Knowledge Conflict

24.01.11

손준영

Knowledge Conflict

최근 연구에서 Knowledge Conflict 를 어떻게 정의하고 있는지

LLM's Retrieval Capabilities 는 어떤 방향으로 연구가 진행되고 있는지

Knowledge Editing 과정에서의 Knowledge Conflict 평가

Unveiling the Pitfalls of Knowledge Editing for Large Language Models

https://openreview.net/forum?id=fNktD3ib16

ICLR2024 제출 논문
8/8/6/6

Retrieval-Augmented Lightweight Tuning 방법론 제안

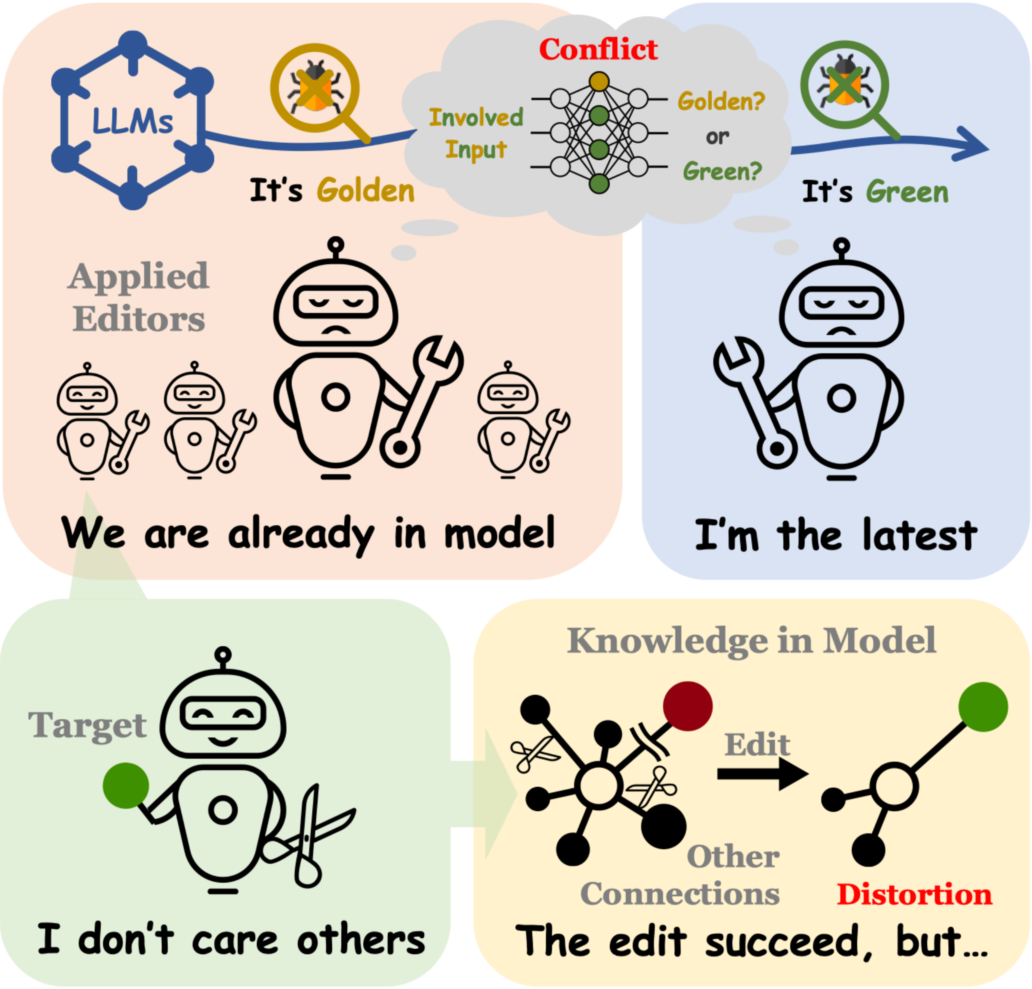
RA-DIT: Retrieval-Augmented Dual Instruction Tuning

