NLP&AI 연구실 세미나 (08/08, Thu)

Hallucination Mitigation wrt RAG

김진성

Natural Language Processing & Artificial Intelligence

고려대학교 KOREA UNIVERSITY

Hallucination Mitigation

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models

S.M Towhidul Islam Tonmoy[1], S M Mehedi Zaman[1], Vinija Jain[3,4]*, Anku Rani[2], Vipula Rawte[2], Aman Chadha[3,4]*, Amitava Das[2]

[1]Islamic University of Technology, Bangladesh
[2]AI Institute, University of South Carolina, USA
[3]Stanford University, USA, [4]Amazon AI, USA

RARR: Researching and Revising What Language Models Say, Using Language Models

Luyu Gao[1◇*] Zhuyun Dai[2*] Panupong Pasupat[2*] Anthony Chen[3◇*] Arun Tejasvi Chaganty[2*] Yicheng Fan[2*] Vincent Y. Zhao[2] Ni Lao[2] Hongrae Lee[2] Da-Cheng Juan[2] Kelvin Guu[2*]

[1]Carnegie Mellon University, [2]Google Research, [3]UC Irvine

luyug@cs.cmu

{zhuyundai,ppasupat,arunchaganty,y

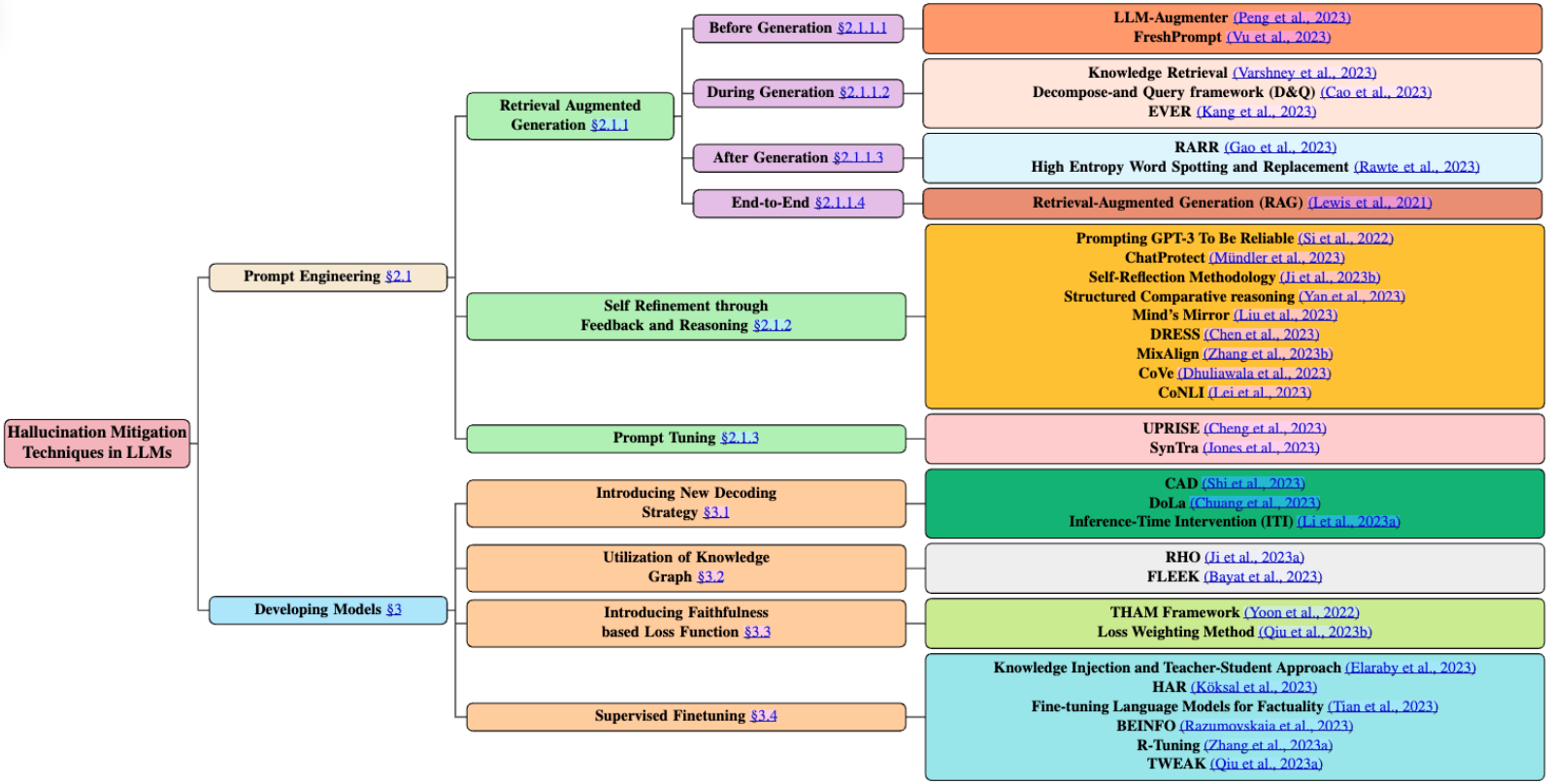
R-Tuning: Instructing Large Language Models to Say 'I Don't Know'

Hanning Zhang♠*, Shizhe Diao♠*, Yong Lin♠*, Yi R. Fung♡, Qing Lian♠, Xingyao Wang♡, Yangyi Chen♡, Heng Ji♡, Tong Zhang♡

♠The Hong Kong University of Science and Technology
♡University of Illinois Urbana-Champaign

Hallucination Mitigation – 범주

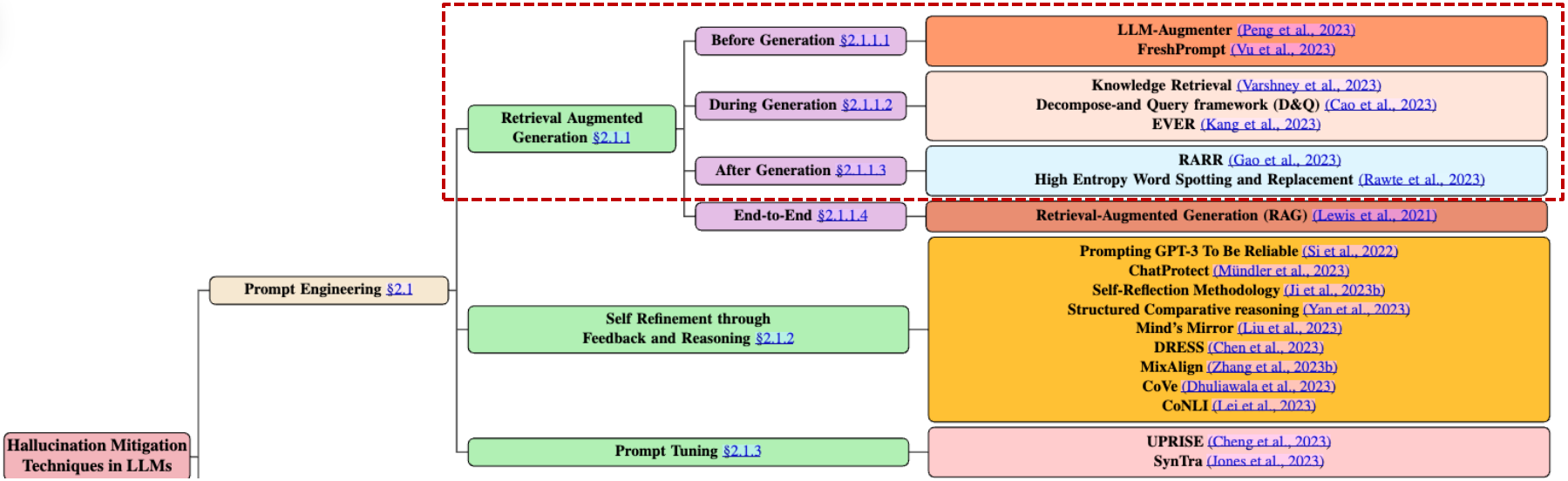
A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024)



Hallucination Mitigation - 범주

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024)

Q1) Tuning 없이 Prompt engineering (RAG 방식)으로 진짜 완화 되는게 맞나? (Halluci. 관점)
Q2) RAG 범주.. Retrieval 을 중요하게 다루는가? 주로 어떤 역할? (Ret. 관점)



Hallucination Mitigation - RAG

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024)

*** 전/중/후 - Before Generation**

- 공통 개념: LLM에 feed 하는 input (prompt)을 retrieval 을 통해 양질화 (→ 전통적)

