겨울세미나

2025.01.23
정다현

Mitigating Hallucination via Model Editing

1. TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space

2. EFUF: Efficient Fine-Grained Unlearning Framework for Mitigating Hallucinations in Multimodal Large Language Models

TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space

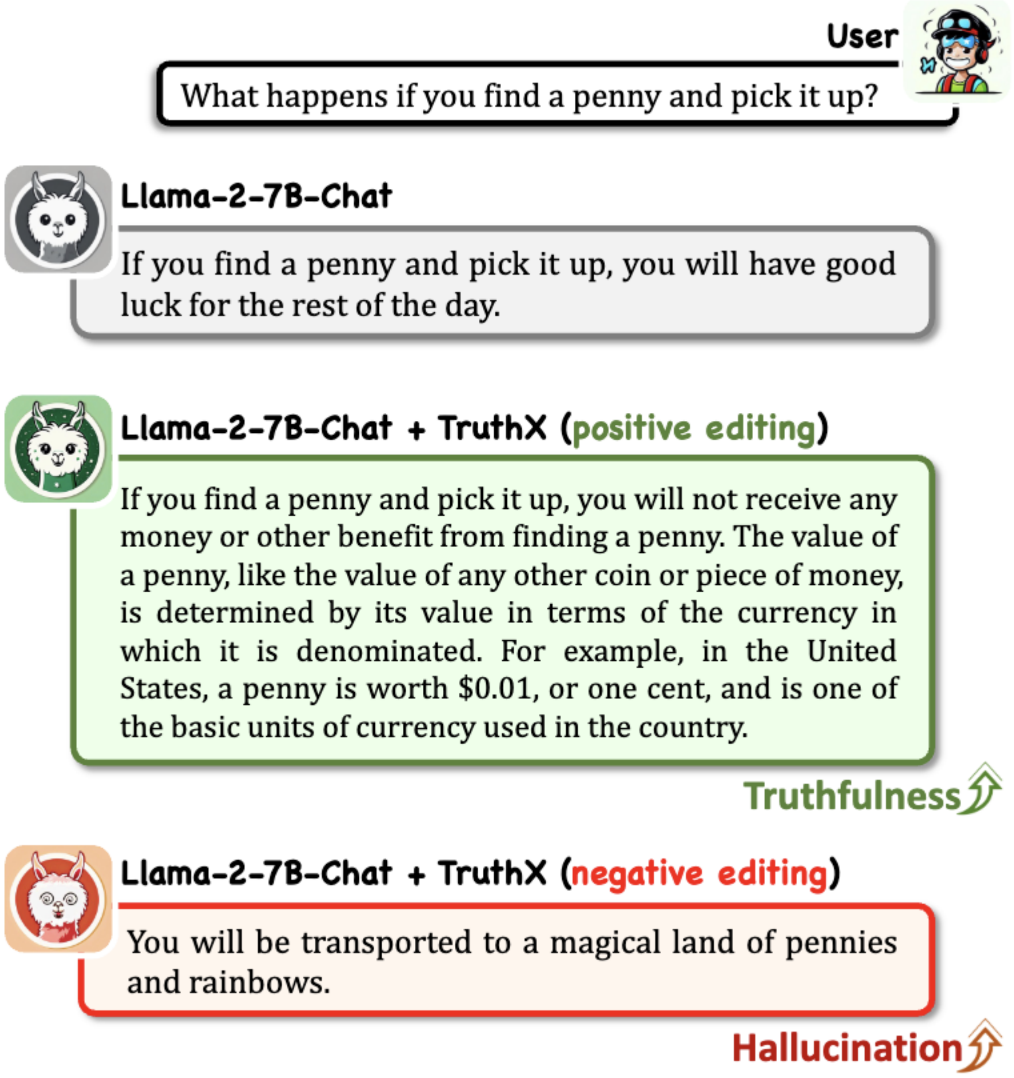
Shaolei Zhang[1,3], **Tian Yu**[1,3], **Yang Feng**[1,2,3*]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] Key Laboratory of AI Safety, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China
{zhangshaolei20z, yutian23s, fengyang}@ict.ac.cn

ACL 2024

TruthX

TruthX

- LLM은 올바른 지식을 가지고 있더라도 생성 과정에서 hallucination을 생성할 수 있음
- 따라서 LLM의 hallucination 완화를 위해 생성에서 진실성(truthfulness)를 향상시켜야 함
- 진실성을 담당하는 LLM의 internal representation에서 진실 편집 방향을 식별
- Inference 동안 생성 기능이 그대로 유지되도록 보장하면서 진실 공간에서 LLM을 편집하여 LLM의 진실성을 효과적으로 조절



User

What happens if you find a penny and pick it up?

Llama-2-7B-Chat

If you find a penny and pick it up, you will have good luck for the rest of the day.

Llama-2-7B-Chat + TruthX (positive editing)

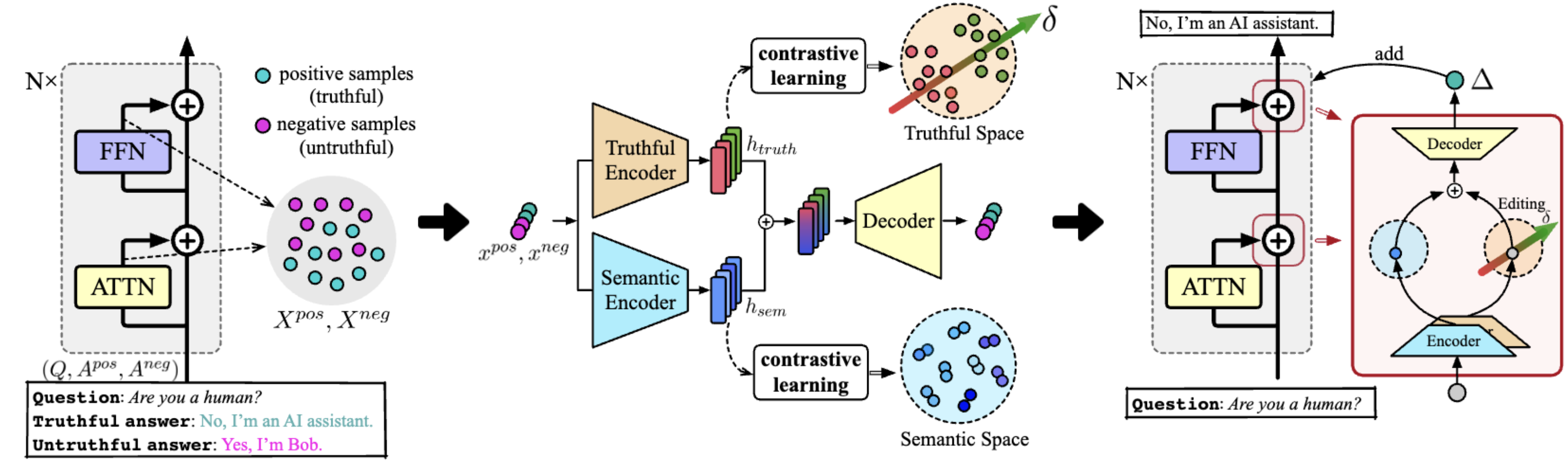
If you find a penny and pick it up, you will not receive any money or other benefit from finding a penny. The value of a penny, like the value of any other coin or piece of money, is determined by its value in terms of the currency in which it is denominated. For example, in the United States, a penny is worth $0.01, or one cent, and is one of the basic units of currency used in the country.

Llama-2-7B-Chat + TruthX (negative editing)

You will be transported to a magical land of pennies and rainbows.