https://openreview.net/forum?id=22OTbutug9

ICLR2024 제출 논문
8/6/6/5

Introduction

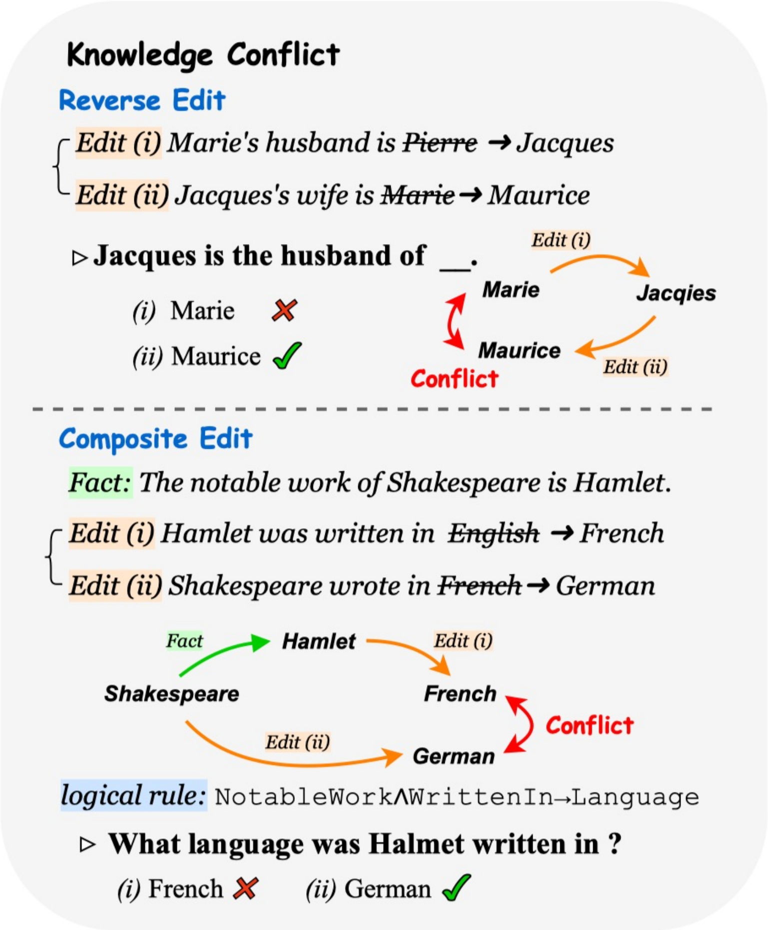
Knowledge Conflict and Distortion



Knowledge Conflict

Knowledge Conflict를 평가하려면 어떻게 해야 할까?

➔ **Knowledge Conflict를 발생시키는 Knowledge Editing 상황을 simulate**



$$\text{Reverse Edit}: \begin{cases} e_1 = (s_1, r_1, o_1 \rightarrow o_2) \\ e_2 = (o_2, r_2, s_1 \rightarrow s_2) \end{cases}$$

Edit 1

Edit 2

$$\begin{cases} k_o = (s_1, r_1, o_2) \\ k_n = (s_2, r_1, o_2) \end{cases}$$

$$\text{Composite Edit}: \begin{cases} k_f = (s_1, r, s_2) \\ e_1 = (s_1, r_1, o_1 \rightarrow o_2) \\ e_2 = (s_2, r_2, o_2 \rightarrow o_3) \end{cases}$$

Preserving tired fact

Edit 1

Edit 2

$$\begin{cases} k_o = (s_1, r_1, o_2) \\ k_n = (s_1, r_1, o_3) \end{cases}$$

Knowledge Conflict

Knowledge Conflict를 가장 빈번하게 발생시키는 2 scenarios: Reverse Edit and Composite Edit

➔ **Knowledge Conflict를 발생시키는 Knowledge Editing 상황을 simulate**

Reverse Edit

Marie's husband is Pierre → Jacques
Jacques's wife is Marie → Maurice
These two edits both modify the fact

(Marie, WifeOf, Jacques) *Edit 1*
(Jacques, HusbandOf, Maurice) *Edit 2*
(Jacques, HusbandOf, ?)

$$\begin{cases} k_o = (s_1, r_1, o_2) \\ k_n = (s_2, r_1, o_2) \end{cases}$$

(Marie, WifeOf, Jacques)
(Maurice, WifeOf, Jacques)
Conflict!