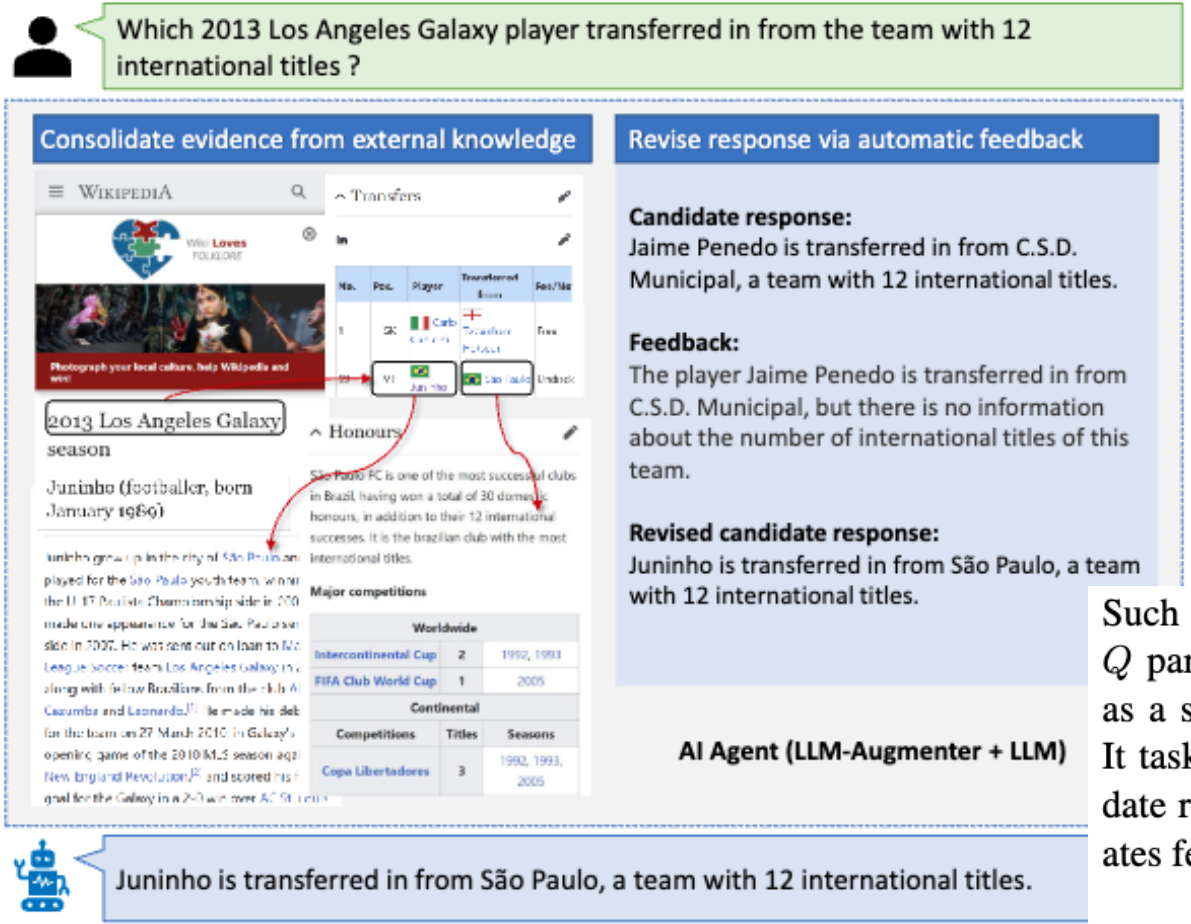
1) Feedback 모듈 활용 [1]

- "Automated feedback 생성 모듈을 통해 input prompts 를 iteratively revise 한다."
- 생성용 LLM 따로, revision 용 feedback 모듈 (set) 따로.
- 생성 LLM : response candidates 생성 (e.g., ChatGPT)
- revision 모듈: query 로 외부 지식 검색하여 evidence chains 를 생성 (knowledge consolidation)
  → utility score 등을 구하고, feedback 생성 (utility 모듈) → 피드백 기반 revised 된 응답 재생성.

2) Search Engine 활용 [2]

- 대부분의 LLMs 지식의 static 한 특성 지적
  → evolving world 에 adapt 하는 능력을 위한 Few-shot prompting 제안!
- up-to-date 정보를 프롬프트에 통합하기 위하여 search engine 활용

[1] Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback (https://arxiv.org/pdf/2302.12813)
[2] FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation (https://arxiv.org/pdf/2310.03214)



uage Models (2024)

양질화 (→ 전통적)

ratively revise 한다."

를 생성 (knowledge consolidation)
vised 된 응답 재생성.

Such a utility function is a text generation model Q parameterized by ψ, and can be implemented as a seq2seq or auto-regression language model. It tasks as input user query q, evidence e, candidate response o and dialog history h_q, and generates feedback in text f as

$$f = Q_\psi(q, e, o, h_q) \quad (2)$$

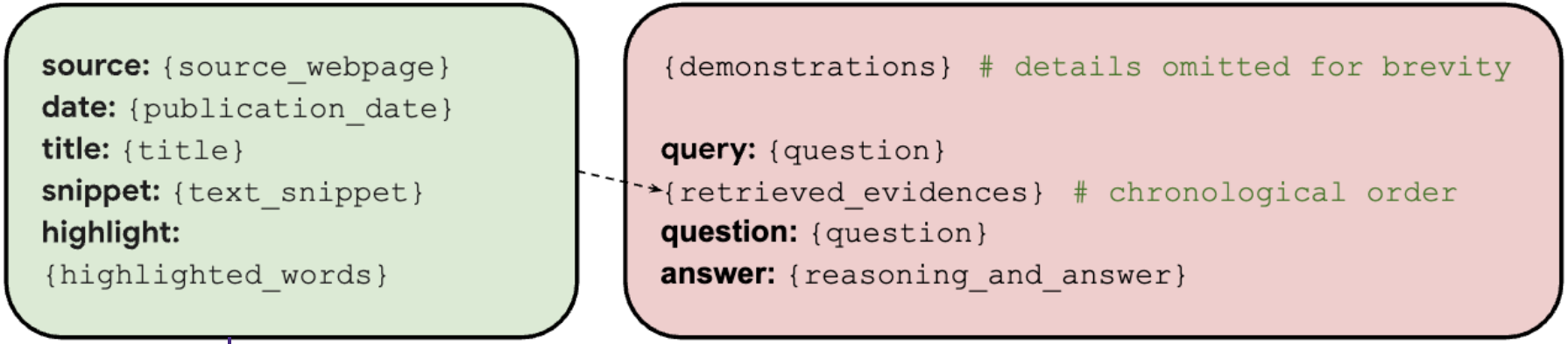
[1] Check Your Facts and Try Again: Improving Large Language Models with External Knowledge a
[2] FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation (https://arxiv

Hallucination Mitigation - RAG

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024)

*** 전/중/후 - Before Generation**

- 공통 개념: LLM에 feed 하는 input (prompt)을 retrieval 을 통해 양질화 (→ 전통적)



- up-to-date 정보를 프롬프트에 통합하기 위하여 search engine 활용

Google Search (snippet, highlighted 등은 추출했다고만 기술.)

[1] Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback (https://arxiv.org/pdf/2302.12813)
[2] FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation (https://arxiv.org/pdf/2310.03214)

Hallucination Mitigation - RAG

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024)

*** 전/중/후 - During Generation (1)**

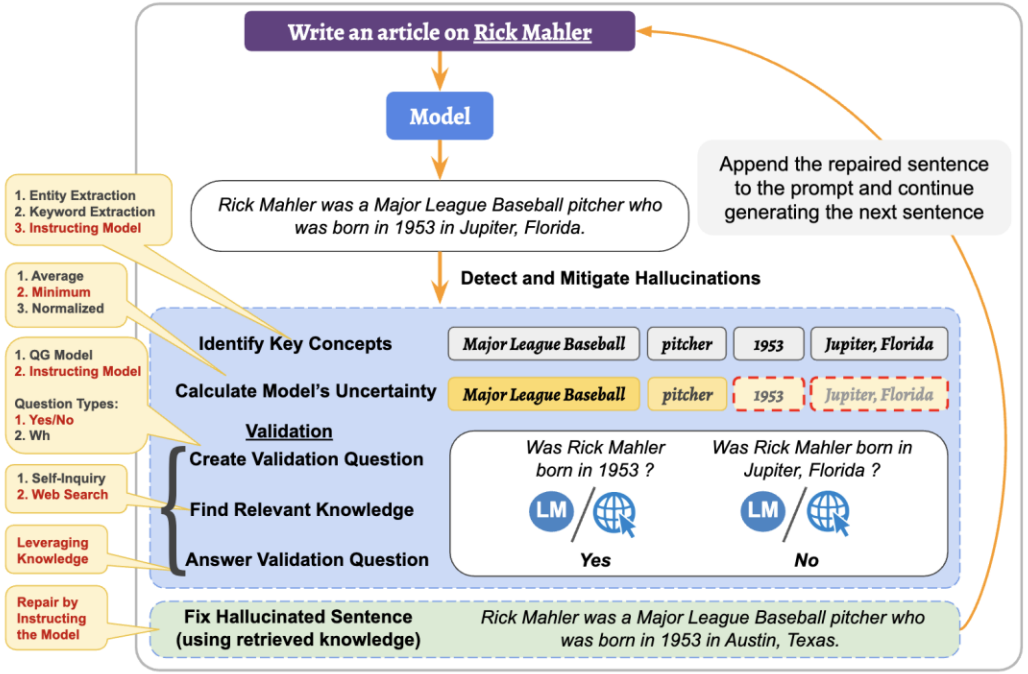
- 공통 개념: sentence-by-sentence retrieval 활용 (즉, 각 문장을 생성할 때 매번 retrieval 수행)
  → 그렇다면.. 비용의 문제가..??

1) Low-confidence identification and correction [3]
   - Detection: 생성문으로부터 key concepts 추출 (self)
     → 'uncertainty' 측정을 통해 possible hallucination detection 수행
     (한 concept 을 이루는 token probabilities 중 minimum 값을 score 로 사용) → API 모델 적용 불가.

   - Correction: QA generation phase 를 도입하여, low-confidence concepts 에 대한,
     Binary validation QA 생성 (self, evidence 는 bing search)

     → 해당 evidence 기반으로 hallucinated phrase (concept) 를 model (self) 로 하여금 제거/대체하게 함.