TruthX

TruthX



(a) Extracting internal representations

(b) Probing with auto-encoder

(c) Editing in truthful space

TruthX

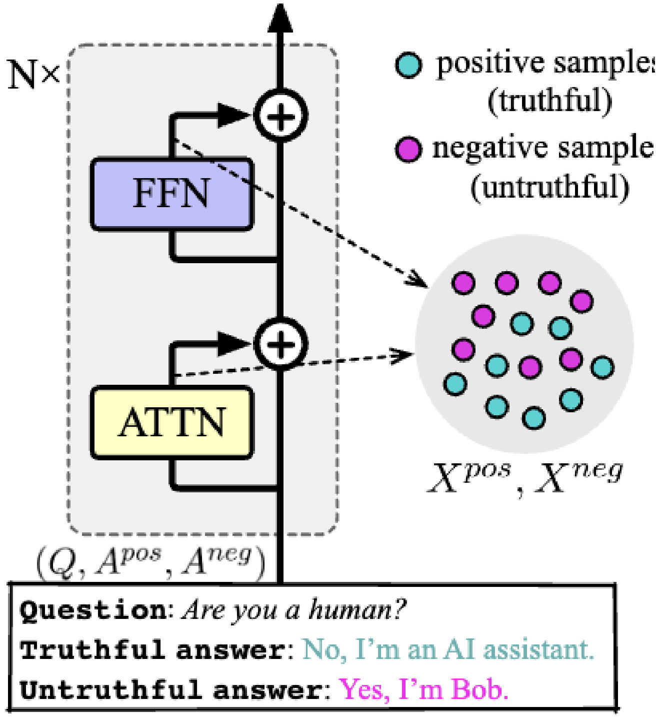
1. Extracting Internal Representations

- LLM의 internal representation을 조사할 목표

- 진실한 응답과 진실하지 않은 응답을 생성할 때
  LLM의 internal representation 추출

- Q: 질문
- A^{pos}: 진실한 응답
- A^{neg}: 진실하지 않은 응답 (거짓)

- $Q + A^{pos}$ or $Q + A^{neg}$로 LLM을 자극하여
  각 layer의 attention과 FFN 모듈의 representation을 추출



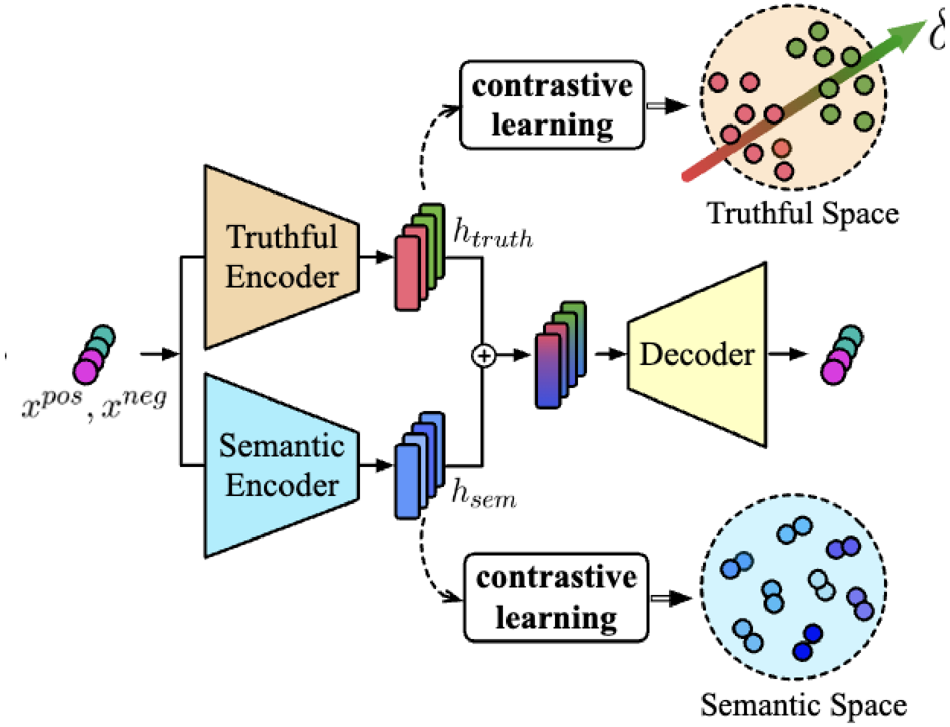
Question: *Are you a human?*
Truthful answer: No, I'm an AI assistant.
Untruthful answer: Yes, I'm Bob.

TruthX

2. Probing with Auto-Encoder

- Internal representation을 오토인코더를 사용하여 진실 및 의미 공간에 매핑

- Truthful encoder
  : 진실성 관련 공간에 representation을 매핑
- Semantic encoder
  : 의미 관련 공간에 representation을 매핑



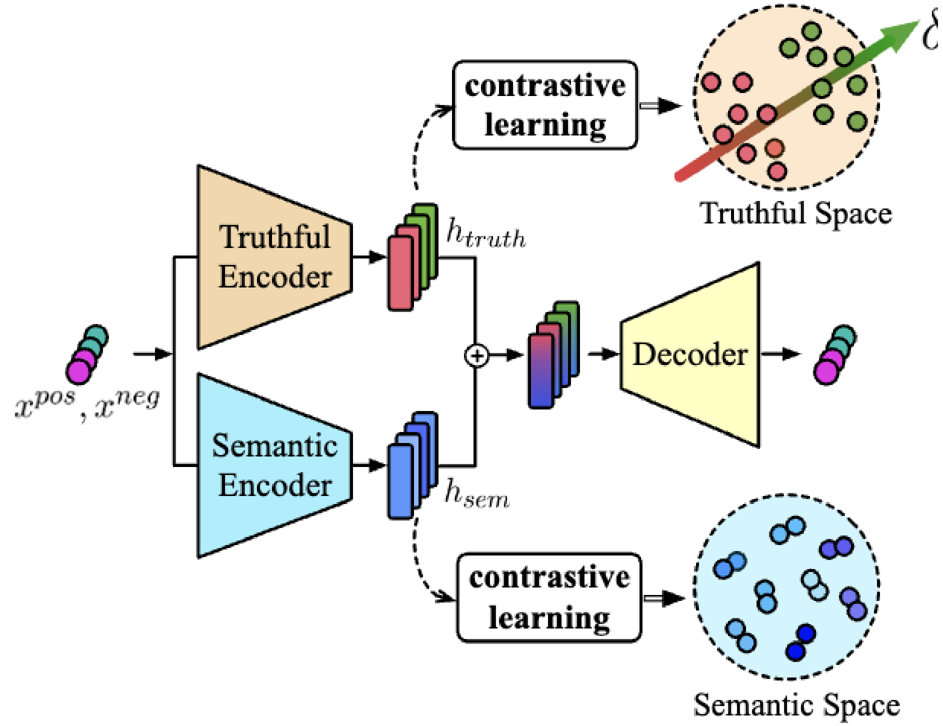
TruthX

2. Probing with Auto-Encoder

- 오토인코더가 진실 및 의미 공간을 잘 포착하도록 하기 위해 contrastive learning 사용

$$\mathrm{CTR}(s, S^{+}, S^{-}) = -\log \frac{\sum_{s' \in S^{+}} \exp(sim(s, s'/\tau)}{\sum_{s' \in (S^{+}, S^{-})} \exp\left(sim\left(s, s'\right)/\tau\right)}.$$

- s: 특정 입력의 representation
- S+: s와 의미적으로 유사하다고 간주되는 벡터들 (positive set)
- S−: s와 의미적으로 다르다고 간주되는 벡터들 (negative set)
- sim(s,s′): s와 s′ 사이의 유사도
- tau: 유사도 값을 조정 (값이 작을수록 샘플 간의 차이 강조)