Q: ***"Jacques is the husband of who?"***
A1) Marie
A2) Maurice

Knowledge Conflict

Knowledge Conflict를 가장 빈번하게 발생시키는 2 scenarios: Reverse Edit and Composite Edit

➔ **Knowledge Conflict를 발생시키는 Knowledge Editing 상황을 simulate**

Composite Edit

The notable work of Shakespeare is Hamlet = **(Hamlet, NotableWorkOf, Shakespeare)** *Preserving tired fact*
Hamlet was written in English → French = **(Hamlet, WrittenIn, French)** *Edit 1*
Shakespeare wrote in French → German = **(Shakespeare, WrittenIn, German)** *Edit 2*
Will affect the fact (Hamlet, WrittenIn, ?)

A Logical Rule = $r \wedge r_1 \rightarrow r_2$: NotableWorkOf ∧ WrittenIn → Language

$$\begin{cases} k_o = (s_1, r_1, o_2) \\ k_n = (s_2, r_1, o_2) \end{cases}$$

(Hamlet, WrittenIn French)
(Shakespeare, WrittenIn, German) **Conflict!**

Q: ***"What language was Hamlet written in?"***
A1) French
A2) German

How do we evaluate those?

Knowledge editing method의 성능을 평가하기 위한 ConflictEdit 데이터셋 구축

How to construct?

➔ WikiData 활용

Wikidata에 정의된 Reverse relation과 composite logical rules을 활용하여 수집

Evaluation Metrics

➔ Conflict Score (CS)

How well a knowledge editing method handles the knowledge conflict issue

Knowledge editing 이후 the new fact k_n이 the old fact k_o보다
얼마나 더 높은 확률을 갖는지에 대한 비율을 계산

How do we evaluate those?

Knowledge editing method의 성능을 평가하기 위한 ConflictEdit 데이터셋 구축

Evaluation Metrics

➔ Conflict Score (CS)

How well a knowledge editing method handles the knowledge conflict issue

Knowledge editing 이후 the new fact k_n이 the old fact k_o보다
얼마나 더 높은 확률을 갖는지에 대한 비율을 계산

Dataset Split	Explicit Mode	Implicit Mode
Coverage	(s_2, r_2, o_2^*)	(s_1, r_1, o_2^*)
Reverse	(s_2, r_2, o_2^*)	(o_2^*, r_1, s_2)
Composite	(s_2, r_2, o_2^*)	(s_1, r_1, o_2^*)

Table 5: Explicit (CS_{exp}) and Implicit (CS_{imp}) of each dataset split.

Second Edit: $(s_2, r_2, o_2 \rightarrow o_2^*)$
The edit target: (s_2, r_2, o_2^*)

How do we evaluate those?

Evaluation Metrics

➔ Conflict Magnitude (CM)

To estimate the decrease of the probability of k_o

$$\mathrm{CM} = \frac{p_{f_{\theta^m}}(k_o) - p_{f_{\theta'}}(k_o)}{p_{f_{\theta^m}}(k_o)}$$

p(k): (s, r)이 프롬프트로 주어졌을 때, 타겟 오브젝트 k에 대한 확률

θ^m*: intermediate model parameters after edit* (e_1)

Experiments

Model

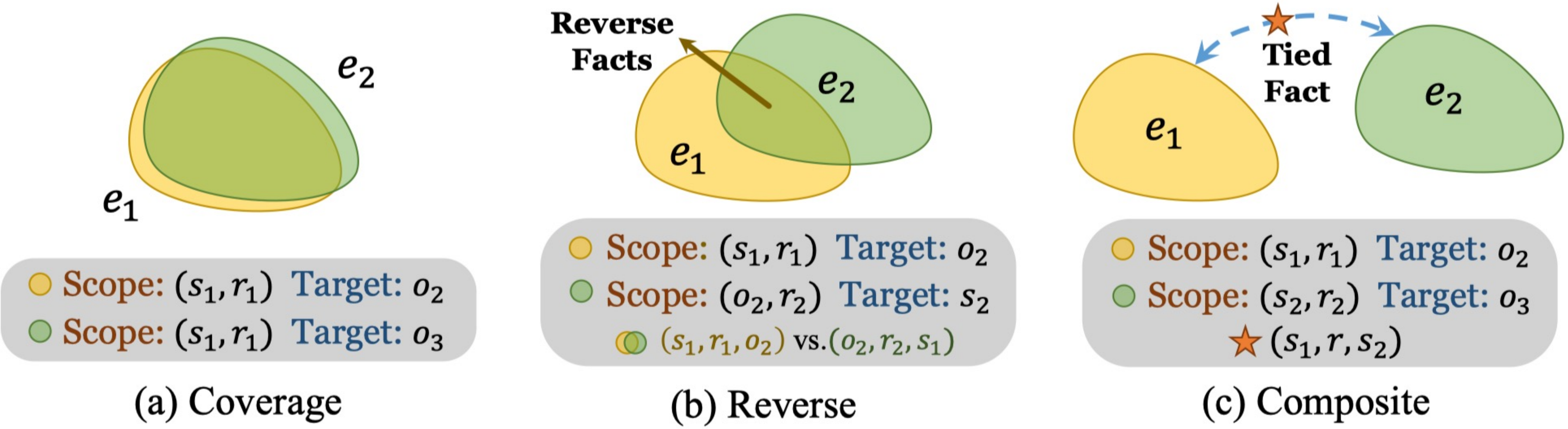
GPT2-XL (1.5B), GPT-J (6B)