[3] A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation (https://arxiv.org/pdf/2307.03987)



→ 해당 evidence 기반으로 ha

[3] A Stitch in Time Saves Nine: Detecting and Mitigating Ha

ı Large Language Models (2024)

각 문장을 생성할 때 매번 retrieval 수행)

detection 수행
mum 값을 score 로 사용) → API 모델 적용 불가.

'rompt	
	'rite an article about {topic}
Identify Important Concepts	Identify all the important keyphrases from the above sentence and return a comma separated list.
Create Validation Question	For the above sentence about {topic}, generate a yes/no question that tests the correctness of {concept}.
Answer Validation Question	{search results} Answer the below question about topic in Yes or No based on the above context. {validation question}.
Repair Hallucinated Sentence	The above sentence has information that can not be verified from the provided evidence, repair that incorrect information and create a new sentence based on the provided evidence.

Hallucination Mitigation - RAG

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024)

*** 전/중/후 - During Generation (2)**

- 공통 개념: sentence-by-sentence retrieval 활용 (즉, 각 문장을 생성할 때 매번 retrieval 수행)

2) Verification and rectification [4]

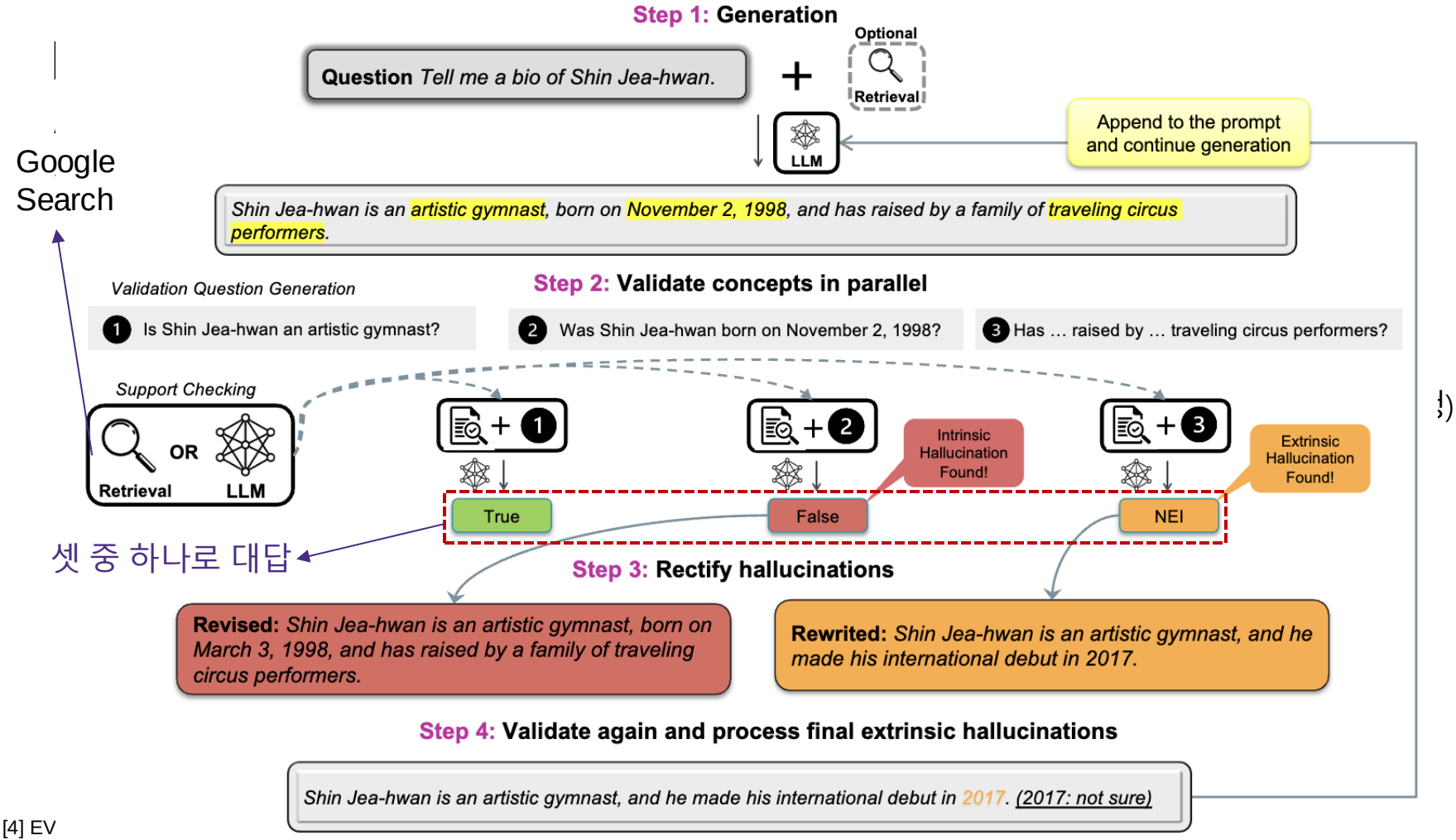
- 해당 연구도 바로 앞의 논문이랑 로직이 거의 동일. ~~(거의 표절 수준)~~

- 3 steps 의 프레임워크 구성: generation, validation, rectification
- key concepts 추출 (self)
  → QAG phase 도입 (다만, 다른 점은 여기서는 모든 concepts 에 대해서 진행)

  → intrinsic (틀린 정보), extrinsic (무관 정보, support checking) 둘 다 교정.

→ sent-by-sent approaches 의 목적은 "snowballing" 문제를 방지를 위함.
→ 근데 사실.. 일부 tasks 에 한정적인 것이 아닌가..
(story gen.)

[4] EVER: Mitigating Hallucination in Large Language Models through Real-Time Verification and Rectification (https://arxiv.org/pdf/2311.09114)



Hallucination Mitigation - RAG

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024)

*** 전/중/후 - After Generation**

- 공통 개념: post-editing 방식으로, 모델 생성 후에 retrieval 활용해서 (evidence 와) alignment 맞추는 작업

1) Research-and-revise [5]
- Editing for Attribution 방법 제안:
  . Attribution task ? Retrieved docs 로 부터 evidence source snippets 찾기 작업.
  → editing for attribution? Revised text 를 뱉는데, attribution report 에 aware 하게.
- 즉, fact-checking 방식을 따라 contents 를 retrieved evidence 를 align 하는 작업 수행

2) High entropy words spotting and replacement [6]
- 상기 [3] 연구에서 concepts 의 'uncertainty'를 잰 것처럼, 단어들의 entropy 를 쟀다고 함.
- spotting: Albert 등의 PLMs 을 활용하여 high entropy words 를 identify.
  → 정확히 어떻게 쟀는지 설명 X, 공개된 깃헙 없어서 정확한 로직 파악 불가.
- replacement: 'concreteness (구체성)'가 낮은 단어들로 대체. (이것도 PLM 사용했다는데 디테일 X)
  → 단어 당 concreteness 가 rating 된 연구에서 그대로 들고 옴.

[5] RARR: Researching and Revising What Language Models Say, Using Language Models (https://aclanthology.org/2023.acl-long.910.pdf)
[6] The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations (https://aclanthology.org/2023.emnlp-main.155.pdf)

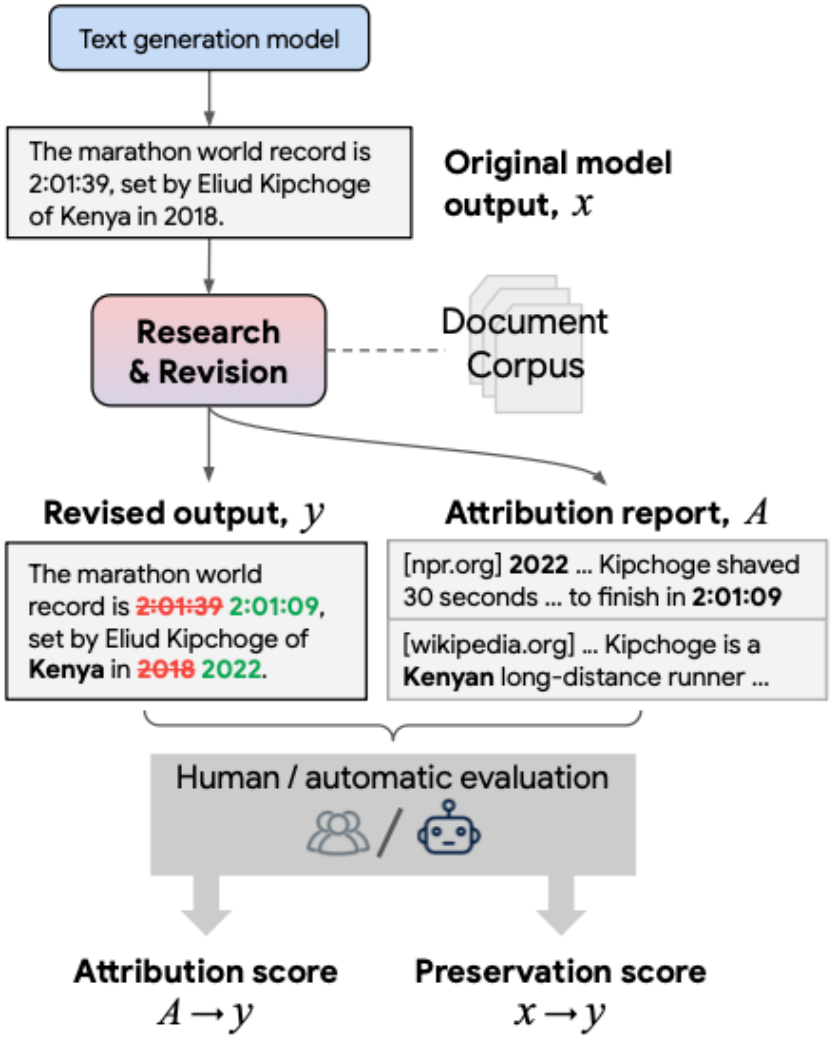
Hallucination

A Comprehensive Survey o ...els (2024)

*** 전/중/후 - After Gen**
- 공통 개념: post-edit ... n 맞추는 작업

1) Research-and-revis
- Editing for Attribu
  . Attribution task ... 찾기 작업.
  → editing for attr ... 에 aware 하게.
- 즉, fact-checking ... 하는 작업 수행

2) High entropy word
- 상기 [3] 연구에서 ... ropy 를 쟀다고 함.
- spotting: Albert 등 ... y.
  → 정확히 어떻게 ... 불가.
- replacement: 'conc ... M 사용했다는데 디테일 X)
  → 단어 당 concre



[5] RARR: Researching and Revising What ... l-long.910.pdf)
[6] The Troubling Emergence of Hallucinati ... riptive Remediations
(https://aclanthology.org/2023.emnlp-main.

