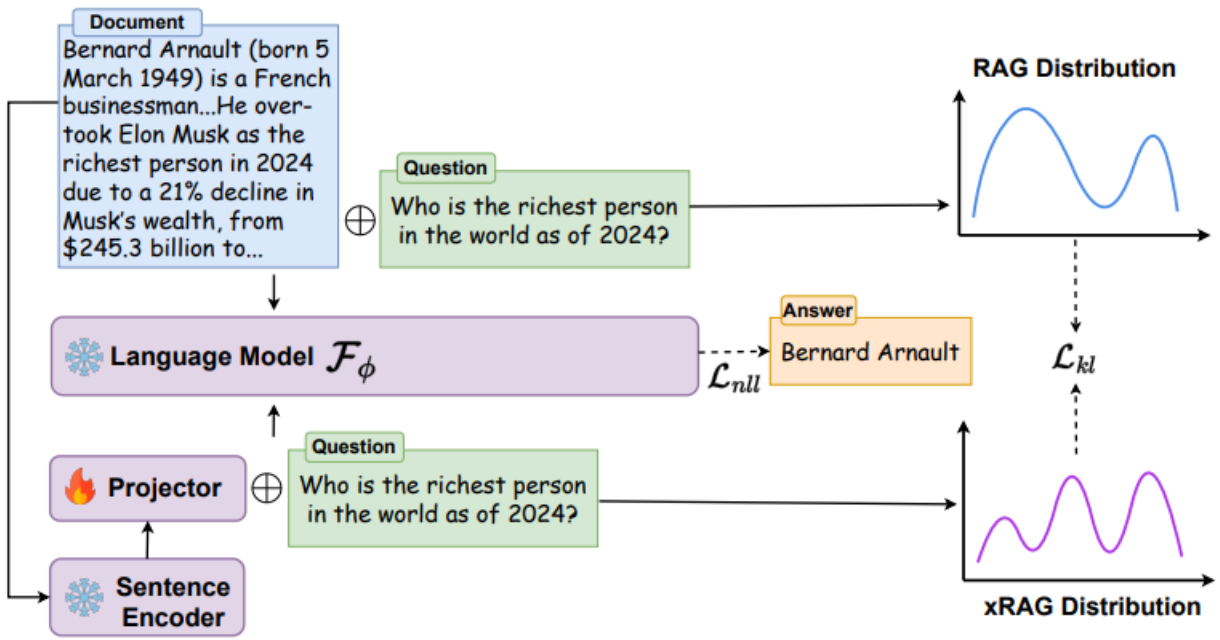


xRAG Training

- **Two-stage training strategy**
  - **Pretraining: Projector가 문서 정보를 잘 압축하도록 학습**
  - **Instruction Tuning: 주어진 Question에 대해서 응답하도록 학습**



(a) Paraphrase Pretraining



(b) Context-aware Instruction Tuning

xRAG Training

Pretraining

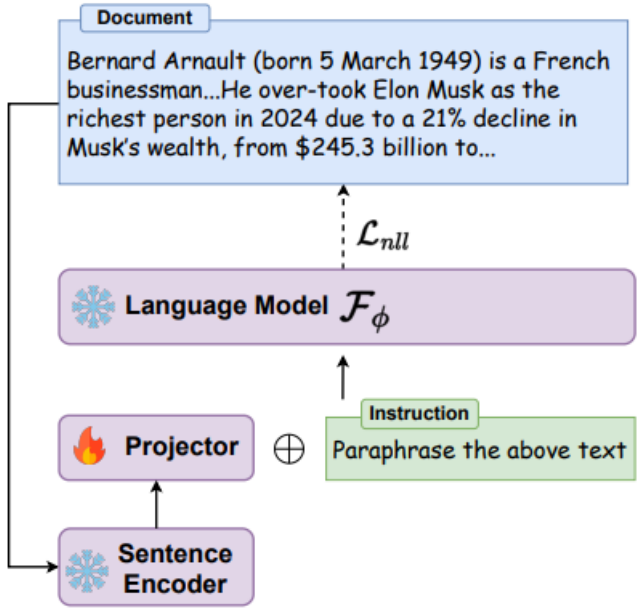
Input

- $W(E)$: 문서의 정보를 담은 벡터
- $X_{instruct}$: 앞의 정보를 보고 문서를 복원하라는 instruction

Output

- D: 원본 문서

$$\mathcal{L}_{nll} = -\sum_{i=1} \log p_\phi(d_i | W(E), X_{instruct}, d_{<i})$$



