Large Language Model's Factuality

2024 여름방학세미나

NLP&AI 강명훈

Natural Language Processing & Artificial Intelligence

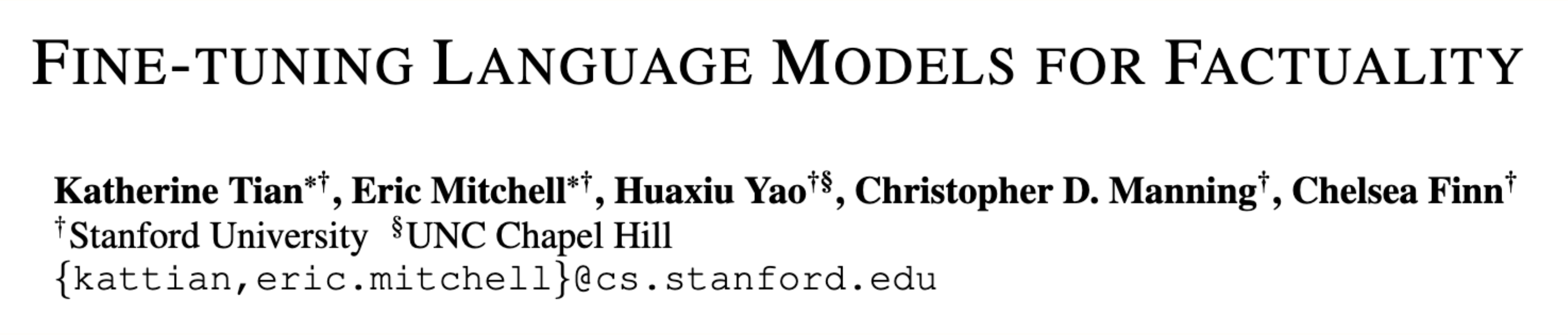
고려대학교 KOREA UNIVERSITY

Theme of the seminar

- How can we fine-tune LLM towards Factuality?
  - 어떻게 LLM이 사실적인 답변을 생성하도록 학습할 수 있을까?

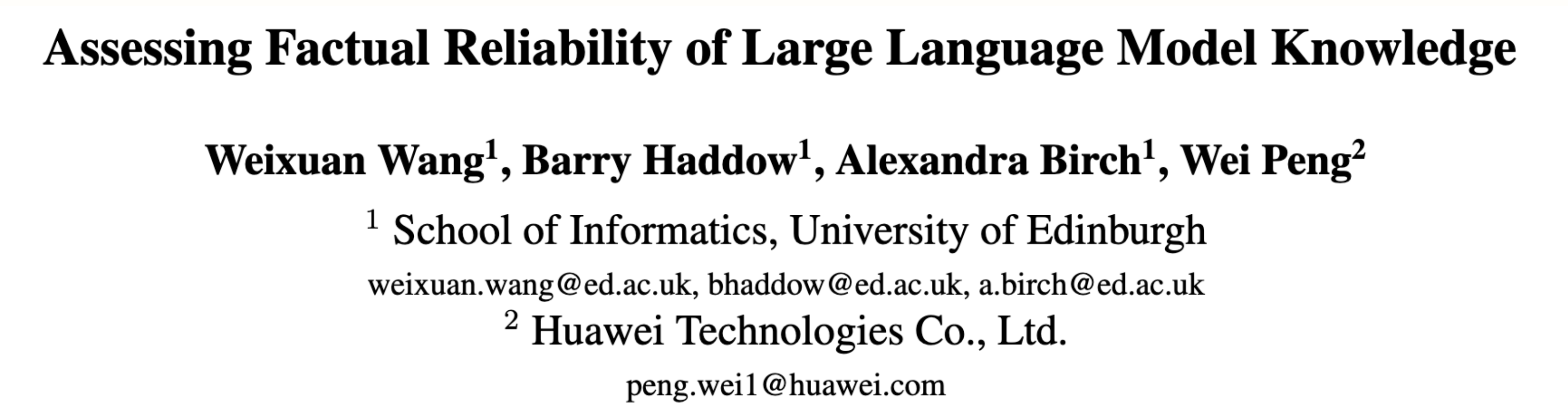
- How can we robustly assess the factuality of the generation output despite prompt & context intervention?
  - Prompt, Context의 변형에 취약한 Accuracy를 대체할 Factuality측정 방법은 없는가?