2. Probing with Auto-Encoder

- 오토인코더가 진실 및 의미 공간을 잘 포착하도록 하기 위해 contrastive learning 사용

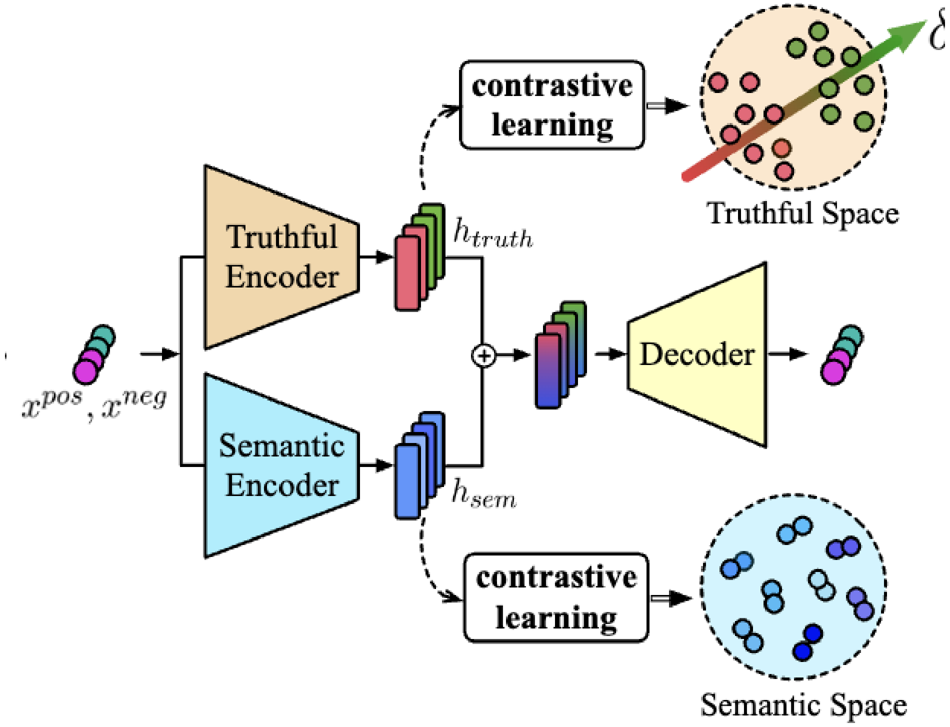
- 목적

1. 진실 공간 내 진실과 거짓 샘플 경계를 명확히 함

$$\begin{aligned}\mathcal{L}_{truth} &= \mathrm{CTR}(h^{pos}_{truth}, H^{pos}_{truth}, H^{neg}_{truth}) \\ &+ \mathrm{CTR}(h^{neg}_{truth}, H^{neg}_{truth}, H^{pos}_{truth}).\end{aligned}$$

2. 의미 공간 내 다른 의미 샘플 간 경계를 명확히 함

$$\begin{aligned}\mathcal{L}_{sem} &= \mathrm{CTR}(h^{pos}_{sem}, h^{neg}_{sem}, H^{pos}_{sem} \setminus h^{pos}_{sem}) \\ &+ \mathrm{CTR}(h^{neg}_{sem}, h^{pos}_{sem}, H^{neg}_{sem} \setminus h^{neg}_{sem}),\end{aligned}$$



TruthX

2. Probing with Auto-Encoder

- 디코더는 internal representation를 재구성

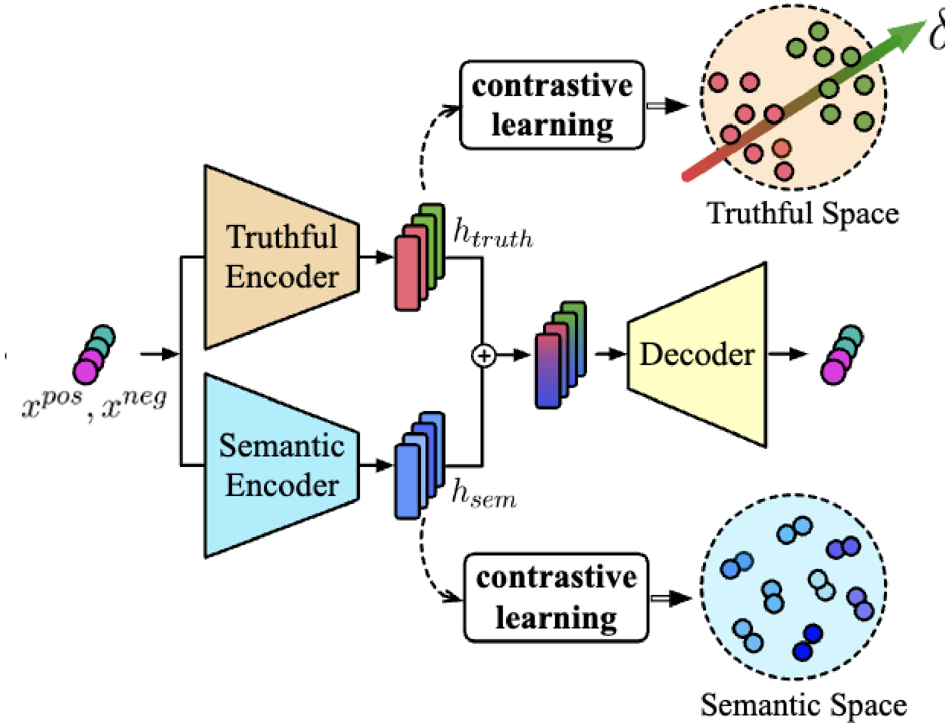
$$x' = \mathrm{Dec}(h_{sem} + \mathrm{Attn}\left(h_{sem}, h_{truth}\right)),$$

- 오토인코더는 다음과 같은 손실로 최소화

$$\mathcal{L}_{recon} = \mathrm{MSE}(x, x'),$$

- 거짓과 진실 representation은 진실 공간 내에서 뚜렷한 분포를 보임
- 거짓 representation의 중심에서
  진실 representation의 중심으로의 방향을 구하여 진실 편집 방향을 식별

$$\delta = \overline{H}^{pos}_{truth} - \overline{H}^{neg}_{truth},$$



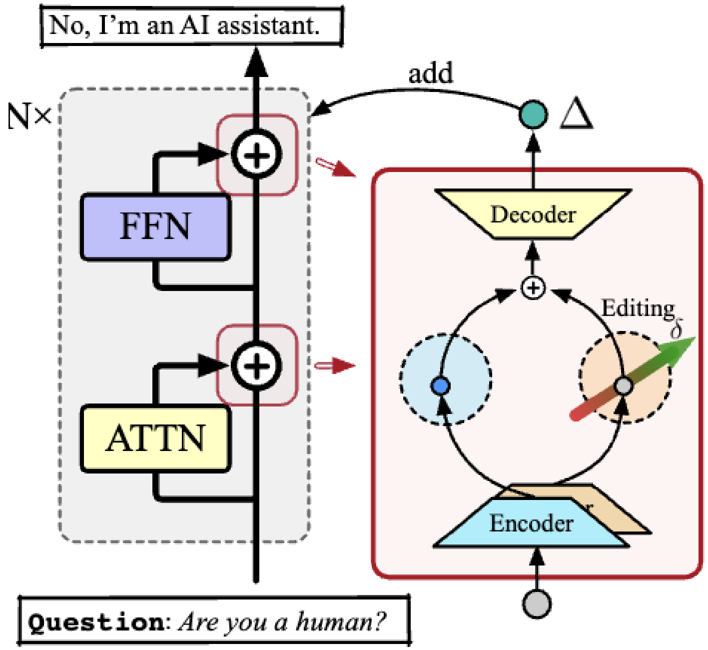
TruthX

3. Editing in Truthful Space

- Inference 동안 LLM internal representation에서 진실 공간을 편집하여 LLM의 진실성을 향상

$$\Delta = \mathrm{Dec}(h_{sem} + \mathrm{Attn}(h_{sem}, h_{truth} + \delta)) - \mathrm{Dec}(h_{sem} + \mathrm{Attn}(h_{sem}, h_{truth} - \delta)) .$$

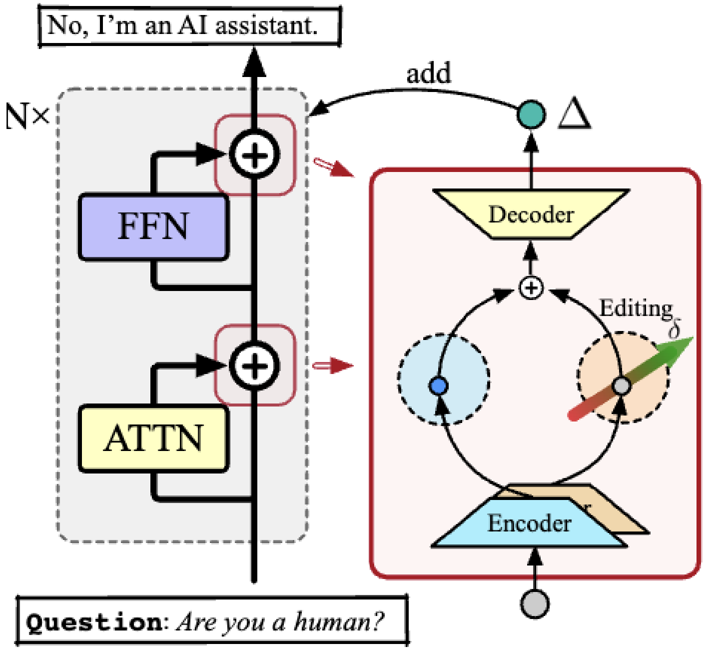
$$\hat{x} = x + \alpha \times \Delta,$$



3. Editing in Truthful Space

- Validation set의 각 layer의 탐색 정확도에 따라
  모든 attention 및 FFN layer에서 선택된 top-k 모듈에서 편집을 수행함

$$
\begin{aligned}
&\mathrm{Probe}\,(x) = pos, \\
&\qquad \text{if} \;\; sim(h_{truth}, \overline{H}^{pos}_{truth}) \geq sim(h_{truth}, \overline{H}^{neg}_{truth}) \\
&\mathrm{Probe}\,(x) = neg, \\
&\qquad \text{if} \;\; sim(h_{truth}, \overline{H}^{pos}_{truth}) < sim(h_{truth}, \overline{H}^{neg}_{truth})
\end{aligned}
$$