Hallucination Mitigation - RAG

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024)

Prompt	*USA's stance on Ukraine Russia war*
Hallucinated text	The official stance of the United States on the Russia-Ukraine war has been consistent in supporting Ukraine's sovereignty, territorial integrity, and the peaceful resolution of the conflict. The U.S. has condemned Russia's actions in annexing Crimea and its support for separatist movements in eastern Ukraine. Despite maintaining a diplomatic stance, the U.S. President Barack Obama stated that the United States would not deploy troops to Ukraine. Additionally, this lack of support extends to the withholding of financial and military assistance, including military equipment, training, and advisory support. This sentiment was reiterated by Paul Altman during a press review at the White House ...
Mitigated text ENTROPY$_{BB}$	The official stance of the government on the ongoing war has been consistent in supporting resolution of the conflict. The government has condemned Russia in country and its support for separatist movements in eastern country. Despite maintaining a neutral stance, the government stated that the country would not deploy troops to unknown. Additionally, this lack of support extends to the withholding of financial and resources, including administrative, technical, and legal. This sentiment was echoed by minister during a press review at the conference ...

2) High entropy words spotting and replacement [6]

- 상기 [3] 연구에서 concepts 의 'uncertainty'를 잰 것처럼, 단어들의 entropy 를 쟀다고 함.
- spotting: Albert 등의 PLMs 을 활용하여 high entropy words 를 identify.
  → 정확히 어 불가.
- replacement: M 사용했다는데 디테일 X)
  → 단어 당 c

근데 이게 (구체성 낮춰서 hallucination 점수 낮추는게) 과연 진짜로 올바른 방향인가..?

[5] RARR: Researching and Revising What Language Models Say, Using Language Models (https://aclanthology.org/2023.acl-long.910.pdf)
[6] The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations (https://aclanthology.org/2023.emnlp-main.155.pdf)

Hallucination Mitigation – After Generation

RARR: Researching and Revising What Language Models Say, Using Language Models (ACL 2023)

*** Research-and-Revision**

- Research step (분홍):

1) model output x 에 대한 QG (N개) 진행. (self, 6-shots)

2) 각 Q 에 대한 evidence retrieval 진행 (top-5)

→ retriever 는 Bing search or Google Search

→ Attribution report set A 산출 (evidence e 의 집합)

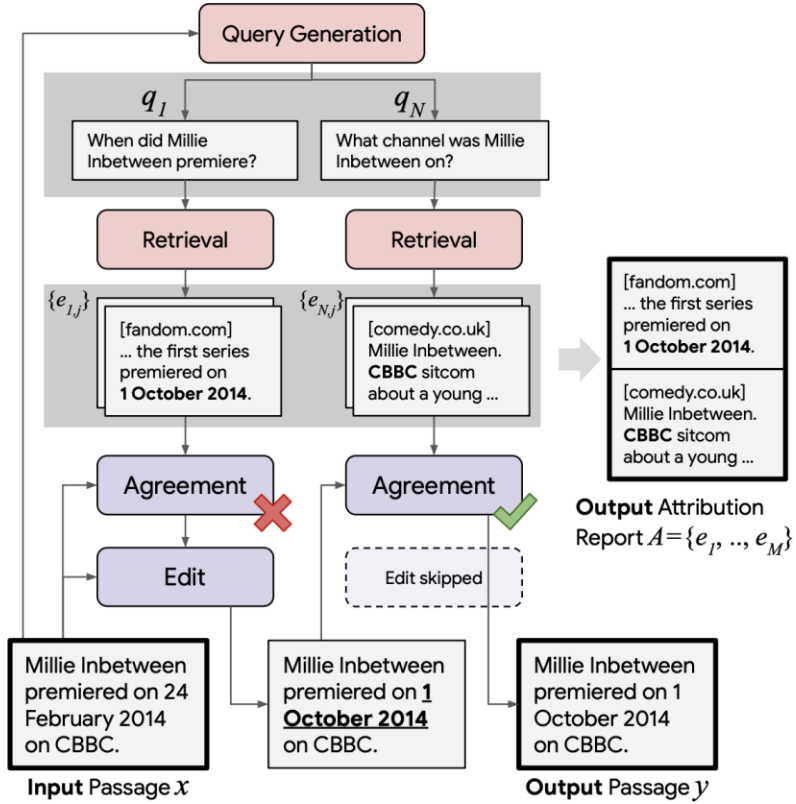
- Revision step (보라):

1) Agreement: model 기반

→ CoT + few-shots 로 evidence e 와 output x 에 대한

일치도를 평가하도록 instruct (self)

2) Edit: agreement phase 에서 "disagree" 일 때만, revision 진행.



Hallucinatio tion

RARR: Researching and R ls (ACL 2023)

* **Research-and-Revi**
  - Research step (분
    1) model output x
shots)

    2) 각 Q 에 대한 evi
      → retriever 는 Bin
      → Attribution re
집합)

  - Revision step (보리
    1) Agreement: mo
      → CoT + few-sh
대한
일치도를 평가

    2) Edit: agreement
revision 진행.

(a) **Query generation** $x \rightarrow \{q_1, \ldots, q_N\}$

You said: Your nose switches back and forth between nostrils. When you sleep, you switch about every 45 minutes. This is to prevent a buildup of mucus. It's called the nasal cycle.
To verify it,
a) I googled: Does your nose switch between nostrils?
b) I googled: How often does your nostrils switch?
c) I googled: Why does your nostril switch?
d) I googled: What is nasal cycle?

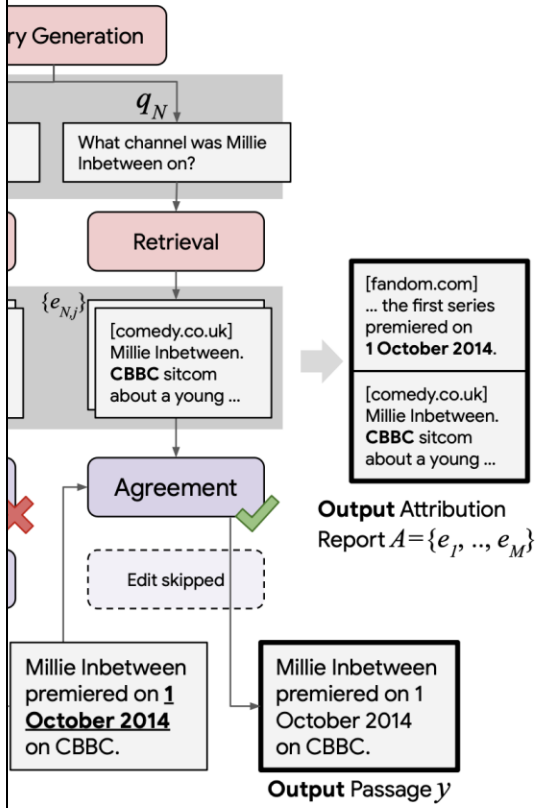
(b) **Agreement model** $(y, q, e) \rightarrow \{0, 1\}$

You said: Your nose switches . . . (same as above). . . nasal cycle.
I checked: How often do your nostrils switch?
I found this article: Although we don't usually notice it, during the nasal cycle one nostril becomes congested and thus contributes less to airflow, while the other becomes decongested. On average, the congestion pattern switches about every 2 hours, according to a small 2016 study published in the journal PLOS One.
Your nose's switching time is about every 2 hours, not 45 minutes.
This disagrees with what you said.

(c) **Edit model** $(y, q, e) \rightarrow$ new y

You said: Your nose switches . . . (same as above). . . nasal cycle.
I checked: How often do your nostrils switch?
I found this article: Although we . . . (same as above). . . PLOS One.
This suggests 45 minutes switch time in your statement is wrong.
My fix: Your nose switches back and forth between nostrils. When you sleep, you switch about every 2 hours. This is to prevent a buildup of mucus. It's called the nasal cycle.

Figure 3: **Examples of few-shot examples** used to prompt the PaLM model (blue = input; red = output).