Papers



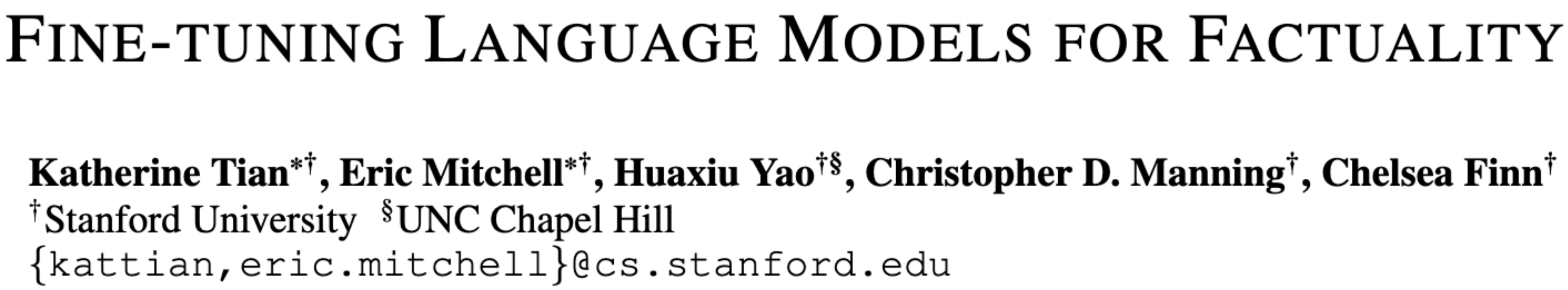
FINE-TUNING LANGUAGE MODELS FOR FACTUALITY

Katherine Tian[*†], Eric Mitchell[*†], Huaxiu Yao[†§], Christopher D. Manning[†], Chelsea Finn[†]
[†]Stanford University [§]UNC Chapel Hill
{kattian,eric.mitchell}@cs.stanford.edu



Assessing Factual Reliability of Large Language Model Knowledge

Weixuan Wang[1], Barry Haddow[1], Alexandra Birch[1], Wei Peng[2]
[1] School of Informatics, University of Edinburgh
weixuan.wang@ed.ac.uk, bhaddow@ed.ac.uk, a.birch@ed.ac.uk
[2] Huawei Technologies Co., Ltd.
peng.wei1@huawei.com



FINE-TUNING LANGUAGE MODELS FOR FACTUALITY

Katherine Tian*[†], Eric Mitchell*[†], Huaxiu Yao[†][§], Christopher D. Manning[†], Chelsea Finn[†]
[†]Stanford University [§]UNC Chapel Hill
{kattian,eric.mitchell}@cs.stanford.edu

ICLR 2024 accepted

(3/3/2, 3/4/2, 3/3/2, 3/4/3)
soundness, presentation, contribution

Problem Setting

Pre-training & RLHF 기반의 학습방법과 Factuality와의 괴리

Pre-training objective 문제

- LLM은 다양한 pre-training 단계에서 Corpus내의 다양한 knowledge를 학습하면서 다양한 task에 engaging한 dialogue를 생성하는 능력을 보유
- 그러나 Maximum likelihood loss 바탕의 학습은 pre-train data distribution에서 벗어나는 token의 생성을 억제하므로 Factual한 생성을 보장할 수 없음

RLHF방법의 문제

- 최근 LLM은 RLHF방법을 이용해 Human preference에 높은 reward를 주는 방향으로 tuning되었음. 이는 Factuality를 높이는 방향으로도 활용됨
- 그러나 Human factual preference, 즉 generation output의 Factuality label을 인간을 통해 얻는 것은 비용이 매우 높음