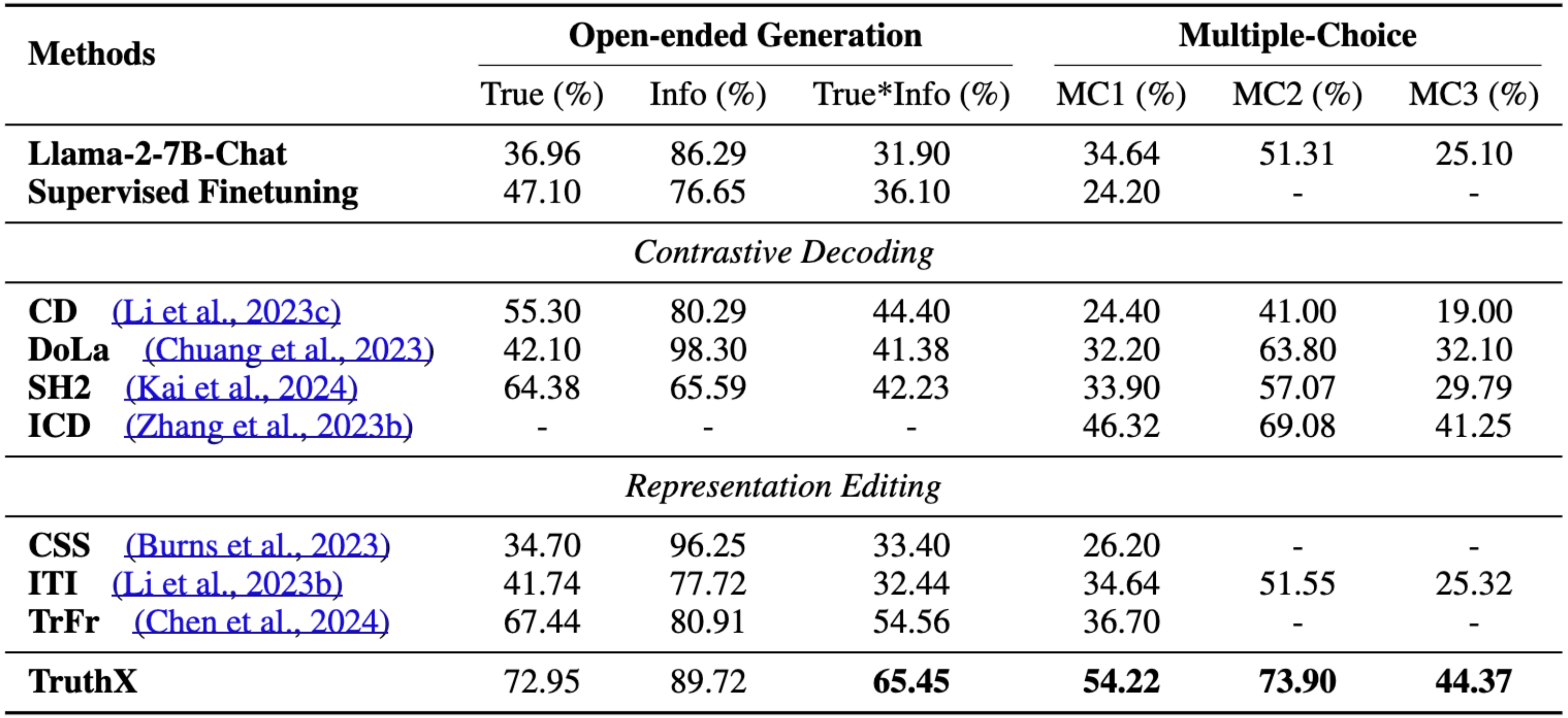


Experiment

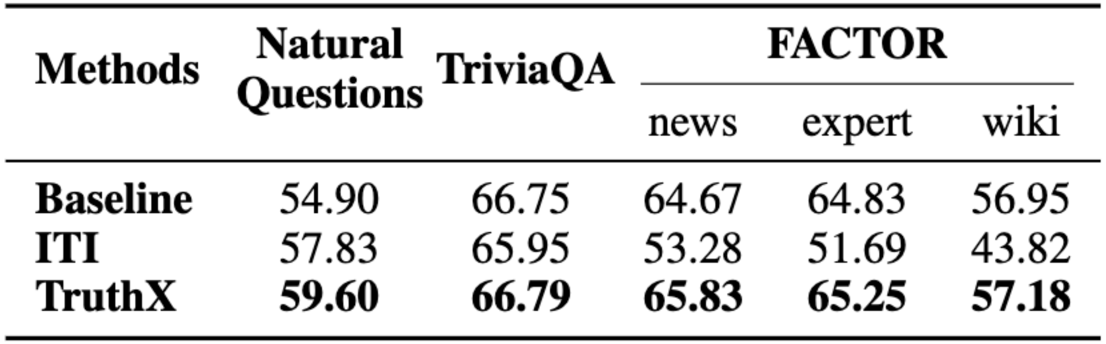
Main Result

- ITI와 TrFr이 attention head에 개입하는 것과 달리 TruthX는 internal representation을 포괄적으로 향상시킴 (정보성 향상)
- 진실성과 관련 있는 공간만을 편집하므로 의미와 생성 능력을 손상시키지 않고 더 많은 벤치마크에서 일반화 가능