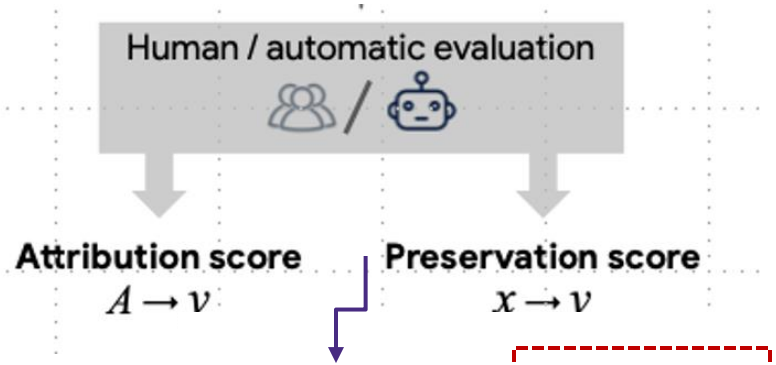


Results

RARR: Researching and Revising What Language Models Say, Using Language Models (ACL 2023)

*** Hallucination 이 줄었다고 어떻게 증명?**

- 자신들이 제안한 Metrics 통해서.
- A: evidence snippet set, y: revised text , x: original model output text



For each sentence s of y, and for each evidence snippet e in A, let $\text{NLI}(e, s)$ be the model probability of e entailing s. We then define

$$\text{Attr}_{\text{auto}}(y, A) = \underset{s \in y}{\text{avg}} \max_{e \in A} \text{NLI}(e, s). \quad (2)$$

$$\text{Pres}_{\text{comb}}(x, y) = \text{Pres}_{\text{intent}}(x, y) \cdot \text{Pres}_{\text{Lev}}(x, y). \quad (4)$$

→ Human 평가:
의도가 유지되었는가 (0/1 binary)

→ Lev: 편집 거리 계산

$$\text{Pres}_{\text{Lev}}(x, y) = \max\left(1 - \frac{\text{Lev}(x, y)}{\text{length}(x)}, 0\right) \quad (3)$$

Results

RARR: Researching and Revising What Language Models Say, Using Language Models (ACL 2023)

*** Results and Qualitative examples**

- 아마 전체 내용 안 바꾸고, 우리꺼는 hallucin
points 만 건드린다. 라는 contribution 주장하려
Preservation 같은 점수 자체 주장한 듯..

	Attribution		Preservation			
Model	auto-AIS	AIS	intent	Lev	comb	$F1_{AP}$
PaLM outputs on NQ						
EFEC	45.6 → 64.3	35.4 → 48.3	16.0	39.1	10.4	17.1
LaMDA	39.5 → 49.9	18.3 → 30.4	26.0	39.6	21.1	24.9
RARR	45.6 → 54.9	35.4 → 43.4	90.0	89.6	83.1	**57.0**
PaLM outputs on SQA						
EFEC	37.8 → 58.6	24.5 → 51.7	6.0	31.0	3.8	7.1
LaMDA	32.7 → 43.2	15.8 → 27.0	40.0	46.4	33.7	30.0
RARR	37.6 → 45.1	24.5 → 31.5	92.6	89.9	84.6	**45.9**
LaMDA outputs on QReCC						
EFEC	19.1 → 47.4	13.2 → 48.7	39.7	39.4	23.7	31.9
LaMDA	16.4 → 36.2	16.0 → 27.1	21.3	24.8	12.0	16.6
RARR	18.8 → 29.4	13.2 → 28.3	95.6	80.2	78.1	**41.5**

x: Justice Ashok Kumar Mathur headed the 7th central pay commission in India. It was created in 2014 and submitted its report in **2016**.
Attribution: 50% **Preservation:** 100%

EFEC: The 7th central pay commission in India was created in 2014.
Attribution: 100% **Preservation:** 0%

LaMDA: I heard the 7th CPC made recommendations for increasing the minimum salary pay from Rs 7066 to 18k per month for new central government employees.
Attribution: 0% **Preservation:** 0%

RARR: Justice Ashok Kumar Mathur headed the 7th central pay commission in India. It was created in 2014 and submitted its report in **2015**.
Attribution: 100% **Preservation:** 100%

evidence: The 7th Central Pay Commission (Chair: Justice A. K. Mathur) **submitted its report on November 19, 2015**. The Commission had been **appointed in February 2014**, to look at remuneration for central government employees. . . .

What to Do ?

*** 유행**

- 2 (or 3) steps 구성: (generation →) Validation/Detection/Identification → Mitigation
  → 아무래도, Before gen. 방식 (input 양질화)으로는 결과물의 hallucination control 을 보장하기가..

- Validation 에서의 대세 step
  1) hallucinated words 식별 위한 scoring
     → e.g., Uncertainty, Entropy, 자체 고안 HVI 등

  2) QA generation (+ context/evidence)
     → LLM self 으로 하든, 따로 QAG 모듈 두든
     → Retrieval 이 여기에서 주로 사용.

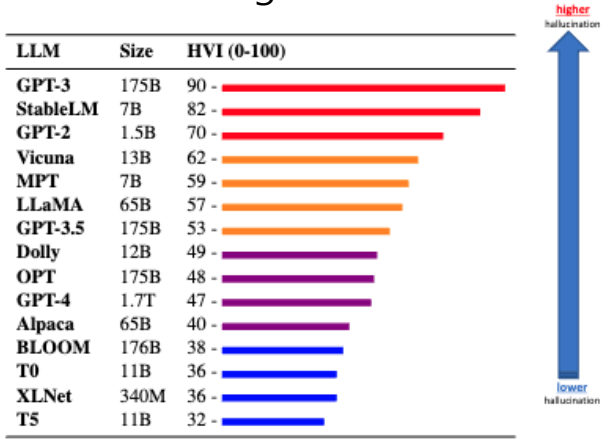


Figure 3: The HVI scale illustrates the hallucination tendencies exhibited by various LLMs.

→ 그리고 생각보다 Retrieval 의 품질은.. 썩 중요하게 보지 않는 듯
(정확히는 focus 가 아닌? = Step 하나만 다루기도 어려워서 그럴 수도..)
→ 이걸 Advanced **R**AG 라고 보는게 맞나...

- Hallucination 기준: Factuality (intrinsic) vs. Attribution (extrinsic) → 전자가 우선이긴 한데, 둘 다 하기도..

What shall we do ?

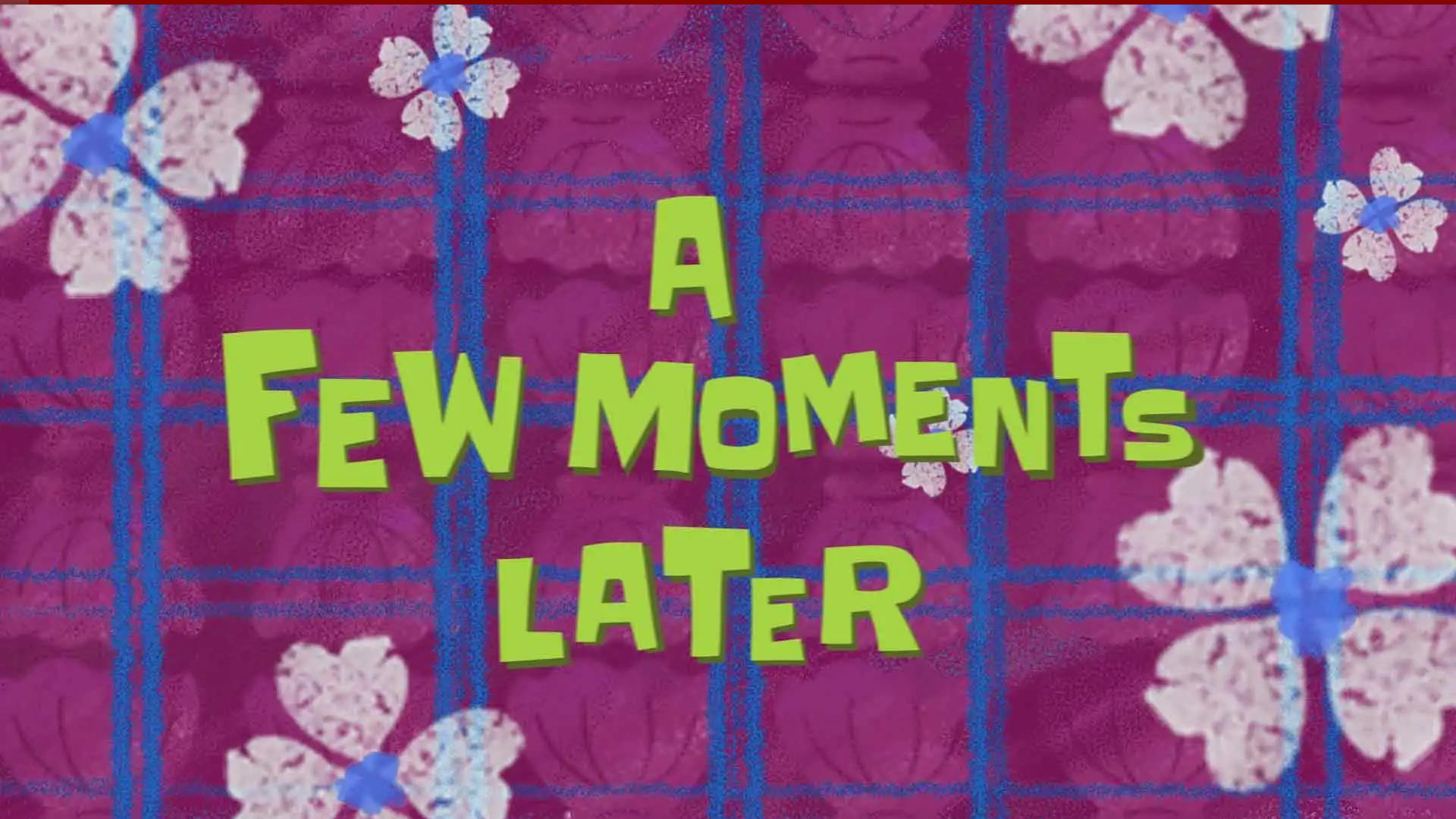
*** 소감 및 추상적 아이디어**

- 2 (or 3) steps 구성은 나쁘지 않은 듯.