↓

Human Intervention없이 LLM의 Factuality를 향상할 수 있는 FactTune 제안

FactTune

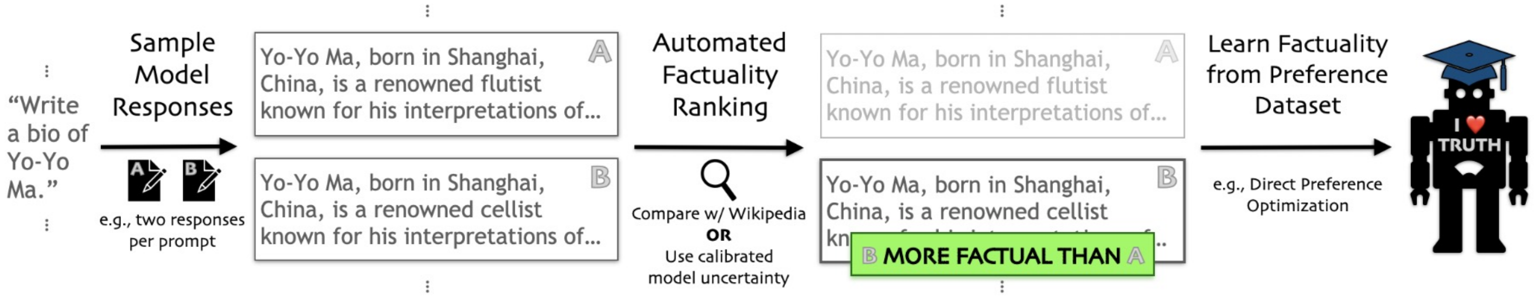

“Write a bio of Yo-Yo Ma.”
Sample Model Responses
e.g., two responses per prompt
Yo-Yo Ma, born in Shanghai, China, is a renowned flutist known for his interpretations of...
A
Yo-Yo Ma, born in Shanghai, China, is a renowned cellist known for his interpretations of...
B
Automated Factuality Ranking
Compare w/ Wikipedia
OR
Use calibrated model uncertainty
B MORE FACTUAL THAN A
Learn Factuality from Preference Dataset
e.g., Direct Preference Optimization
I ❤ TRUTH

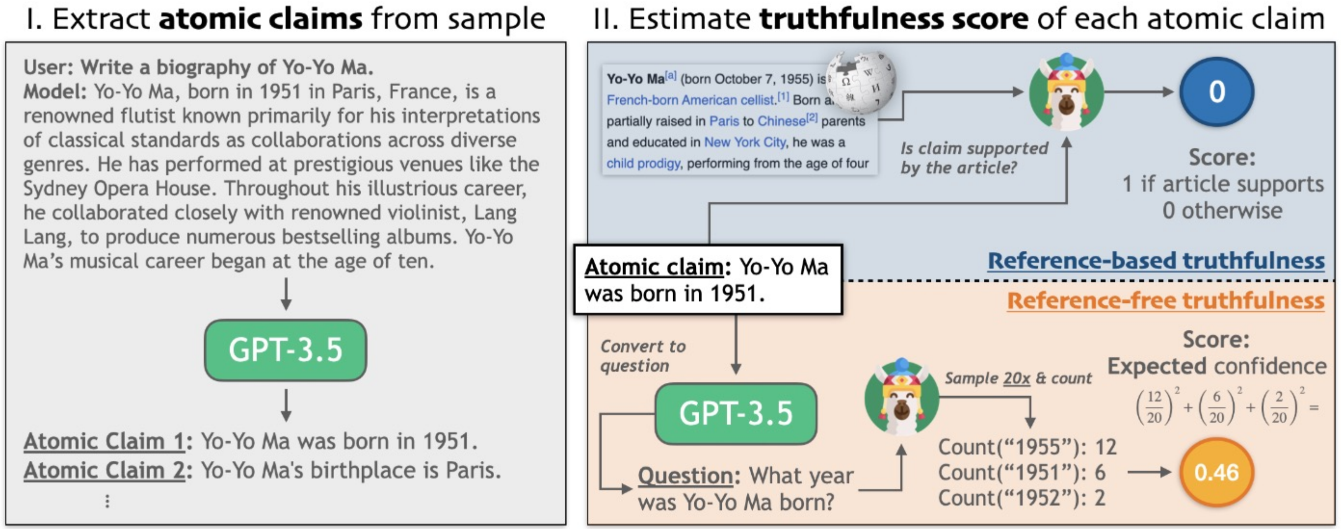
Figure 1: Our approach aims to improve the factuality of language models, specifically focusing on long-form generation (e.g. writing a biography). We develop two different approaches for estimating factuality of a passage (center), each of which allows us to generate a preference dataset (right). We then fine-tune the language model to optimize these factuality preferences (far right).