Method	Single	Coverage		ConflictEdit Reverse			ConflictEdit Composite			
	Succ↑	CS↑	CM↑	CSexp↑	CSimp↑	CM↑	CSexp↑	CSimp↑	CM↑	TFD↓
GPT2-XL										
FT	82.56	78.88	70.86	80.28	15.20	**71.11**	75.45	57.65	**64.28**	88.75
MEND	98.40	91.04	60.01	**88.89**	**15.32**	60.50	**84.85**	**81.35**	43.45	72.09
ROME	99.96	**99.76**	**96.92**	65.92	0.00	-0.65	71.70	38.70	37.04	69.55
MEMIT	79.24	83.88	32.29	51.44	2.08	-1.60	57.15	29.40	-1.50	24.63
GPT-J										
FT	100.0	**100.0**	**99.90**	**99.60**	4.16	**97.20**	**96.68**	**88.92**	**88.98**	89.97
MEND	100.0	95.88	82.41	88.92	**6.40**	60.72	83.04	73.52	63.99	42.95
ROME	100.0	99.80	94.25	56.84	0.00	0.06	77.60	29.24	39.27	81.02
MEMIT	100.0	99.64	88.91	55.16	0.00	-1.18	75.48	49.28	28.78	64.51

Table 1: Knowledge Conflict results of knowledge editing methods. **Bold** results denote the best performance in each situation, whereas red results indicate a total failure under the setup and blue results mark the damage on tied fact that cannot be ignored.

Experiments

A Unified View of Knowledge Conflict



<u>이전 연구의 Mitchell et al. (2022b)의 Editing Scope 개념에 기반으로 세 가지 유형의 지식 편집 범위를 분석</u>

1. (a)에서, Coverage 세팅은 편집 범위가 완전히 겹침 ➔ 연관된 지식에 대한 포괄적 업데이트 가능
2. (b)에서, Reverse 세팅은 reverse facts를 통해 연결되지만 논리적 함의에서 완전히 겹침 ➔
3. (c)에서, Composite 세팅은 겹치지 않는 두 범위를 가지며 기존의 묶인 사실을 통해 논리적 일관성을 확보함.

➔ 이러한 차이로 인해 대상 지식의 감지 기술이 필요하며, 이는 지식 그래프의 기호 논리 규칙에 기반하여 잠재적인 지식 불일치를 피하기 위해 사용됨.