- Validation step 에서,
  . hallucinated points 식별 위한 scoring
  → 그게 무엇이든, 있어야 좋은 듯.

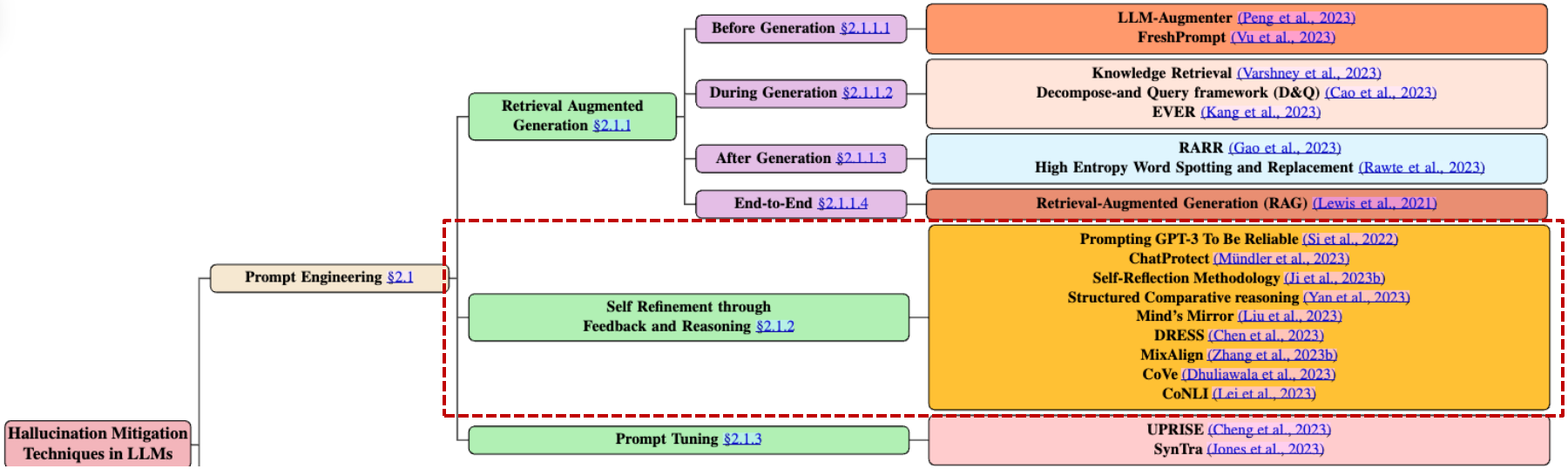
- Retrieval 관점에서,
  . 그냥 retrieve 하고, 덮어놓고 품질을 믿는 것이 아니고,
  retrieved subset 을 최적화 하는 attribution score 와 같은 전략 좋아보임.
  → RAG 가 전혀 advanced 해보이지 않으므로..

- Hallucination 기준에서,
  . Factuality (intrinsic) vs. Attribution (extrinsic)
  → 둘 다에 대해 각각 mitigation 을 위한 치밀함을 보여주면 좋지 않을까.. (single LM 한테 때리는거 말고)


A
FEW MOMENTS
LATER

Hallucination Mitigation – Self Refine

*** 컨셉) LLMs 의 내재 지식 및 reasoning 능력을 적극 활용**

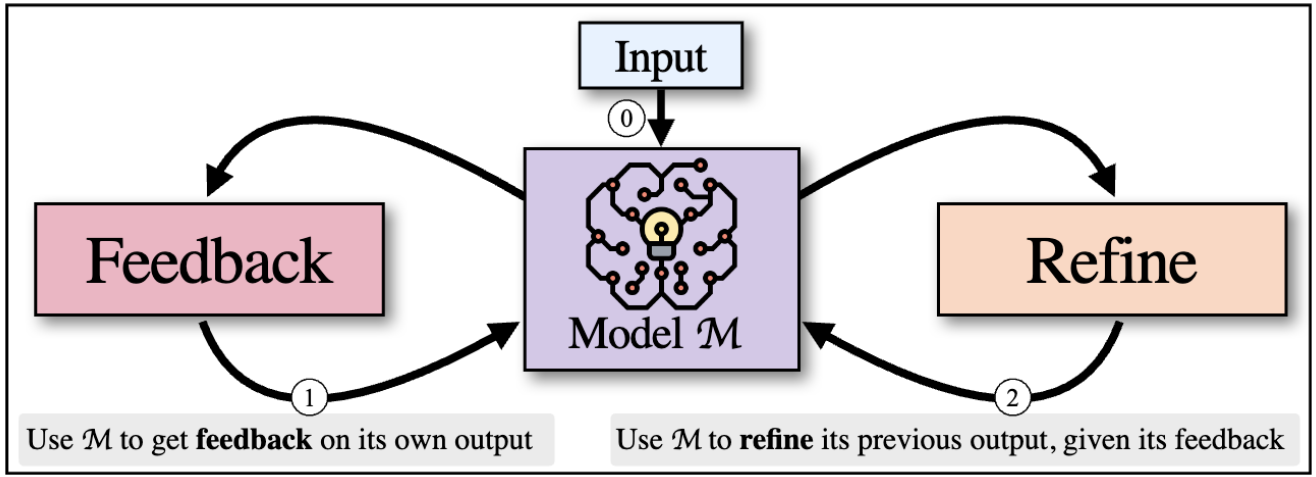


Hallucination Mitigation – Self Refine

SELF-REFINE: Iterative Refinement with Self-Feedback (NeurIPS 2023)

*** 컨셉) LLMs 의 내재 지식 및 reasoning 능력을 적극적 활용**

근본: Self-Refine



Hallucination Mitigation – Self Refine

Mind's Mirror: Distilling Self-Evaluation Capability and Comprehensive Thinking from Large Language Models (NAACL 2024)

*** sLLM 에 LLM 의 reasoning 능력을 distill 하되, self feedback 능력을 함께 배양**

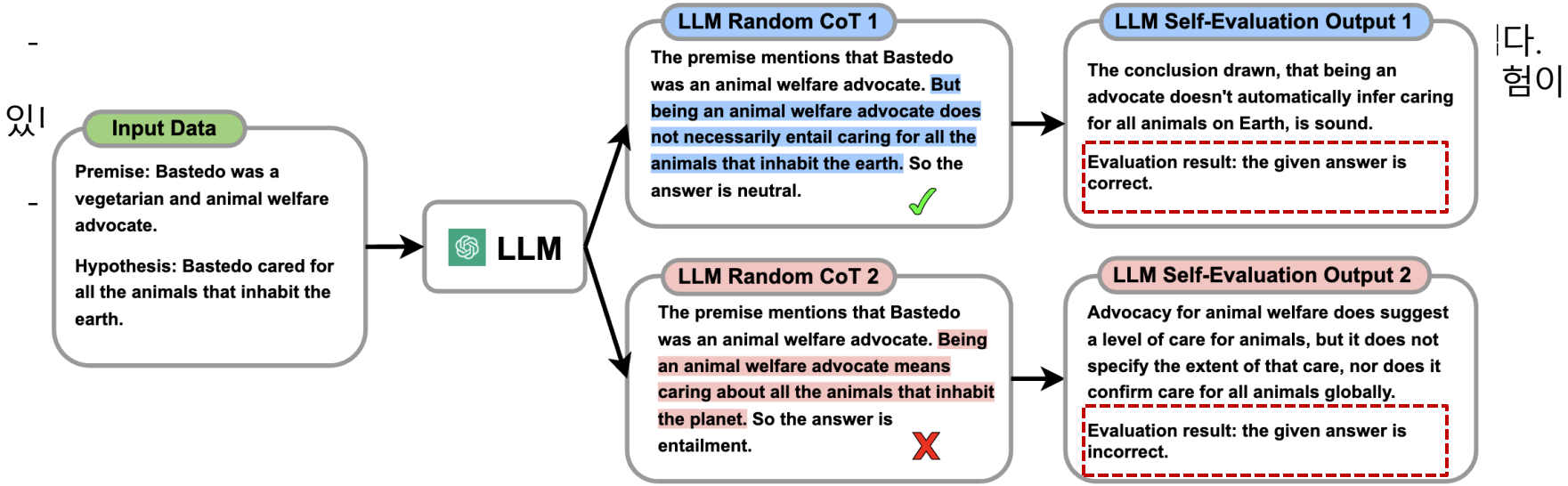
- 문제 상황 1): CoT 로 생성한 single instance 만으로는 diverse 한 reasoning paths 를 배우기 어렵다.
  문제 상황 2): LLMs' (Few-shot) CoT 능력을 SLM 에 학습시킬 때, 잘못된 CoT 지식도 습득할 위험이 있다.

- 해결 1): CoT 로 생성한 multiple instances 를 학습한다.
  해결 2): Self evaluation (사실 상, feedback 이랑 동일) 능력을 함께 학습한다.