FactTune: DPO 기반 자동화된 훈련데이터 생성을 통한 Factuality 향상 학습 방법

1) 자동화된 훈련데이터 생성
   - Reference-based method: Generation output이 주어진 wikipedia knowledge를 support하는지를 측정하는 FactScore방식을 차용하여 chosen, rejected set 결정
   - Reference-free method: Generation output sampling 기반 모델의 생성 확률을 이용한 confidence score를 이용하여 chosen, rejected set 결정

2) DPO 활용 Factual preference학습

FactTune



I. Extract **atomic claims** from sample

User: Write a biography of Yo-Yo Ma.
Model: Yo-Yo Ma, born in 1951 in Paris, France, is a renowned flutist known primarily for his interpretations of classical standards as collaborations across diverse genres. He has performed at prestigious venues like the Sydney Opera House. Throughout his illustrious career, he collaborated closely with renowned violinist, Lang Lang, to produce numerous bestselling albums. Yo-Yo Ma's musical career began at the age of ten.

GPT-3.5

Atomic Claim 1: Yo-Yo Ma was born in 1951.
Atomic Claim 2: Yo-Yo Ma's birthplace is Paris.
⋮

II. Estimate **truthfulness score** of each atomic claim

Yo-Yo Ma[a] (born October 7, 1955) is a French-born American cellist.[1] Born and partially raised in Paris to Chinese[2] parents and educated in New York City, he was a child prodigy, performing from the age of four

Is claim supported by the article?

0

Score:
1 if article supports
0 otherwise

Atomic claim: Yo-Yo Ma was born in 1951.

Reference-based truthfulness

Reference-free truthfulness

Convert to question

GPT-3.5

Question: What year was Yo-Yo Ma born?

Sample 20x & count

Score:
Expected confidence

$\left(\frac{12}{20}\right)^2 + \left(\frac{6}{20}\right)^2 + \left(\frac{2}{20}\right)^2 =$

Count("1955"): 12
Count("1951"): 6
Count("1952"): 2

0.46

Figure 2: We estimate the factuality of a generation by first extracting claims (left) and then evaluating each claims' truthfulness (right). For the latter, we consider: a *reference-based* (top right) method that uses a fine-tuned Llama model to check if the fact is supported by Wikipedia (Min et al., 2023) and a *reference-free* (bottom right) method that uses the model's confidence in its most likely answer to estimate its truthfulness.