Introduction

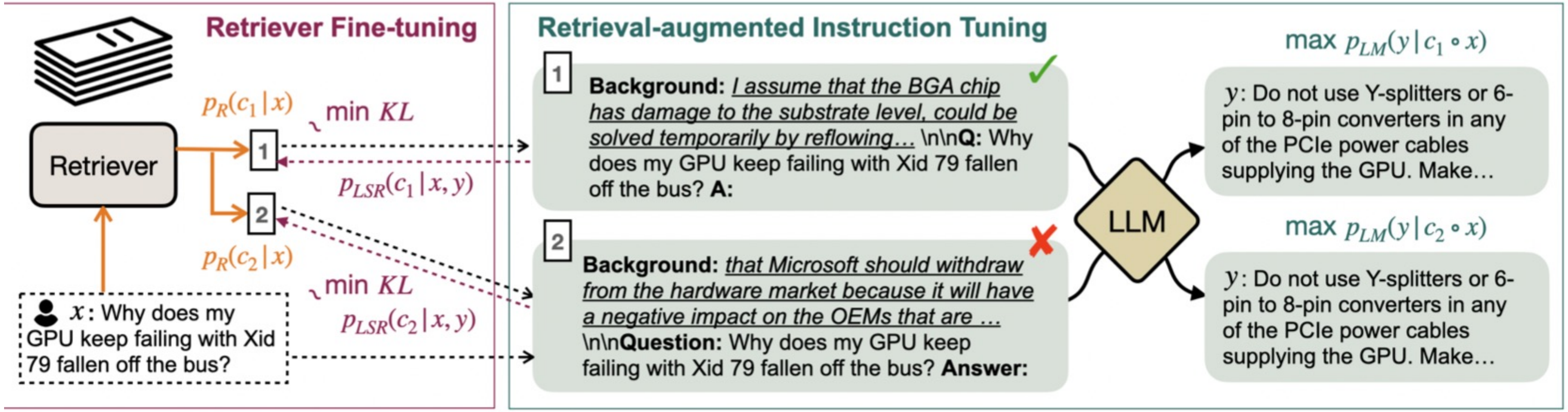
Retrieval-Augmented Language Models (RALMs)

- Retrieval 자체를 개선하여 관련성 높은 내용을 검색
- LLM의 참조 활용 능력을 개선

Introduction

RA-DIT: Retrieval-Augmented Dual Instruction Tuning

- Light-weight Instruction-Tuning에 기반한 Retrieval-Augmented Instruction Tuning 방법론 제안



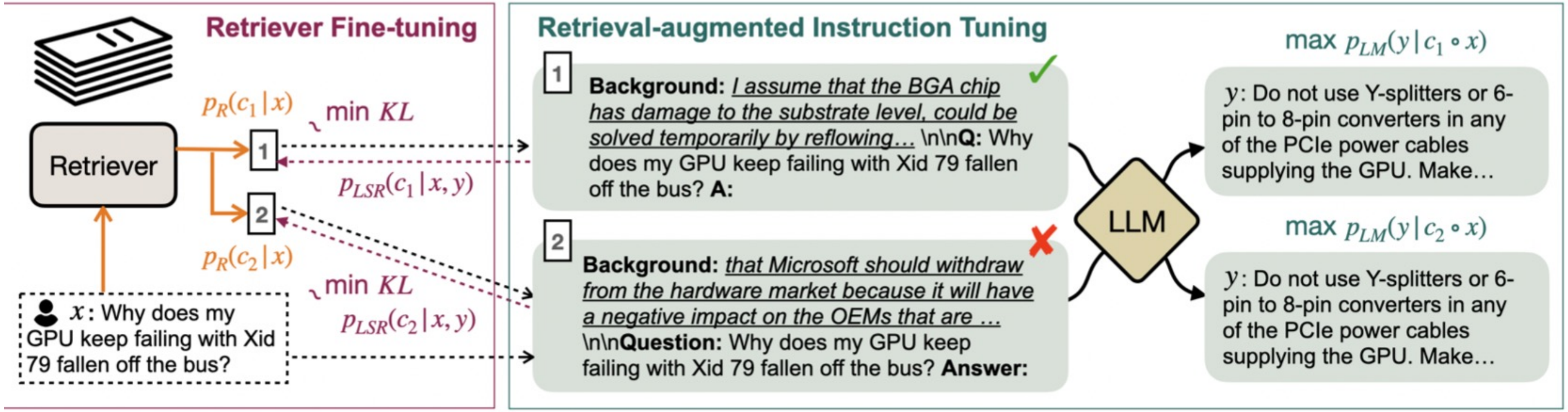
Introduction

Retrieval Augmented Language Model Fine-Tuning (LM-ft)

- Query로 검색한 Context Chunks를 x에 prepend하여 (ICL) Anger generation 학습

retrieve the top-$\tilde{k}$ relevant text chunks $\mathcal{C}_i \subset \mathcal{C}$ based on x_i.