(a) Paraphrase Pretraining

Projector는 Language Model이 이해할 수 있도록 Context 벡터 $W(E)$를 생성

xRAG Training

Instruction Tuning 1 (Language Modeling)

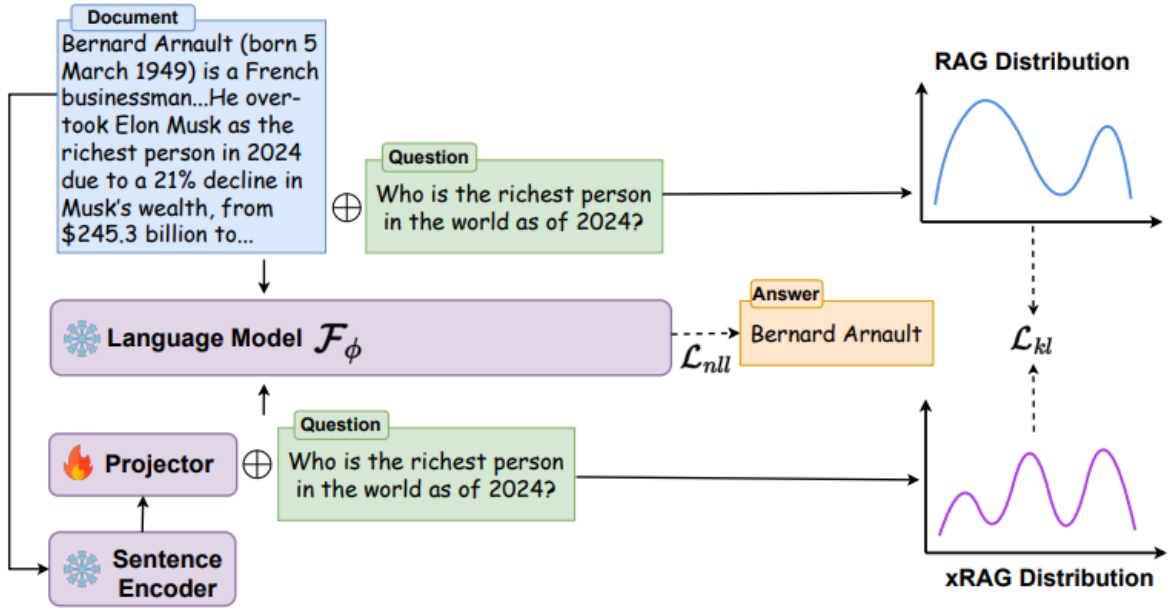
Input

- $W(E)$: 문서의 정보를 담은 벡터
- $X_{question}$: QA 데이터의 질문

Output

- X_{answer}: QA 데이터의 정답

$$\mathcal{L}_{nll} = -\sum_{i=1} \log p_\phi(X_{anwer,i}|W(E), X_{instruct}, X_{answer,<i})$$