- Dataset generation procedure

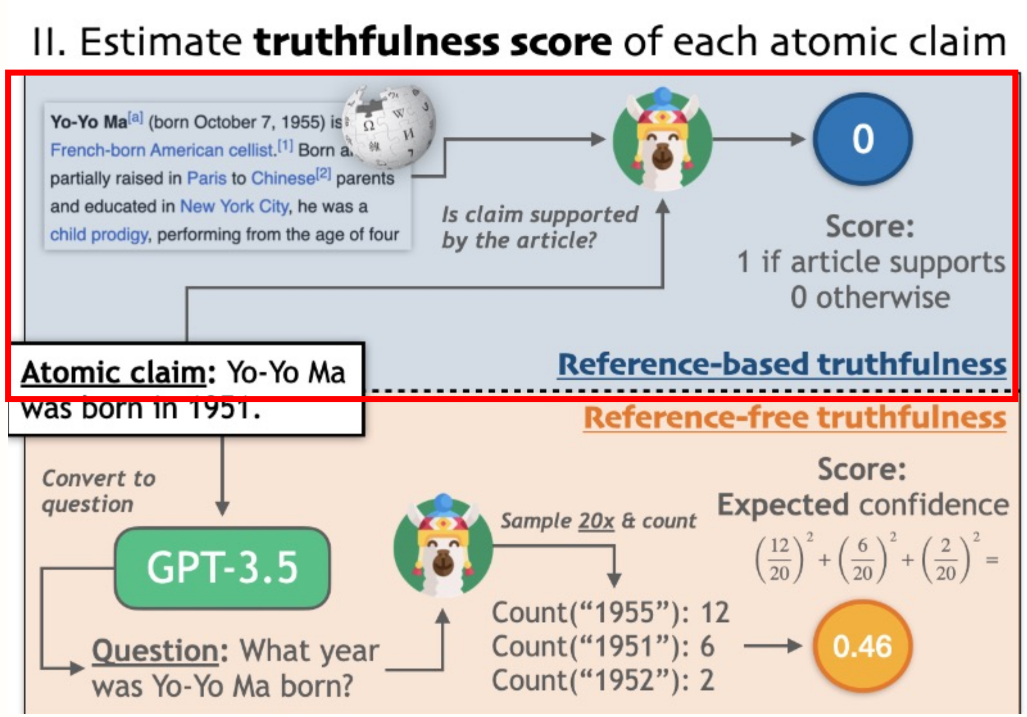
1. Extract atomic claims from sample

- ChatGPT(GPT-3.5)를 이용해 LLM의 생성 결과들을 각각 m개의 atomic claim으로 분해

2. Estimate truthfulness score of each atomic claim

- 분해된 개별 atomic claim의 진실성(truthfulness)를 reference-based, reference-free 방법을 활용하여 측정
- n개 atomic claim의 truthfulness score를 합산하여서 두 생성결과의 chosen, rejected label을 할당

FactTune


II. Estimate **truthfulness score** of each atomic claim
Yo-Yo Ma[a] (born October 7, 1955) is a French-born American cellist.[1] Born a... partially raised in Paris to Chinese[2] parents and educated in New York City, he was a child prodigy, performing from the age of four
Is claim supported by the article?
0
Score:
1 if article supports
0 otherwise
Atomic claim: Yo-Yo Ma was born in 1951.
Reference-based truthfulness
Reference-free truthfulness
Convert to question
GPT-3.5
Question: What year was Yo-Yo Ma born?
Sample 20x & count
Count("1955"): 12
Count("1951"): 6
Count("1952"): 2
Score:
Expected confidence
$\left(\frac{12}{20}\right)^2 + \left(\frac{6}{20}\right)^2 + \left(\frac{2}{20}\right)^2 =$
0.46

- **Reference-based truthfulness**

- Factscore 방식을 차용하여서 진행
- 분해된 개별 atomic claim이 mapping된 Wikipedia passage를 support하는지를 학습된 LLaMA1-7b 모델로 이진분류 (NLI 방식)
- 만약 support할 경우 1, 그렇지 않을 경우 0으로 할당
- 따라서 어떤 generation output의 truthfulness score는 추출된 총 atomic claim개수 대비 truthfulness score 합의 비율

단점

- 적절한 reference를 mapping할 수 있는 retrieval 필요
- atomic claim과 reference의 matching을 판단하는 적절한 판별 모델 필요