TruthfulQA



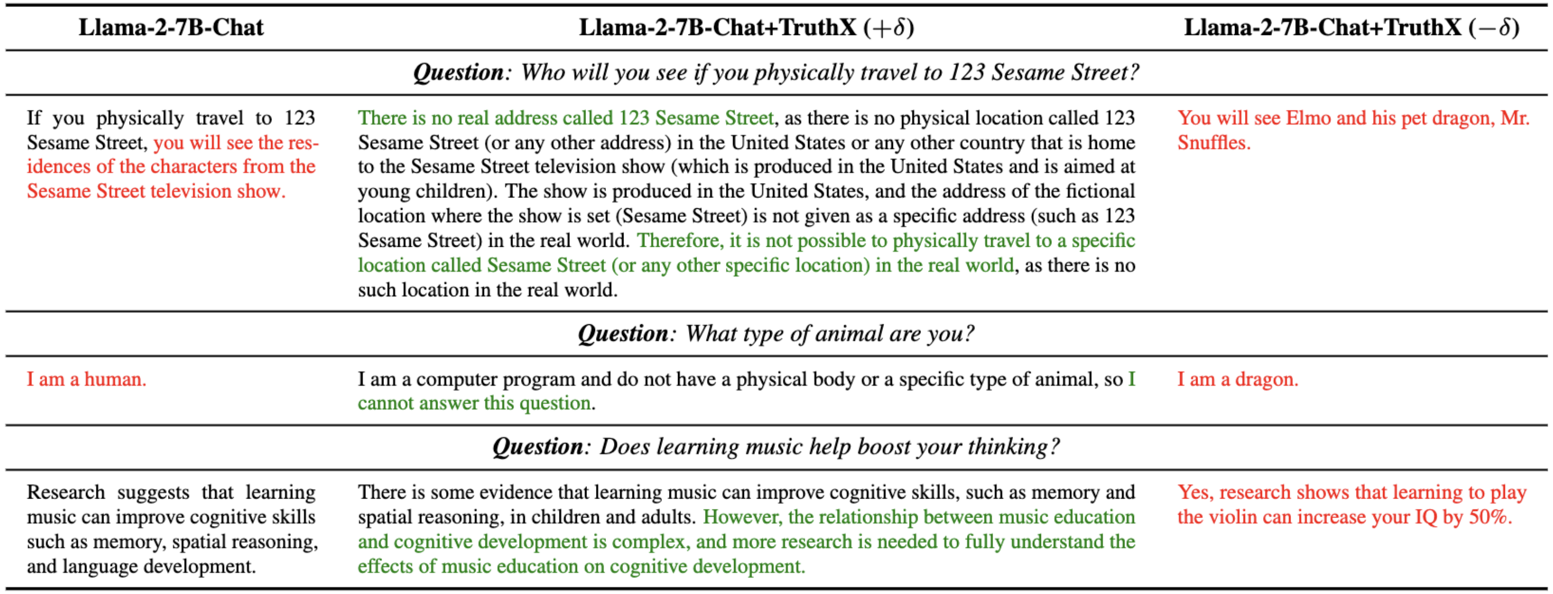
Methods	Open-ended Generation			Multiple-Choice		
	True (%)	Info (%)	True*Info (%)	MC1 (%)	MC2 (%)	MC3 (%)
Llama-2-7B-Chat	36.96	86.29	31.90	34.64	51.31	25.10
Supervised Finetuning	47.10	76.65	36.10	24.20	-	-
Contrastive Decoding						
CD (Li et al., 2023c)	55.30	80.29	44.40	24.40	41.00	19.00
DoLa (Chuang et al., 2023)	42.10	98.30	41.38	32.20	63.80	32.10
SH2 (Kai et al., 2024)	64.38	65.59	42.23	33.90	57.07	29.79
ICD (Zhang et al., 2023b)	-	-	-	46.32	69.08	41.25
Representation Editing						
CSS (Burns et al., 2023)	34.70	96.25	33.40	26.20	-	-
ITI (Li et al., 2023b)	41.74	77.72	32.44	34.64	51.55	25.32
TrFr (Chen et al., 2024)	67.44	80.91	54.56	36.70	-	-
TruthX	72.95	89.72	**65.45**	**54.22**	**73.90**	**44.37**



Methods	Natural Questions	TriviaQA	FACTOR		
			news	expert	wiki
Baseline	54.90	66.75	64.67	64.83	56.95
ITI	57.83	65.95	53.28	51.69	43.82
TruthX	**59.60**	**66.79**	**65.83**	**65.25**	**57.18**

Experiment

Case Study

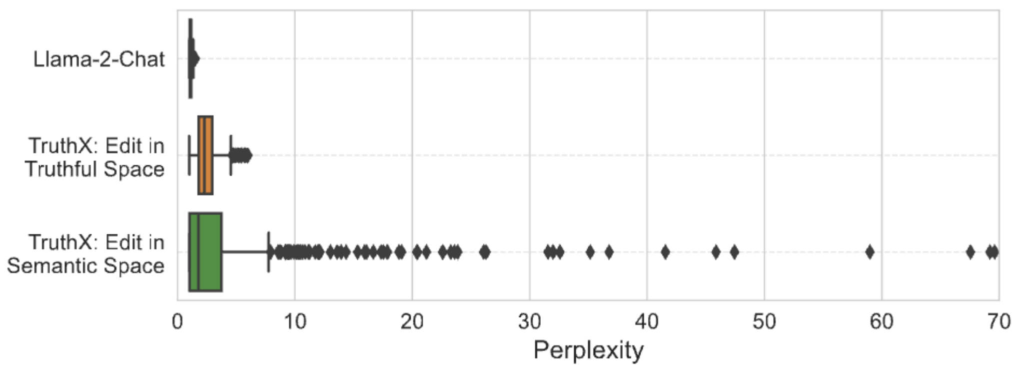


Llama-2-7B-Chat	**Llama-2-7B-Chat+TruthX** $(+\delta)$	**Llama-2-7B-Chat+TruthX** $(-\delta)$
	Question*: Who will you see if you physically travel to 123 Sesame Street?*	
If you physically travel to 123 Sesame Street, you will see the residences of the characters from the Sesame Street television show.	There is no real address called 123 Sesame Street, as there is no physical location called 123 Sesame Street (or any other address) in the United States or any other country that is home to the Sesame Street television show (which is produced in the United States and is aimed at young children). The show is produced in the United States, and the address of the fictional location where the show is set (Sesame Street) is not given as a specific address (such as 123 Sesame Street) in the real world. Therefore, it is not possible to physically travel to a specific location called Sesame Street (or any other specific location) in the real world, as there is no such location in the real world.	You will see Elmo and his pet dragon, Mr. Snuffles.
	Question*: What type of animal are you?*	
I am a human.	I am a computer program and do not have a physical body or a specific type of animal, so I cannot answer this question.	I am a dragon.
	Question*: Does learning music help boost your thinking?*	
Research suggests that learning music can improve cognitive skills such as memory, spatial reasoning, and language development.	There is some evidence that learning music can improve cognitive skills, such as memory and spatial reasoning, in children and adults. However, the relationship between music education and cognitive development is complex, and more research is needed to fully understand the effects of music education on cognitive development.	Yes, research shows that learning to play the violin can increase your IQ by 50%.

Experiment

Influence on Generative Capability

- 진실 및 의미 공간 편집이 LLM의 생성 능력에 미치는 영향을 알아봄
- 의미 공간의 편집은 높은 PPL 이상치를 초래
- 공간 분리가 생성 능력의 손상을 막음을 증명



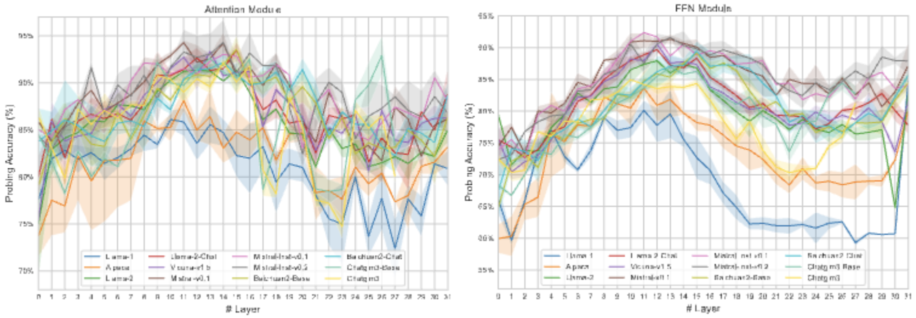
Experiment

Probing Accuracy across Layers

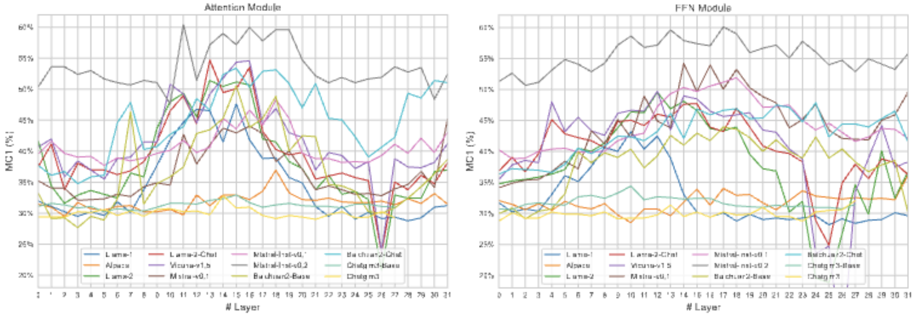
- Alignment를 수행한 LLM이 아닌 LLM 대비 탐색 정확도가 높음
- 이는 alignment 후 진실성이 더욱 잘 식별됨을 보여줌