(b) Context-aware Instruction Tuning

- $W(E)$ 정보를 활용하여, 질문에 대한 답을 생성하도록 학습

xRAG Training

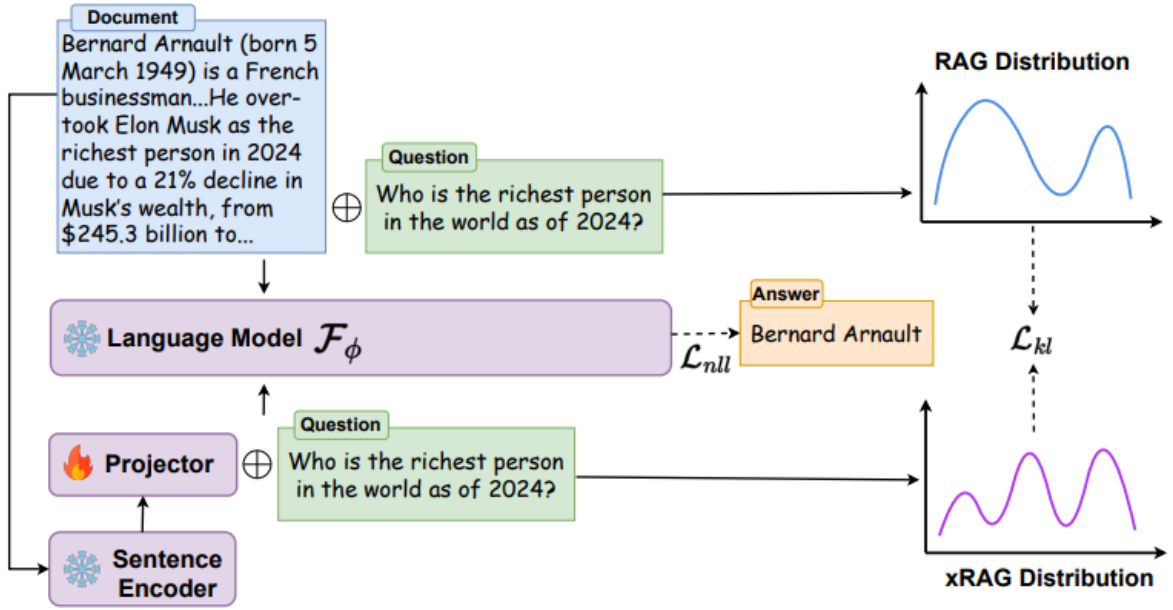
Instruction Tuning 2 (Self-Distillation)

Input

- $W(E)$: 문서의 정보를 담은 벡터
- $X_{context}$: 문서 원본
- $X_{question}$: QA 데이터의 질문

Output

- X_{answer}: QA 데이터의 정답



(b) Context-aware Instruction Tuning

$$\mathcal{L}_{kl} = D_{KL}(p_\phi(X_{answer}|X_{question}, X_{context})||p_\phi(X_{anwer}|X_{question}, W(E))$$

- $W(E)$ 정보를 활용한 확률 분포와 원본 문서를 활용한 확률 분포가 동일해지도록 학습

Experiments Setting

- **Dataset**
  - **Open-Domain QA**
    - NQ, TriviaQAm, WebQuestions
  - **Multi-hop QA**
    - HotpotQA
  - **Long-form QA**
    - TruthfulQA
  - **Fact-checking**
    - FactKG

- **Evaluation metric**
  - **EM (Exact Match), Accuracy (classification), F1 score, Rouge-L**

Main Results

Knowledge Intensive Tasks

Task Type	NQ	TriviaQA	WebQA	HotpotQA	TrutufulQA		FactKG	Average	# Doc Length
	Open-Domain QA (EM)			Multihop QA (EM)	Long-form QA (F1/R-L)		Fact Checking (Acc)		
Mistral-7b									
w/o retrieval	30.25	57.08	34.89	27.02	26.23	25.51	54.78	36.54 (0.0%)	0
w retrieval	**42.71**	**65.88**	37.84	**38.79**	26.50	25.92	**67.76**	**43.63** (19.4%)	175.1
***with Compression**									
LLMLingua †	30.64	57.94	32.63	29.91	25.70	25.10	64.17	38.01 (4.0%)	98.6
LLMLingua ‡	28.81	57.09	32.33	29.13	26.10	25.39	63.57	37.48 (2.5%)	61.1
TF-IDF	30.25	58.49	35.43	26.62	26.33	25.83	59.56	37.49 (2.6%)	1
xRAG	39.10	65.77	**39.40**	34.05	**28.10**	**27.71**	63.08	42.46 (16.2%)	1
Mixtral-8x7b									
w/o retrieval	41.99	71.10	40.31	32.87	25.60	24.90	62.64	42.76 (0.0%)	0
w retrieval	45.15	70.34	41.26	**43.46**	27.10	25.80	**70.42**	46.22 (8.0%)	175.1
***with Compression**									
LLMLingua†	37.65	67.70	36.02	35.66	25.99	25.39	67.98	42.32 (-1.0%)	96.6
LLMLingua‡	37.81	67.81	35.78	35.27	25.68	25.00	68.03	44.17 (-1.3%)	61.1
TF-IDF	41.19	69.94	41.63	32.05	26.80	26.00	66.17	43.41 (1.4%)	1
xRAG	**47.28**	**74.14**	**44.50**	39.66	**27.80**	**26.64**	68.20	**46.91** (9.7%)	1

Computational Efficiency

- **CUDA Time & GFLOPs**
  - CUDA Time: GPU를 사용하는 물리적 시간
  - FLOPs: 연산량 (FLOPs, GFLOPs, TFLOPs, ...)