FActScore

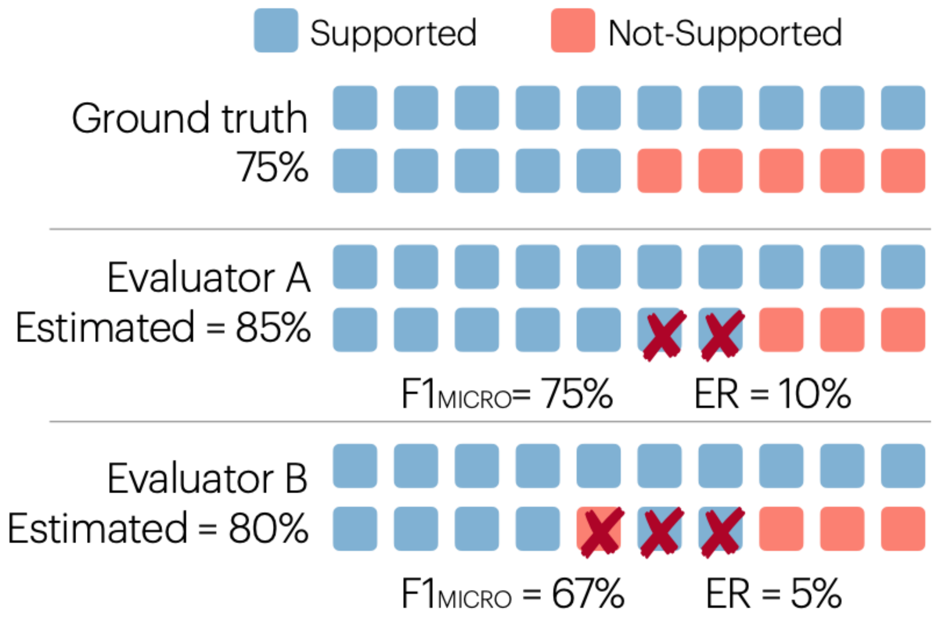


Figure 4: A case in which $F1_{MICRO}$ and Error Rate (ER) rank two evaluators differently. Evaluator A is better in $F1_{MICRO}$, and Evaluator B is better in ER.

- ## Estimating Factscore

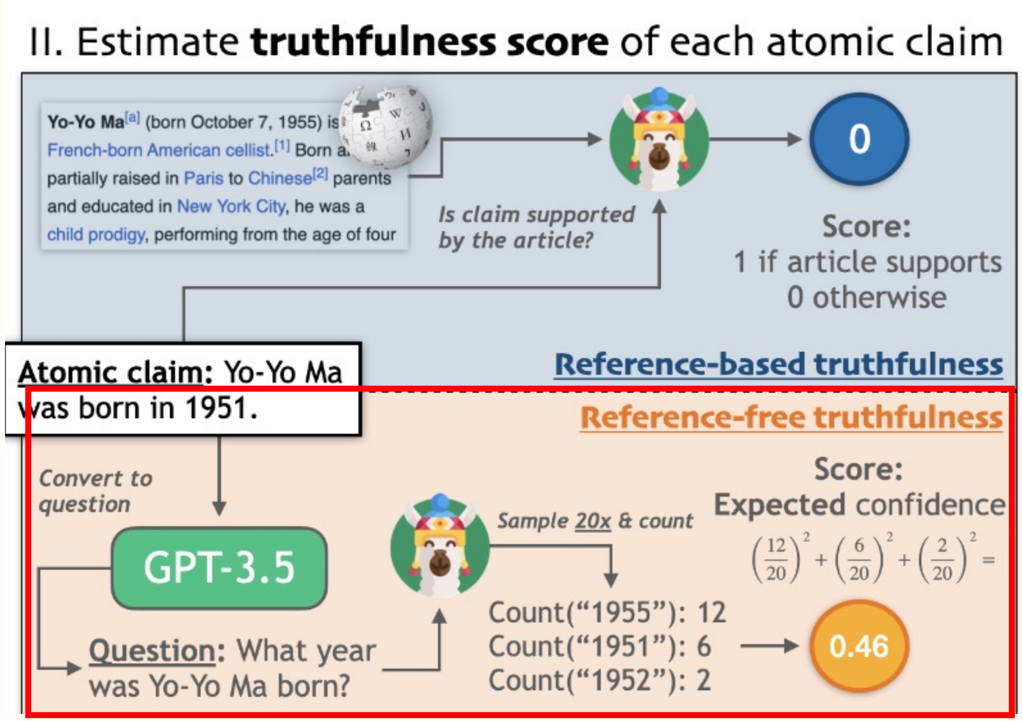
Inst-LLaMA, ChatGPT를 이용하여서 자동화된 Factscore 측정을 시도

Ground truth Factscore와 Estimated Factscore의 차이 ER(error rate)을 기준으로 가장 타당한 방법 탐색

4개의 prompting 방법으로 Factscore 측정 방식 다변화

- No-context LM: <atomic-fact> True or False?
- Retrieve→LM: GTR을 이용해 top-5 passage + <atomic-fact> 를 prompt로 사용하여 True, False 판별
- NP: MLM 기반의 모델을 사용하여 <atomic-fact>의 평균 masking 복원 확률을 이용하여 True, False 판별
- Retrieve→LM + NP: Retrieve→LM과 NP의 판별 방식을 결합하되 두 방식 각각으로 판별했을 때만 True로 예측

FactTune



II. Estimate **truthfulness score** of each atomic claim

Yo-Yo Ma[a] (born October 7, 1955) is French-born American cellist.[1] Born a partially raised in Paris to Chinese[2] parents and educated in New York City, he was a child prodigy, performing from the age of four

Is claim supported by the article?