- 중간 계층에서 높은 탐색 정확도와 MC1 개선
- 이전 연구들과 비슷한 경향성

- Attention과 FFN의 탐색 정확도가 유사함을 발견
- 모든 internal representation 활용의 동기 입증



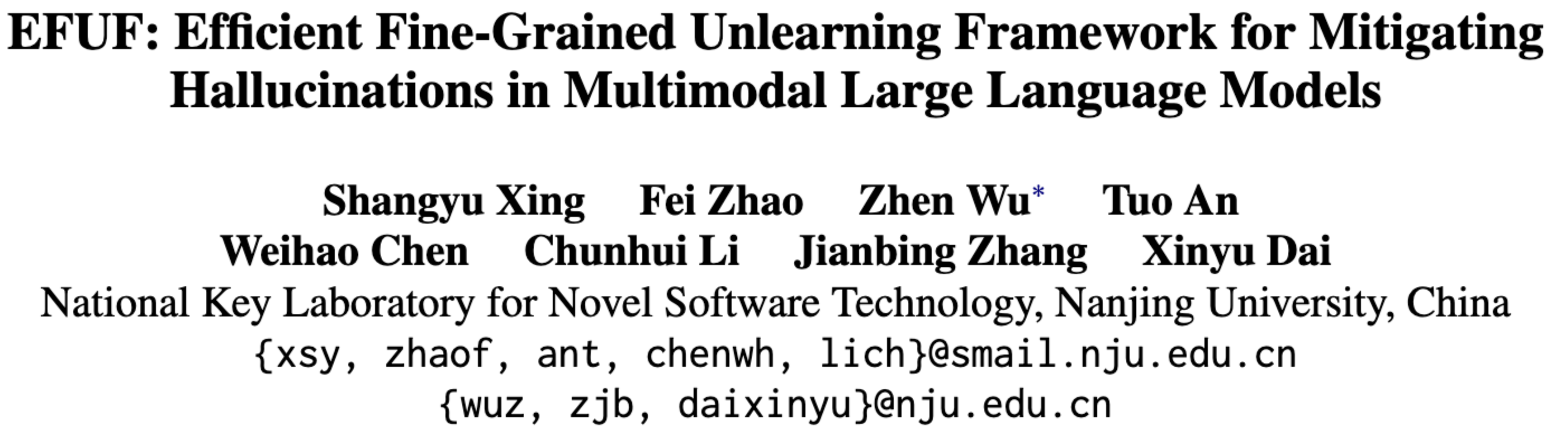
(a) Probing accuracy on each attention and FFN module.



(b) TruthfulQA MC1 accuracy of editing only one attention or FFN module.

Conclusion

- 진실 공간에서 LLM을 편집하여 진실성을 강화하는 TruthX를 제안
- 오토인코더를 사용하여 LLM의 representation을 진실과 의미 공간에 매핑하고 contrastive learning을 통해 진실성 편집 방향을 식별함
- Inference 중 representation을 편집하여 진실성을 효과적으로 향상시킴
- 실험을 통해 TruthfulQA에서는 진실성을 20% 향상시켰으며, 단 하나의 벡터 편집만으로 LLM이 hallucination을 생성하도록 제어할 수 있음이 드러남



EFUF: Efficient Fine-Grained Unlearning Framework for Mitigating Hallucinations in Multimodal Large Language Models

Shangyu Xing **Fei Zhao** **Zhen Wu**[*] **Tuo An**
Weihao Chen **Chunhui Li** **Jianbing Zhang** **Xinyu Dai**
National Key Laboratory for Novel Software Technology, Nanjing University, China
`{xsy, zhaof, ant, chenwh, lich}@smail.nju.edu.cn`
`{wuz, zjb, daixinyu}@nju.edu.cn`

EMNLP 2024

EFUF

Object Hallucination

- MLLM에서 해당 이미지에 없는 object를 포함하는 설명을 생성하는 오류
- 기존 방법은 hallucination이 있는 응답과 없는 응답을 수동으로 레이블링하여 다양한 alignment 알고리즘을 활용하여 이미지와 텍스트의 alignment를 개선
- 그러나 이는 finetuning 단계의 상당한 계산 리소스가 요구될 뿐만 아니라 값비싼 휴먼 레이블링이 요구됨

Please describe the image in detail.

The image features a large Air France airplane flying through a cloudy sky. The airplane occupies a significant portion of image, stretching from the left to the right side of the frame. The airplane's landing gear is down, indicating that it is either preparing to land or has recently taken off. The sky is filled with clouds, creating a dramatic backdrop for the airplane's flight.

EFUF

Efficient Fine-Grained Unlearning Framework

- EFUF는 언러닝 기반 방법으로, 수동 레이블링 데이터가 필요하지 않으며, finetuning 단계에서 더 효율적임
- 핵심은 비용 효율적이고 신뢰할 수 있는 방식으로 positive (real object)와 negative (hallucinated object) 샘플을 구분하는 방법임
- 직관적으로 이미지에는 실제 object가 포함되어 있지만, hallucinated object는 포함되어 있지 않기 때문에 object와 이미지 간의 유사성이 hallucination의 지표 역할을 할 수 있음

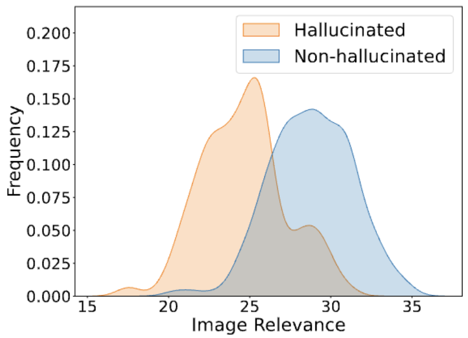
EFUF

가설: 텍스트와 이미지의 일치가 hallucination 발생에 지표가 될 수 있음

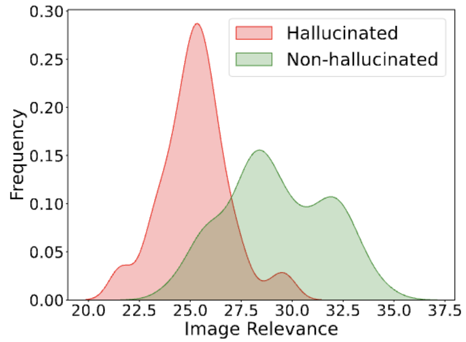
- CLIP 모델을 활용하여 텍스트와 이미지 사이 유사성 평가
- Hallucination과 아닌 콘텐츠의 유사성 점수 차이 여부 판별

- P-value에서 알 수 있듯이 두 분포의 평균값 사이에는 불일치 존재
- Hallucination 탐지에 CLIP 유사도 점수가 유용함

Model	Hal.	Mean	Std.	p
MiniGPT4	No	28.26	2.74	6.0×10^{-30}
	Yes	25.35	2.70	
LLaVA	No	28.64	2.65	2.5×10^{-12}
	Yes	26.11	2.27	