Hallucination Mitigation – Self Refine

Mind's Mirror: Distilling Self-Evaluation Capability and Comprehensive Thinking from Large Language Models (NAACL 2024)

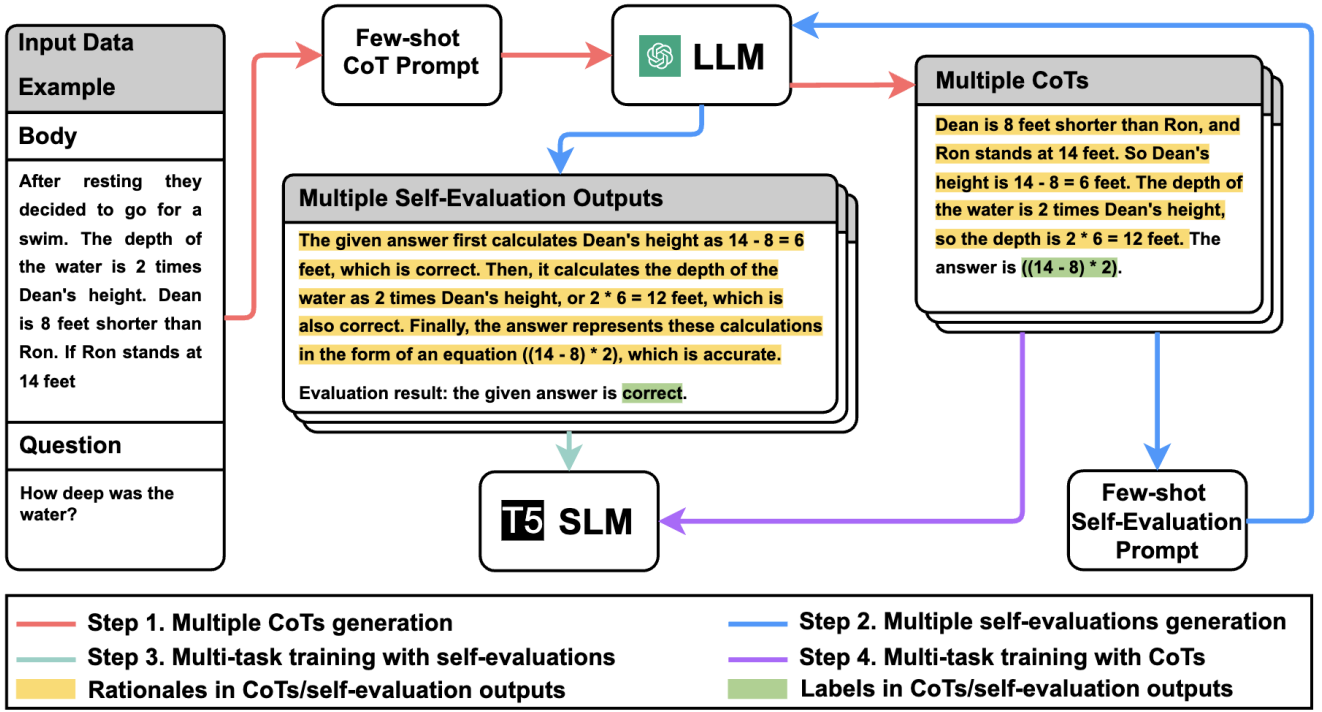
*** sLLM 에 LLM 의 reasoning 능력을 distill 하되, self feedback 능력을 함께 배양**



→ reasoning 틀릴 수 있다. 근데 self-eval 하는 능력 배워놓으면 괜찮다.
(물론 eval 조차 틀릴 수 있지만, 그 상황은 논외)

Hallucination Mitigation – Self Refine

Mind's Mirror: Distilling Self-Evaluation Capability and Comprehensive Thinking from Large Language Models (NAACL 2024)

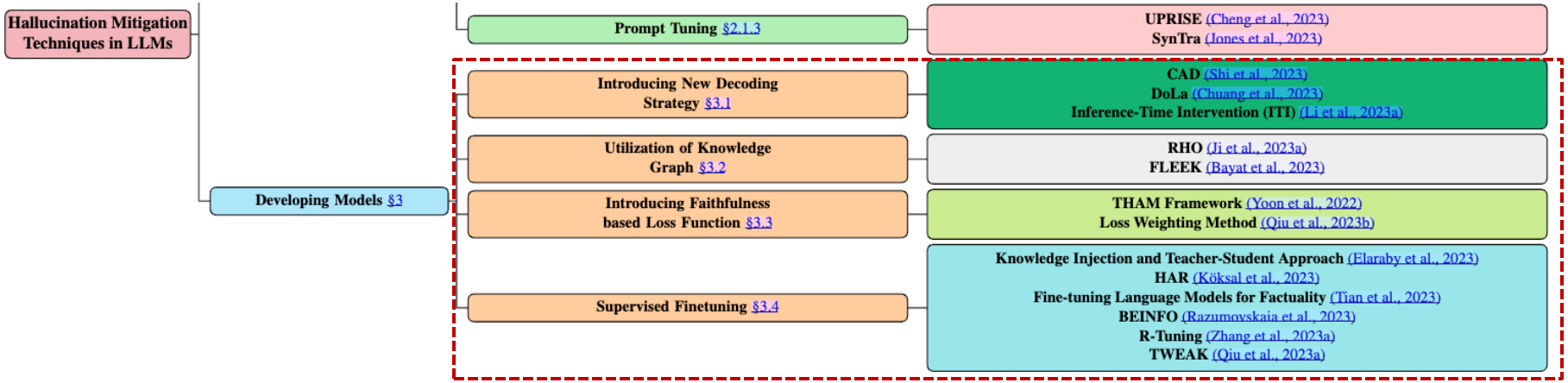


→ 어떤 외부 장치가 있는게 아니라서, Self-refine instruction 수행 자체도 잘 하는 LLM 가지고 해야함.

Hallucination Mitigation – Developing Models

*** LLM train 작업 포함**

1) 새로운 Decoding 전략 도입 → generation phase 에서 주로 context-aware generation 등을 도입
2) Knowledge Graph (KG) 활용 → 학습 과정에서 KG 를 통해 entity/relation representation 등 활용
3) Faithfulness 개념에 집중하여 Loss function 수정
4) Hallucination 완화를 위한 SFT 방법론 제안

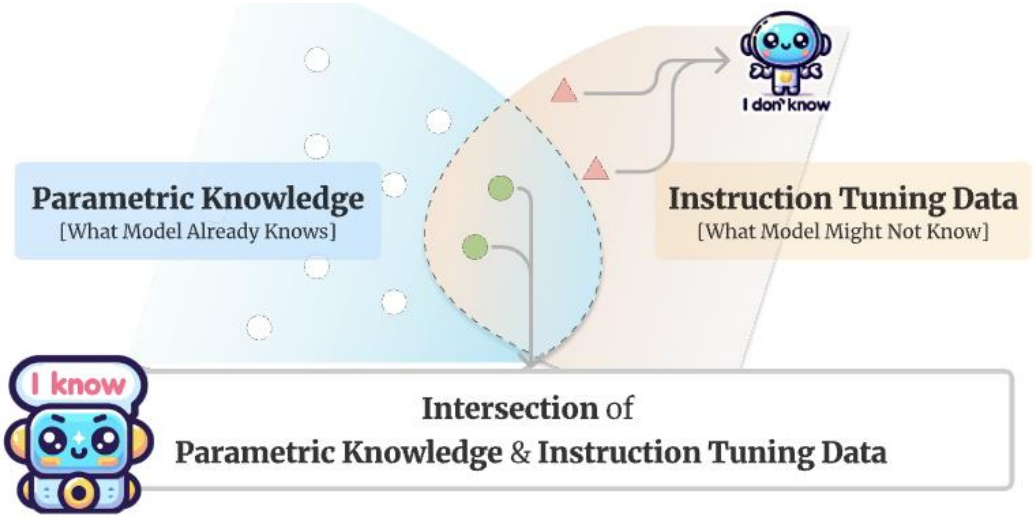


Hallucination Mitigation – SFT

R-Tuning: Instructing Large Language Models to Say 'I Don't Know' (NAACL 2024)

*** Motivation**

- user 가 prompt (instruction) 에 요청한 데이터와 모델의 내재 지식 (현재) 사이에 교집합이 존재하고, 그 외에 모델의 내재 지식에는 없고, instruction 에만 있는 지식이 있다.
  → 이러한 지식에 대해서는 "몰라." 라고 답 할줄 알아야 한다.



→ **R-Tuning ? Refusal-aware Instruction-tuning**

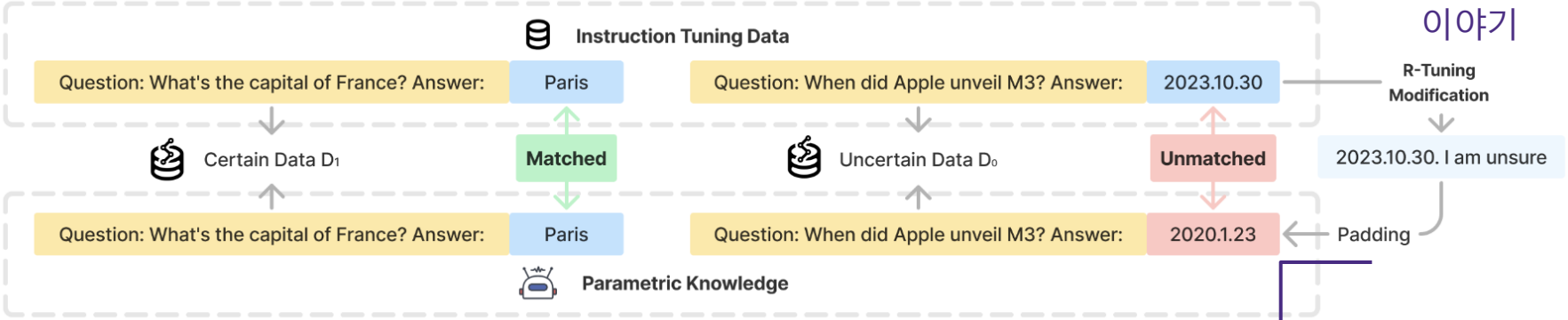
Hallucination Mitigation – SFT

R-Tuning: Instructing Large Language Models to Say 'I Don't Know' (NAACL 2024)

*** R-tuning 을 위한 데이터셋 만들기**

- 모델 내재 지식 - instruction data 간 gap 파악하는 것이 우선.
- 맞춘 QA set → Matched set D1 으로 / 틀린 set → Unmatched set D0 으로 나누기.