0

Score:
1 if article supports
0 otherwise

Reference-based truthfulness

Atomic claim: Yo-Yo Ma was born in 1951.

Reference-free truthfulness

Convert to question

GPT-3.5

Question: What year was Yo-Yo Ma born?

Sample 20x & count

Count("1955"): 12
Count("1951"): 6
Count("1952"): 2

Score:
Expected confidence

$$\left(\frac{12}{20}\right)^2 + \left(\frac{6}{20}\right)^2 + \left(\frac{2}{20}\right)^2 =$$

0.46

- Reference-free truthfulness

- 분해된 개별 atomic claim을 GPT-3.5를 이용해 question으로 변환
- LLaMA1-7b모델에게 QA를 실시, 20번 sampling하여 총 20개의 answer를 도출
- 유사한 정답이 있을 수 있으므로 GPT-3.5혹은 Heuristic string match로 유사한 정답을 하나의 정답군으로 묶음
- (전체 sample 개수 대비 각 정답군의 비율)^2 값의 합으로 truthfulness score를 산출

FactTune

- Factuality Tuning: Putting it all together

  - 최종 데이터셋 생성 방안은 먼저 주어진 1개의 prompt에 대하여 n개의 sample을 생성
  - 생성된 n개의 sample들 간 2개의 pair를 만드는 조합 $\binom{n}{2}$ 구성, 각 조합 내에서

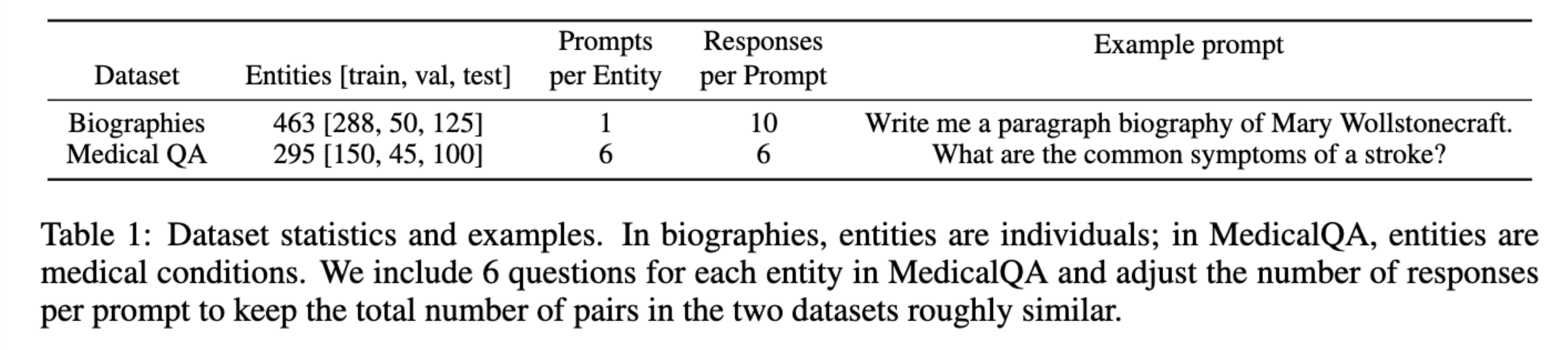
    chosen은 truthfulness score가 더 높은 sample로 할당, rejected는 반대의 sample로 할당
  - 총 m개의 prompt에 대해서 수행하므로 최종 생성 훈련데이터 수는 $m\binom{n}{2} - k$, 이 때 k는 truthfulness

    score가 동률인 경우에 해당하여 제외하는 경우를 의미
  - 최종 생성 훈련데이터를 DPO framework를 활용하여 Fine-tuning 진행

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

Experiment Setup

- Dataset generation



Dataset	Entities [train, val, test]	Prompts per Entity	Responses per Prompt	Example prompt
Biographies	463 [288, 50, 125]	1	10	Write me a paragraph biography of Mary Wollstonecraft.
Medical QA	295 [150, 45, 100]	6	6	What are the common symptoms of a stroke?