Table 2: Comparison of RAG and xRAG performance in CUDA Time and GFLOPS.

	CUDA Time (ms)			GFLOPs		
	RAG	**xRAG**	**Improvement**	**RAG**	**xRAG**	**Improvement**
FactKG	431.5	215.6	x2.01	4683.8	1289.5	x3.63
NQ	918.7	611.3	x1.51	1338.6	384.0	x3.48
TriviaQA	807.1	512.1	x1.57	1667.2	492.3	x3.38
WebQA	872.6	577.3	x1.51	1405.1	386.8	x3.63
Average			**x1.64**			**x3.53**

RAG effectiveness

Resilience & Boost

- Resilience: **외부 문서 없이 맞춘 문제**를 외부 문서가 주어졌을 때 여전히 맞추는가?
- Boost: **외부 문서 없이 틀린 문제**를 외부 문서가 주어졌을 때 맞출 수 있는가?

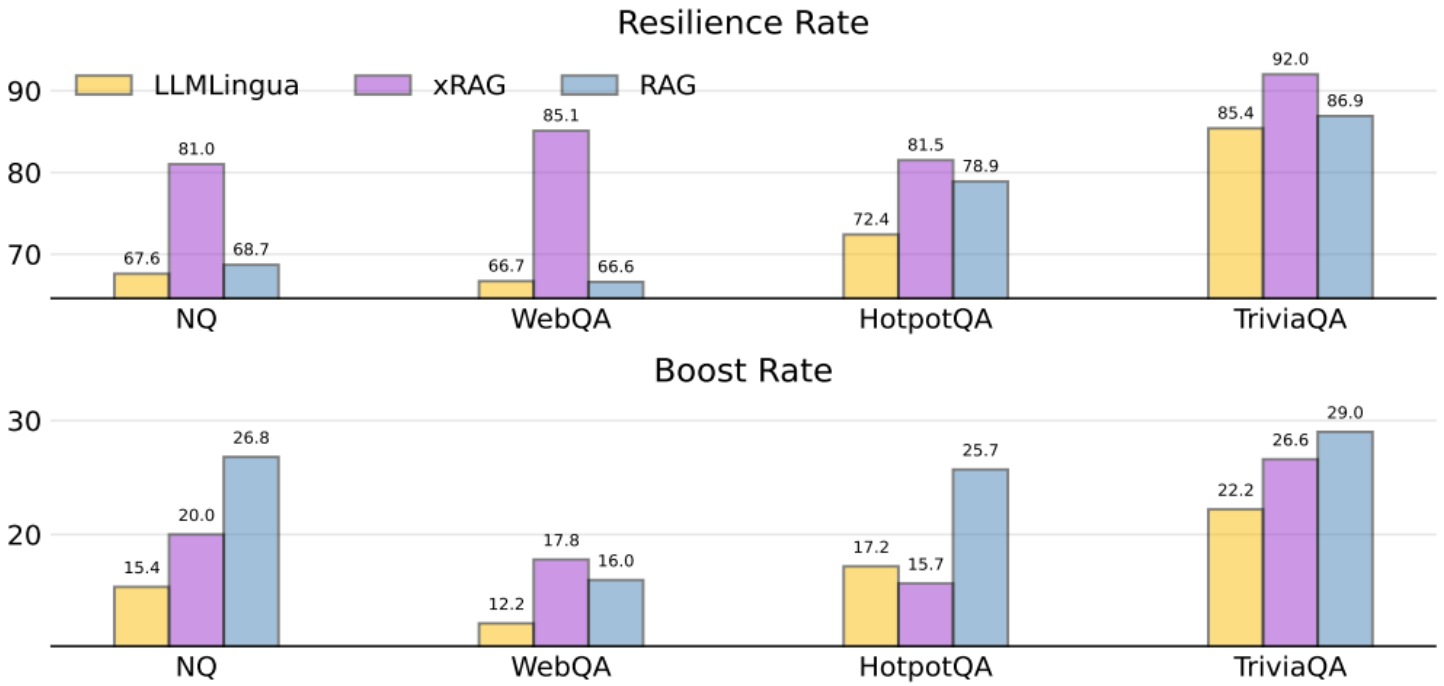


Figure 4: Resilience rate and boost rate of three augmentation methods: LLMLinuga, xRAG and RAG over a Mixtral-8x7b baseline without retrieval augmentation.