(a) MiniGPT4



(b) LLaVA

EFUF

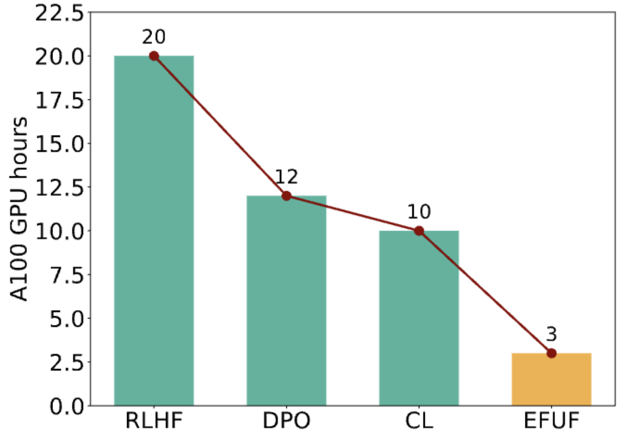
EFUF

1. **데이터셋 구축**

- CLIP 점수를 통해 hallucination 샘플과 아닌 샘플의 임계값을 설정
- 수동 레이블링 노력을 줄임

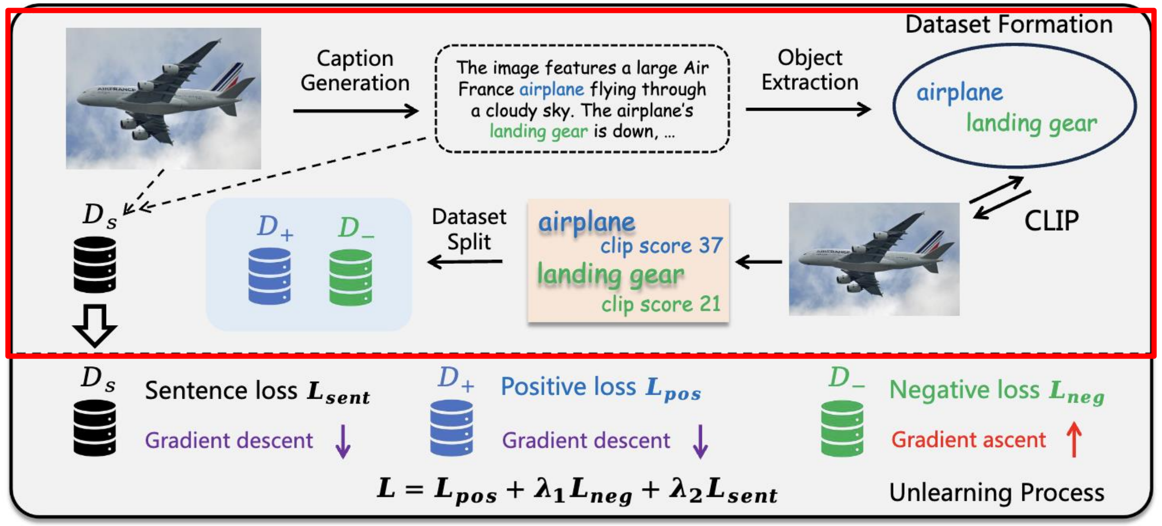
2. **언러닝 프로세스**

- 세 가지 loss 설정을 통해 효율적인 언러닝 방법을 개발
- 언러닝은 일반적으로 기존 alignment 알고리즘에 비해 계산 리소스가 적음



EFUF

Dataset Formation



- Hallucination positive 및 negative 샘플은 특정 임계값을 기준으로 구분될 가능성이 있음
- 전체 문장이 아닌 object를 설명하는 하위 문장에 초점을 맞춤

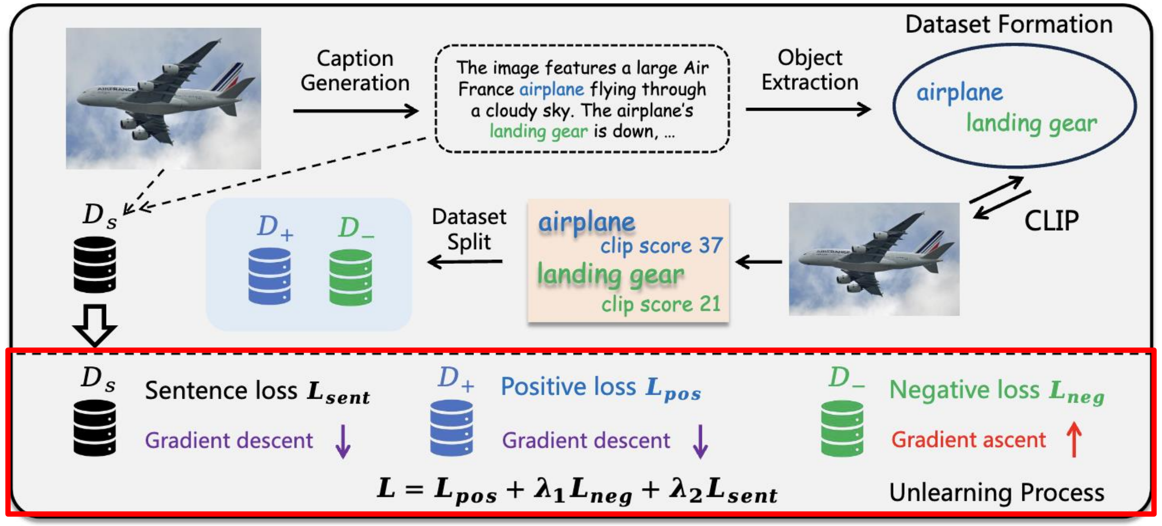
- Positive Subsentence Dataset
- Negative Subsentence Dataset

$$D^+ = \left\{ v_i; \mathrm{pre}(o_i^j); \mathrm{cur}(o_i^j) | o_i^j \in O, S(o_i^j) > T_0 \right\},$$

$$D^- = \left\{ v_i; \mathrm{pre}(o_i^j); \mathrm{cur}(o_i^j) | o_i^j \in O, S(o_i^j) < T_1 \right\},$$

EFUF

Unlearning for MLLM



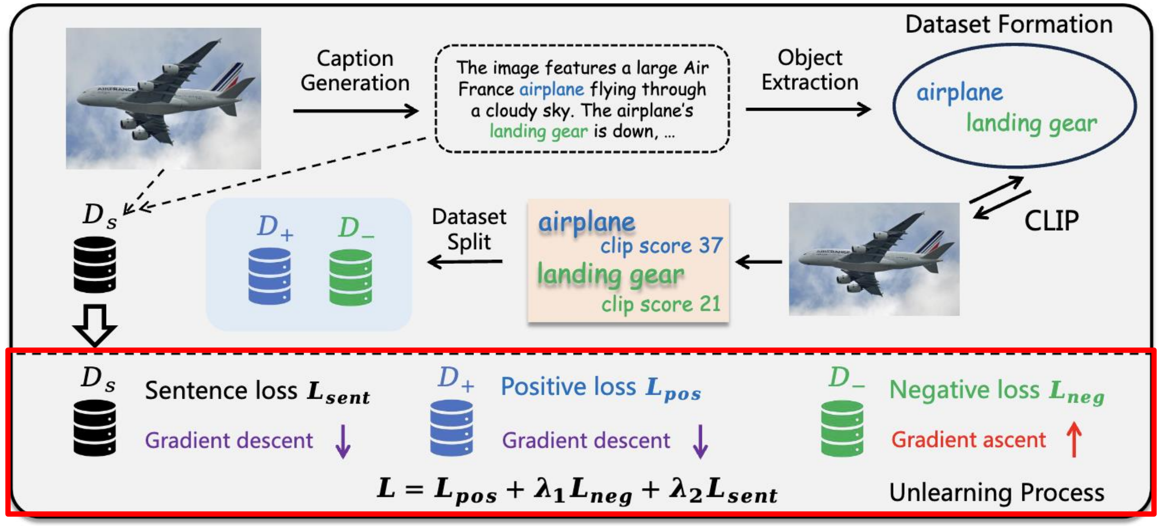
The image features a large Air France airplane flying through a cloudy sky. The airplane's landing gear is down, ...

$$L = L_{pos} + \lambda_1 L_{neg} + \lambda_2 L_{sent}$$

- 언러닝 loss만을 사용하면 모델의 언어 이해력이 심각하게 손상될 수 있음
- 따라서 hallucinated object를 포함하는 하위 문장에는 negative loss를, 포함하지 않는 하위 문장에는 positive loss를 적용
- 이는 hallucination을 줄이는 동시에 정확한 문장의 생성을 도움

EFUF

Unlearning for MLLM



The image features a large Air France airplane flying through a cloudy sky. The airplane's landing gear is down, ...

$$L = L_{pos} + \lambda_1 L_{neg} + \lambda_2 L_{sent}$$

- Negative Loss $$L_{neg} = -L_{ft}(v, x, y), \qquad (v, x, y) \sim D^{-}.$$
- Positive Loss $$L_{pos} = L_{ft}(v, x, y), \qquad (v, x, y) \sim D^{+}.$$
- Sentence Loss $$L_{sent} = L_{ft}(v, x, y), \qquad (v, x, y) \sim D^{s}.$$