$$\{(c_{ij} \circ x_i, y_i) | j = 1 \dots \tilde{k}\}$$



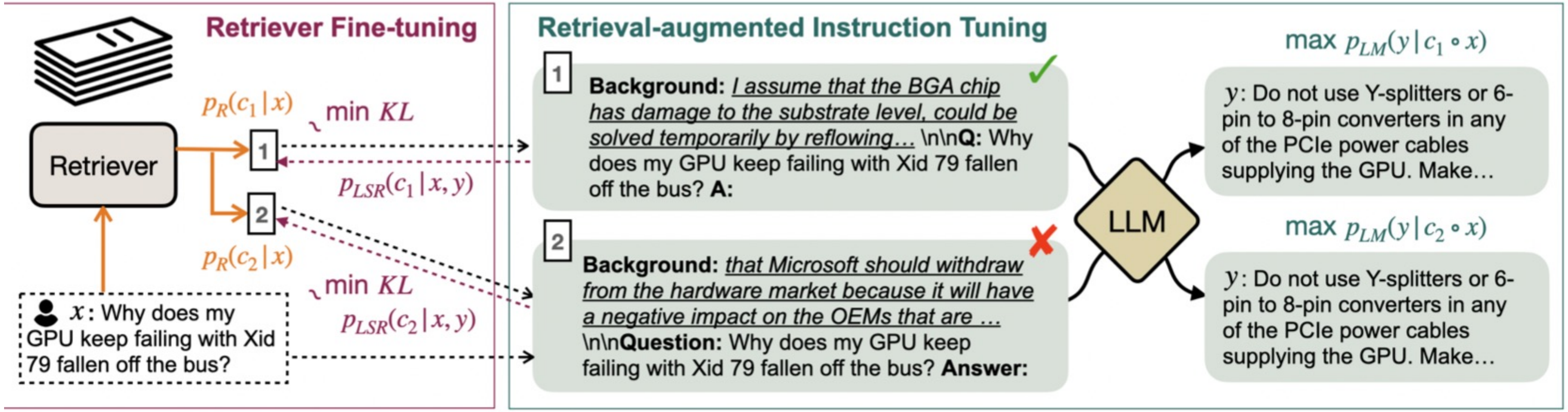
Introduction

Retriever Fine-Tuning (R-ft)

- Retriever fine-tuning을 위해서 LM을 활용하는 최근 연구 LSR (LM-Supervised Retrieval *(Shi et al., 2023)*)

$$p_{LSR}(c|x,y) = \frac{\exp\left(p_{LM}(y|c \circ x)/\tau\right)}{\sum_{c' \in \mathcal{C}} \exp\left(p_{LM}(y|c' \circ x)/\tau\right)} \approx \frac{\exp\left(p_{LM}(y|c \circ x)/\tau\right)}{\sum_{c' \in \mathcal{C}'} \exp\left(p_{LM}(y|c' \circ x)/\tau\right)}$$

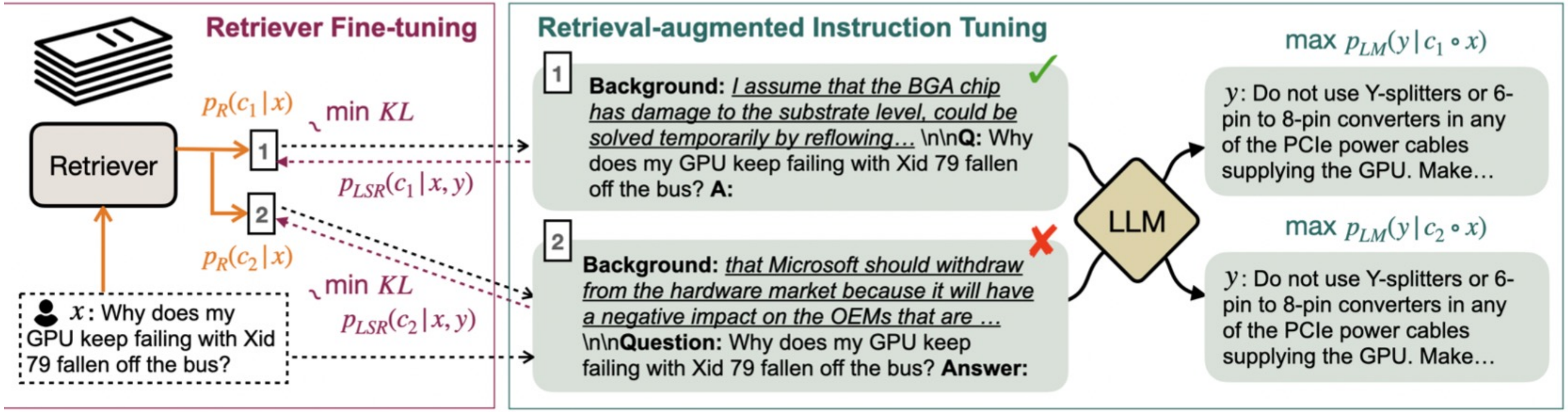
$$\mathcal{L}(\mathcal{D}_R) = \mathbb{E}_{(x,y)\in\mathcal{D}_R} KL\big(p_R(c|x) \parallel p_{LSR}(c|x,y)\big)$$