Ablation Study

Ablation Study

Table 3: Ablation on different training strategy for xRAG.

	NQ	TriviaQA	WebQA	HotpotQA	Averaged Performance	Resilience	Boost
Mistral-7b							
xRAG	39.10	65.77	39.40	34.05	**44.58**	82.3%	22.2%
w/o finetune	30.14	59.48	35.19	26.70	37.87	66.6%	20.8%
w/o pretrain	31.25	59.07	41.19	24.32	38.95	79.8%	14.1%
w/o nll	35.46	65.27	39.57	31.80	43.02	83.7%	19.4%
w/o self-kd	34.99	64.33	39.22	27.45	41.49	76.2%	20.8%
w LoRA	35.71	60.14	40.45	22.91	39.80	76.0%	18.0%
Mixtral-8x7b							
xRAG	47.48	74.14	44.50	39.66	**51.45**	84.9%	20.0%
w/o finetune	34.46	64.08	34.89	30.43	40.96	65.9%	17.8%
w/o pretrain	42.54	71.17	47.44	31.23	48.09	85.0%	14.2%
w/o nll	45.10	72.85	45.03	37.11	50.02	84.8%	18.9%
w/o self-kd	42.38	72.26	44.73	32.41	47.94	79.8%	18.9%

Case Study

Case Study

RAG

Refer to the background document and answer the question:

Background: The Old Man and the Key | " several ""washed-up"" celebrities, was composed by Alf Clausen and written by Vitti. Even though Vitti received sole credit for writing the song, parts of the lyrics were written by Simpsons writer Carolyn Omine, which were then revised by fellow writer Matt Selman. Selman was especially satisfied with the rhyme sung by Mr. T; he stated that whenever he feels down, he thinks of that rhyme and that it "boosts [him] up inside." The episode features American actress Olympia Dukakis as Grampa's love interest Zelda. American comedian Bill Saluga also makes an appearance, as his television character Ray J. Johnson. The owner of the minivan is portrayed by series regular Hank Azaria, who imitated the voice of Clark Gable for the character." ✗

Question: Which US comedian/actor is credited with saying 'A man is only as old as the woman he feels'?

Output: Bill Saluga. ✗

w/o Retrieval

Answer the question:

Question: Which US comedian/actor is credited with saying 'A man is only as old as the woman he feels'?

Output: Groucho Marx. ✓

xRAG

Refer to the background document and answer the question:

Background: [X]

Question: Which US comedian/actor is credited with saying 'A man is only as old as the woman he feels'?

Output: Groucho Marx. ✓

Figure 5: Given the misleading document, RAG model tend to generate a wrong answer based on the document, while xRAG demonstrate its robustness by leveraging the internal knowledge of LLM.

Conclusion

- **RAG에 특화된 Context 압축 기법인 xRAG를 제안**
  - 단일 토큰으로 context를 압축하여 매우 효율적이면서도 동시에 높은 성능을 달성
- **효율적인 시스템 구조**
  - Offline-constructed 문서 임베딩을 활용할 수 있다는 점에서 효율적인 구조 (정보검색 시스템의 임베딩 등)
- **Limitation**
  - 여러 문서를 활용하는 RAG 시스템에 적용이 어려움

Thank you

Dooyoung Kim

(kdysunleo98@gmail.com)

질문 https://forms.gle/LtvyMJ7BFwMKpWtz8 **피드백** https://forms.gle/PAmxLQnRBZVhAMaw8

응답 https://docs.google.com/spreadsheets/d/1uWyc0pUfQOwTTZUDyY3gxlmU5xcFro90kKJKOcDitnk/edit?usp=sharing