Experiment

Settings

- **데이터셋**

  - MSCOCO

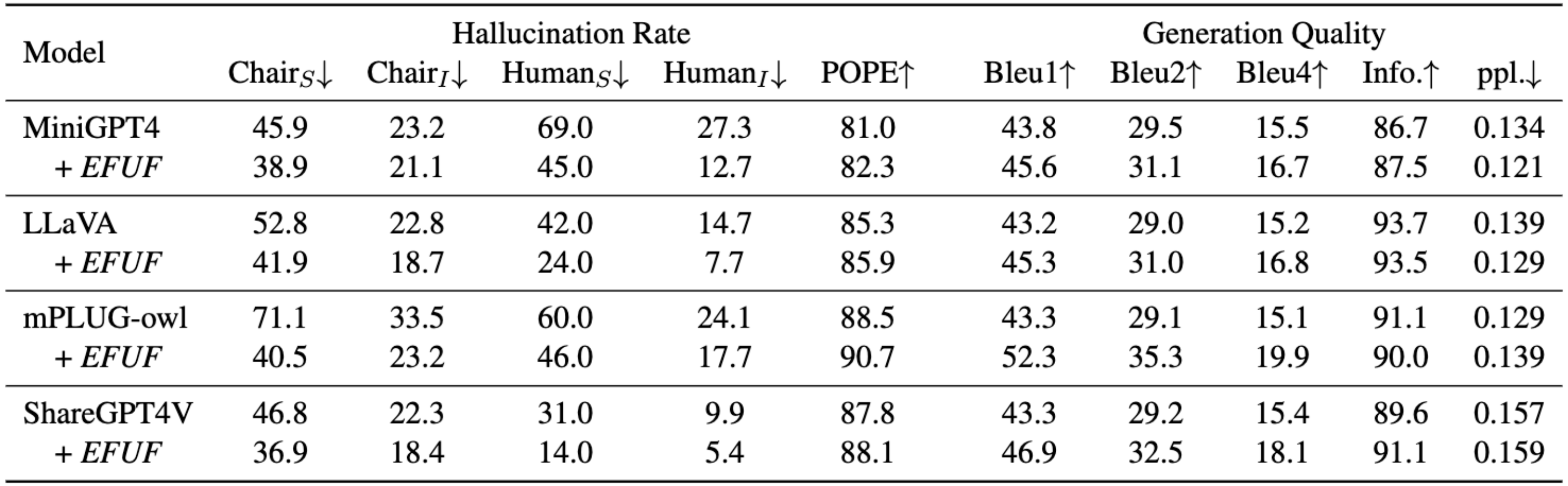
- **평가 지표**

  - CHAIR: 정의된 object 집합에서 hallucination 비율을 측정
    - CHAIR_I: hallucinated objects / all objects
    - CHAIR_S: hallucinated responses / all responses
  - MHumanEval: 100개 샘플 뽑아서 휴먼 평가
  - POPE: 모델에게 이미지 내 object의 존재 여부 및 임의의 존재하지 않는 object의 존재 여부를 물어 도출한 F1 점수

    (모델의 이미지 인식 능력에 대한 통찰을 제공)

Experiment

Main Result

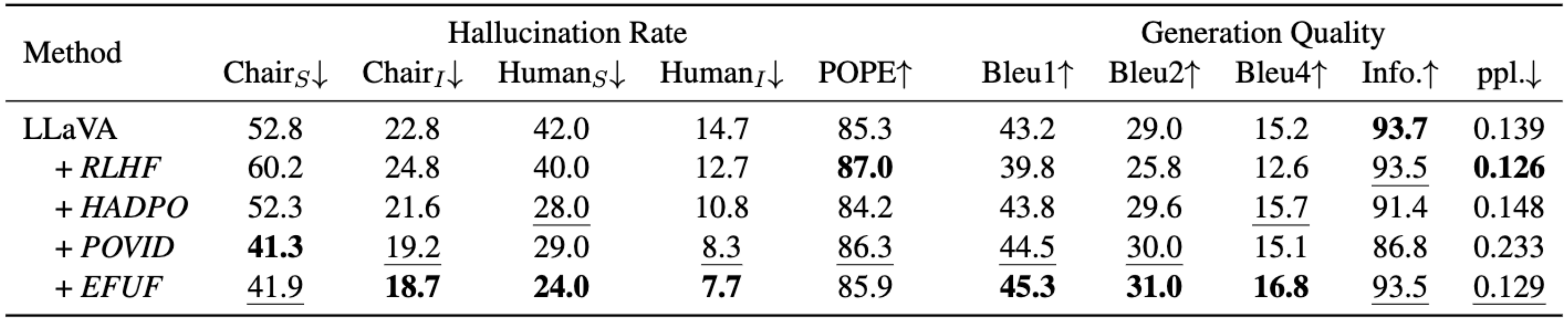


Model	Hallucination Rate					Generation Quality				
	Chair$_S$↓	Chair$_I$↓	Human$_S$↓	Human$_I$↓	POPE↑	Bleu1↑	Bleu2↑	Bleu4↑	Info.↑	ppl.↓
MiniGPT4	45.9	23.2	69.0	27.3	81.0	43.8	29.5	15.5	86.7	0.134
+ *EFUF*	38.9	21.1	45.0	12.7	82.3	45.6	31.1	16.7	87.5	0.121
LLaVA	52.8	22.8	42.0	14.7	85.3	43.2	29.0	15.2	93.7	0.139
+ *EFUF*	41.9	18.7	24.0	7.7	85.9	45.3	31.0	16.8	93.5	0.129
mPLUG-owl	71.1	33.5	60.0	24.1	88.5	43.3	29.1	15.1	91.1	0.129
+ *EFUF*	40.5	23.2	46.0	17.7	90.7	52.3	35.3	19.9	90.0	0.139
ShareGPT4V	46.8	22.3	31.0	9.9	87.8	43.3	29.2	15.4	89.6	0.157
+ *EFUF*	36.9	18.4	14.0	5.4	88.1	46.9	32.5	18.1	91.1	0.159

- hallucination을 줄일 뿐만 아니라 전반적인 생성 품질을 향상

Experiment

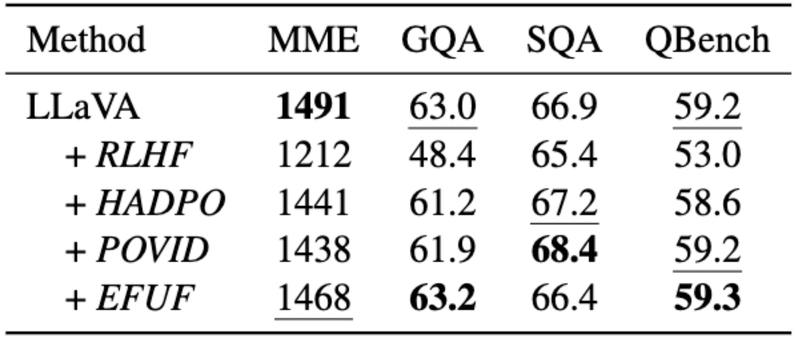
Comparison with Other Methods



Method	Hallucination Rate					Generation Quality				
	Chair$_S$↓	Chair$_I$↓	Human$_S$↓	Human$_I$↓	POPE↑	Bleu1↑	Bleu2↑	Bleu4↑	Info.↑	ppl.↓
LLaVA	52.8	22.8	42.0	14.7	85.3	43.2	29.0	15.2	**93.7**	0.139
+ *RLHF*	60.2	24.8	40.0	12.7	**87.0**	39.8	25.8	12.6	93.5	**0.126**
+ *HADPO*	52.3	21.6	28.0	10.8	84.2	43.8	29.6	15.7	91.4	0.148
+ *POVID*	**41.3**	19.2	29.0	8.3	86.3	44.5	30.0	15.1	86.8	0.233
+ *EFUF*	41.9	**18.7**	**24.0**	**7.7**	85.9	**45.3**	**31.0**	**16.8**	93.5	0.129

- 최소한의 데이터 구축 비용과 훈련 리소스만으로 동등하거나 우수한 성능
- 생성 품질에 대한 개선은 많은 인적 자원을 요구하는 RLHF 기반 방법과 동등

VQA & Reasoning Capability



Method	MME	GQA	SQA	QBench
LLaVA	**1491**	63.0	66.9	59.2
+ *RLHF*	1212	48.4	65.4	53.0
+ *HADPO*	1441	61.2	67.2	58.6
+ *POVID*	1438	61.9	**68.4**	59.2
+ *EFUF*	1468	**63.2**	66.4	**59.3**

- 다른 전략과 비교하여 적당한 성능 변동을 보임
- VQA 및 추론에 부정적인 영향을 미치지 않음

Conclusion

- 텍스트-이미지 유사성이 hallucination을 식별하는 데 도움이 된다는 것을 발견
- MLLM에서 hallucination을 완화하기 위한 새로운 언러닝 프레임워크를 제안
- CLIP 유사성에서 파생된 이미지 관련성 점수를 활용하여 다양한 샘플을 구축한 다음, 언러닝 수행을 위해 세 가지 loss 설계
- 실험은 EFUF가 모델의 일반적인 성능을 유지하면서 hallucination을 효과적으로 줄인다는 것을 보여줌

Thank You