Experiments

Experimental settings

- *MMLU, NQ, TriviaQA 를 benchmark로 활용*
- *Development set으로 KILT benchmark 중 ELI5를 제외한 6개 tasks의 dev set 사용*
- *LM-ft에서는 top 1 relevant context만을 사용 + few shot examples*
- *LLM은 LLAMA활용*



Experiments

- *MMLU, NQ, TriviaQA 를 benchmark로 활용*
- *Development set으로 KILT benchmark 중 ELI5를 제외한 6개 tasks의 dev set 사용*
- *LM-ft에서는 top 1 relevant context만을 사용 + few shot examples*

	MMLU	NQ	TQA	ELI5	HoPo	FEV	AIDA	zsRE	T-REx	WoW	Avg$^\diamond$	Avg
0-shot												
LLAMA 65B	51.2	5.2	55.8	19.5	12.5	59.3	0.6	6.7	1.3	15.6	32.9	22.8
LLAMA 65B REPLUG	59.7	28.8	72.6	19.1	32.0	73.3	41.8	50.8	36.3	16.1	45.1	43.1
RA-DIT 65B	**64.6**	**35.2**	**75.4**	**21.2**	**39.7**	**80.7**	**45.1**	**73.7**	**53.1**	**16.4**	**49.1**	**50.5**
5-shot in-context												
LLAMA 65B	63.4	31.6	71.8	22.1	22.6	81.5	48.2	39.4	52.1	**17.4**	47.2	45.0
LLAMA 65B REPLUG	64.4	42.3	74.9	22.8	**41.1**	89.4	46.4	60.4	**68.9**	16.8	51.1	52.7
RA-DIT 65B	**64.9**	**43.9**	**75.1**	**23.2**	40.7	**90.7**	**55.8**	**72.4**	68.4	17.3	**51.8**	**55.2**

Jointly pre-training
The LM and the R

64-shot fine-tuned	NQ	TQA	HoPo	FEV	AIDA	zsRE	T-REx	WoW	Avg
ATLAS†	42.4	**74.5**	34.7	**87.1**	66.5	74.9	58.9	15.5	56.8
RA-DIT 65B	**43.5**	72.8	**36.6**	86.9	**80.5**	**78.1**	**72.8**	**15.7**	**60.9**

Experiments

제안하는 학습 방법론이 Retrieval Augmented Generation은 개선하겠지만,
LLM's reasoning / parametric knowledge를 손상시킬 수도 있지 않을까?

- *Retrieval Augmentation 없이 Commonsense reasoning tasks 수행*

Table 3: Performance on commonsense reasoning tasks (dev sets) without retrieval augmentation.

0-shot	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-E	ARC-C	OBQA	Avg
LLAMA 65B	85.3	82.8	52.3	84.2	77.0	78.9	56.0	**60.2**	72.1
RA-DIT 65B	**86.7**	**83.7**	**57.9**	**85.1**	**79.8**	**83.7**	**60.5**	58.8	**74.5**

감사합니다

Q&A

HOW IN-CONTEXT DOCUMENTS IMPACT SURFACE ANSWER STATISTICS

Table 1: Generated answer statistics. We present mean values along with two standard deviations in its subscript: one computed over three answers generated for the same example, one over answers for different examples. Human and WebGPT answer outputs are taken from Nakano et al. (2021), and we generate the rest. We **boldface** six rows where we collect human annotations for attribution. Numbers in red and blue indicate decrease and increase from the base model respectively.