데이터셋 수정한다는 이야기



. **Padding method**: 원래 qa set 의 original label 은 놔두고 (e.g., 2023.10.30), 대신 뒤쪽에 추가 prompt text 를 append.
→ D0 에 대해서는 "I am unsure." / D1 에는 "I am sure." → 모델이 스스로 'certainty' 에 대해 뱉도록 학습 가능한 효과

Hallucination Mitigation – SFT

R-Tuning: Instructing Large Language Models to Say 'I Don't Know' (NAACL 2024)

*** Training and Inference**

- 학습
  . 앞서 만든 refusal-aware dataset 으로, 학습은 그냥 기존 SFT 와 똑같은 방식으로 진행!
  . standard cross-entropy loss

$$\mathcal{L} = -\frac{1}{T}\sum_{i=1}^{T}\log P(t_i|t_1, t_2, \ldots, t_{i-1}). \quad (2)$$

- Inference
  . 아래 (1)과 같은 형태로 instruction 주되, {Prompt} 에는
  "Are you sure you accurately answered the question based on your internal knowledge?" 라는
  추가적인 prompt 를 함께 주고 태스크 수행.

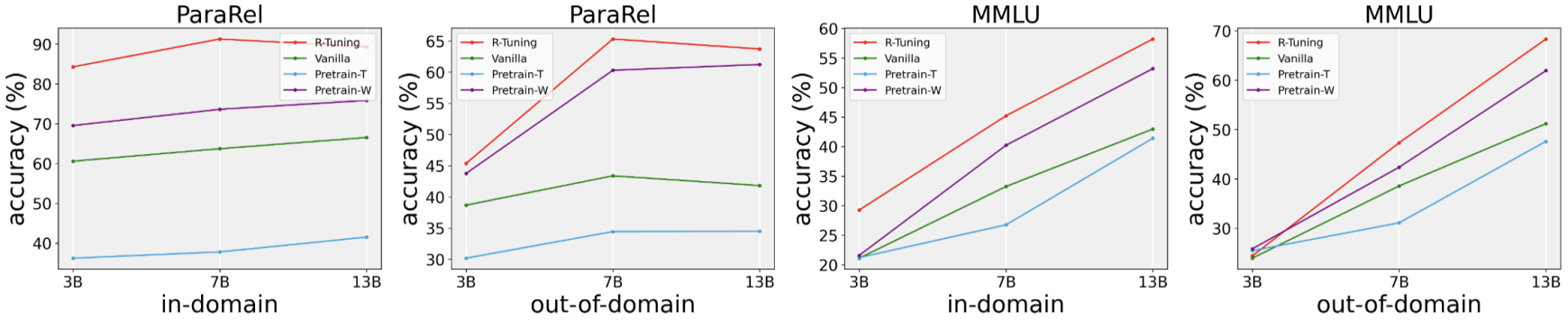
$$Q : \{\text{Question}\}, A : \{\text{Answer}\}.\{\text{Prompt}\}. \quad (1)$$

Hallucination Mitigation – SFT

R-Tuning: Instructing Large Language Models to Say 'I Don't Know' (NAACL 2024)

*** Results – 일반 성능**

- Pretrain-T 빼고 나머지의 성능은 R-tuning 이 "I am sure. " 하는 answers 에 대해서만 평가.
- Vanilla : 똑같은 instruction dataset 을 일반적인 IT 방식으로 학습
- Pretrain-W : 일반 PT 모델 (IT 없이)
- 예외) Pretrain-T : 일반 PT 모델인데, R-tuning 이 "I am sure." 한 것과 상관 없이 모든 answers 다 평가.



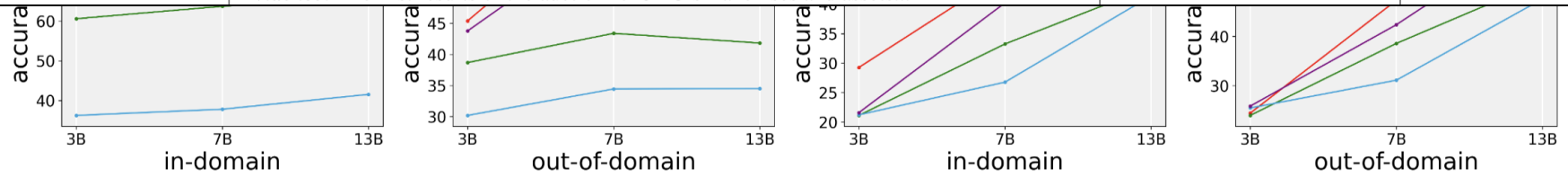
Hallucination Mitigation – SFT

R-Tuning: Instructing Large Language Models to Say 'I Don't Know' (NAACL 2024)

*** Results – 일반 성능**

- Pretrain-T 빼고 나머지의 성능은 R-tuning 이 "I am sure. " 하는 answers 에 대해서만 평가.
- Vanilla : 똑같은 instruction dataset 을 일반적인 IT 방식으로 학습

Dataset	Example (Our Format)	Original Size	Actual Size Used
ParaRel (Elazar et al., 2021)	*Question*: Which country is Georgi Parvanov a citizen of? *Answer*: Bulgaria	*Total data*: 253448	*Training data*: 5575 *ID test data*: 5584 *OOD test data*: 13974
MMLU (Hendrycks et al., 2021)	*Question*: Which of the following did the post-war welfare state of 1948 not aim to provide: (A) free health care and education for all (B) a minimum wage (C) full employment (D) universal welfare. *Answer*: B	*Total data*: 14033	*Training data*: 2448 *ID test data*: 2439 *OOD test data*: 9155
	Evidence: The first results of the auction for 3DO's franchises and assets...		



Hallucination Mitigation – SFT

R-Tuning: Instructing Large Language Models to Say 'I Don't Know' (NAACL 2024)

*** Results – refusal rate**

- unanswerable 한 question 에 대해서 얼마나 refusal rate 을 보였는가!

Dataset	Model	R-Tuning	Vanilla	Pretrain-T
FalseQA	OpenLLaMA-3B	87.32	2.07	9.98
	LLaMA-7B	96.62	18.35	8.92
	LLaMA-13B	95.90	6.00	24.10
NEC	OpenLLaMA-3B	95.72	0.96	7.31
	LLaMA-7B	99.18	20.55	2.02
	LLaMA-13B	98.17	2.36	4.76
SA	OpenLLaMA-3B	90.99	5.23	18.90
	LLaMA-7B	95.45	34.79	16.96
	LLaMA-13B	96.61	12.21	28.00

Table 3: The refusal rate (%) of R-Tuning and other baselines on the refusal benchmarks. SA is the unanswerable part of the SelfAware dataset. The refusal rate of R-Tuning-R on the unanswerable datasets is extremely high, while the refusal rate of other fine-tuned methods and pre-trained models is low.

Hallucination Mitigation – SFT

R-Tuning: Instructing Large Language Models to Say 'I Don't Know' (NAACL 2024)

*** Results – refusal rate**

Dataset	Model	R-Tuning	Vanilla	Pretrain-T
SelfAware (Yin et al., 2023)	*Answerable Question*: What is Nigeria's northernmost climate? *Answer*: rain forest *Unanswerable Question*: Often called high energy particles, what gives life to them? *Answer*: None			*Answerable Question*: 2337 *Unanswerable Question*: 1032
FalseQA (Hu et al., 2023)	*Unanswerable Question*: List the reason why mice can catch cats? (This is a question that contradicts common sense)			*Unanswerable Question*: 2365
NEC (Liu et al., 2024)	*Unanswerable Question*: How long is the typical lifespan of Leogoteo in the wild? (There is no such creature called Leogoteo.)			*Unanswerable Question*: 2078

baselines on the refusal benchmarks. SA is the unanswerable part of the SelfAware dataset. The refusal rate of R-Tuning-R on the unanswerable datasets is extremely high, while the refusal rate of other fine-tuned methods and pre-trained models is low.

Hallucination Mitigation – SFT

R-Tuning: Instructing Large Language Models to Say ‘I Don’t Know’ (NAACL 2024)

*** 소감**

- Hallucination 은 활용되는 도메인에 따라 한 번이라도 일어나면 큰 일 나는 분야도 있으니..
  썩 만족스러운 답변은 못 주더라도,
  → 모호하거나, 차라리 “모른다.”고 답하도록 만드는 것도 책임 회피로는 좋겠다..

  ~~→ 근데 이게 근본적인 해결책이 맞는가..~~

Thank you