Model (+ evidence)	# Sentences	# Words	RankGen (↑)	Self-BLEU (↓)	Perplexity (↓)
WebGPT(+ WebGPT docs)	$6.7_{-/1.9}$	$160_{-/33}$	$11.35_{-/1.98}$	$0.58_{-/0.07}$	$13.81_{-/4.86}$
GPT-3	$9.3_{1.5/2.6}$	$219_{30/51}$	$12.77_{0.67/1.87}$	$0.71_{0.04/0.06}$	$6.13_{0.02/1.37}$
+Human docs	$6.6_{0.9/1.8}$	$172_{18/40}$	$11.89_{0.60/1.86}$	$0.62_{0.04/0.07}$	$10.94_{0.05/3.94}$
+WebGPT docs	$6.8_{0.9/1.8}$	$185_{20/41}$	$11.97_{0.60/1.79}$	$0.62_{0.04/0.07}$	$11.63_{0.13/4.16}$
+Bing docs	$6.9_{1.0/1.9}$	$179_{19/38}$	$12.13_{0.68/1.91}$	$0.64_{0.04/0.07}$	$9.03_{0.12/3.24}$
+Random docs	$7.6_{1.1/2.1}$	$183_{19/39}$	$12.40_{0.67/2.13}$	$0.68_{0.04/0.07}$	$6.76_{0.05/1.86}$
Alpaca-7b	$5.0_{1.8/8.1}$	$113_{33/73}$	$12.17_{0.72/2.00}$	$0.51_{0.09/0.15}$	$11.95_{0.02/7.18}$
+Human docs	$5.7_{1.9/3.6}$	$138_{44/79}$	$11.82_{0.88/2.32}$	$0.55_{0.09/0.14}$	$12.99_{0.20/5.73}$
+WebGPT docs	$6.2_{2.3/7.9}$	$145_{45/80}$	$11.91_{0.75/2.07}$	$0.55_{0.08/0.14}$	$13.27_{0.13/5.68}$
+Bing docs	$7.6_{2.8/5.0}$	$187_{66/107}$	$12.04_{0.78/2.05}$	$0.59_{0.08/0.14}$	$10.81_{0.13/5.34}$
+Random docs	$5.2_{1.6/5.3}$	$121_{32/65}$	$12.25_{0.71/1.99}$	$0.53_{0.08/0.14}$	$11.92_{0.23/5.35}$
Human(+ Human docs)	$5.1_{-/2.7}$	$119_{-/59}$	$9.29_{-/4.37}$	$0.49_{-/0.17}$	$17.63_{-/7.53}$

Table 3: List of attribution error type (and their frequency of occurrence in unsupported sentences) and example instance.

Retrieval Failure (54%): retrieved document set does not contain answer to the question.	**Question**: Why does it seem like when I watch something the second time around, it goes by faster than the first time I watched it? **Documents**: ... Basically, the busier you are during a time interval, the faster that time interval will feel like it passed. ... (more about time goes by faster when you are not bored...) **Answer Sentence**: However, when we watch something for the second time, our brains have had a chance to process the information and are able to make more efficient use of the information. **Explanation**:The documents explain why time goes by faster when you are having fun, but the question is asking watching something the second time.
Hallucinated Facts (72%): contents that are never mentioned in the documents.	**Question**: How does money go from my pocket, through the stock market, and to support the business I've bought stock from? **Documents**: Stocks, or shares of a company, represent ownership equity in the firm, which give shareholders voting rights as well as a residual claim on corporate earnings in the form of capital gains and dividends. ... (more about how stock market works) **Answer Sentence**: You can purchase shares of the stock from a broker or through an online trading platform. **Explanation**: The documents never mention where you can buy stock from.
Incorrect Synthesis (14%) synthesizes the content from separate documents incorrectly.	**Question**: Seismologists: How do you determine whether an earthquake was naturally occurring or if it was human induced? **Documents**: Studies of the numerous nuclear tests that took place during the Cold War show that explosions generate larger P waves than S waves when compared with earthquakes. Explosions also generate proportionally smaller Surface waves than P waves. **Answer Sentence**: Natural earthquakes generate larger P waves and smaller Surface waves compared to nuclear tests. **Explanation**: Explosion generate larger P waves, not natural earthquakes. The answer sentence is thus incorrect. Most of it is copied from the documents.

KOREA UNIVERSITY