Table 1: Dataset statistics and examples. In biographies, entities are individuals; in MedicalQA, entities are medical conditions. We include 6 questions for each entity in MedicalQA and adjust the number of responses per prompt to keep the total number of pairs in the two datasets roughly similar.

- Reference-based truthfulness score를 측정할 수 있어야 하므로 Reference가 존재하는 Biographies, Medical QA(Wikipedia page) 데이터셋을 활용함
- Seed prompt는 GPT-3.5개 생성했으며 sample은 LLaMA1-7b모델의 few-shot prompt를 주어서 추출
- 위 데이터셋을 기반으로 3개의 모델을 학습
  1) FactTune-FC: Reference-based truthfulness score를 이용하여 DPO training set 구축 및 학습
  2) FactTune-MC: Reference-free truthfulness score의 max값을 이용하여 DPO training set 구축 및 학습
  3) FactTune-EC: Reference-free truthfulness score의 Estimation값을 이용하여 DPO training set 구축 및 학습

Experiment Setup

- 사용 Model & Evaluation setting

Baseline

- DOLA: Decoding by Contrasting Layers

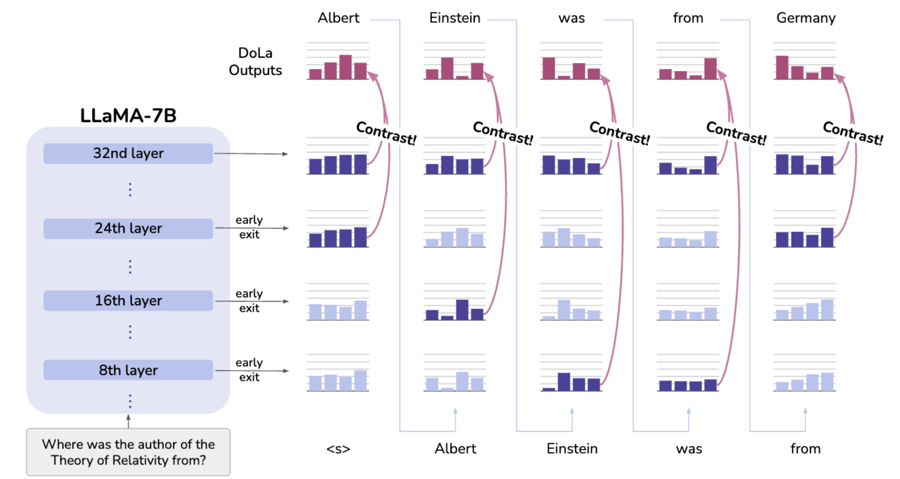


Figure 3: The illustration of how dynamic premature layer selection works.

output next-word probability is obtained from the difference in logits between a higher layer versus a lower layer

Baseline

- ITI: Inference-time intervention

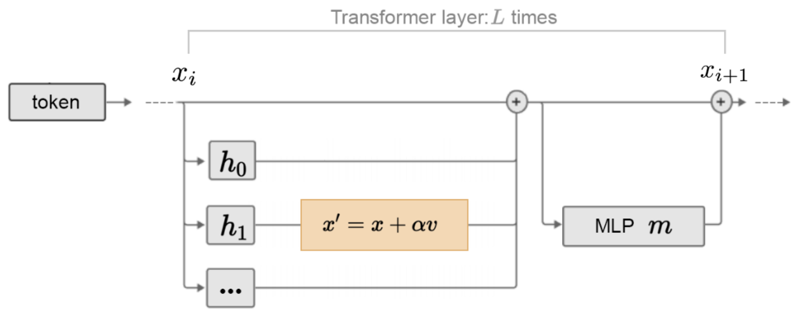
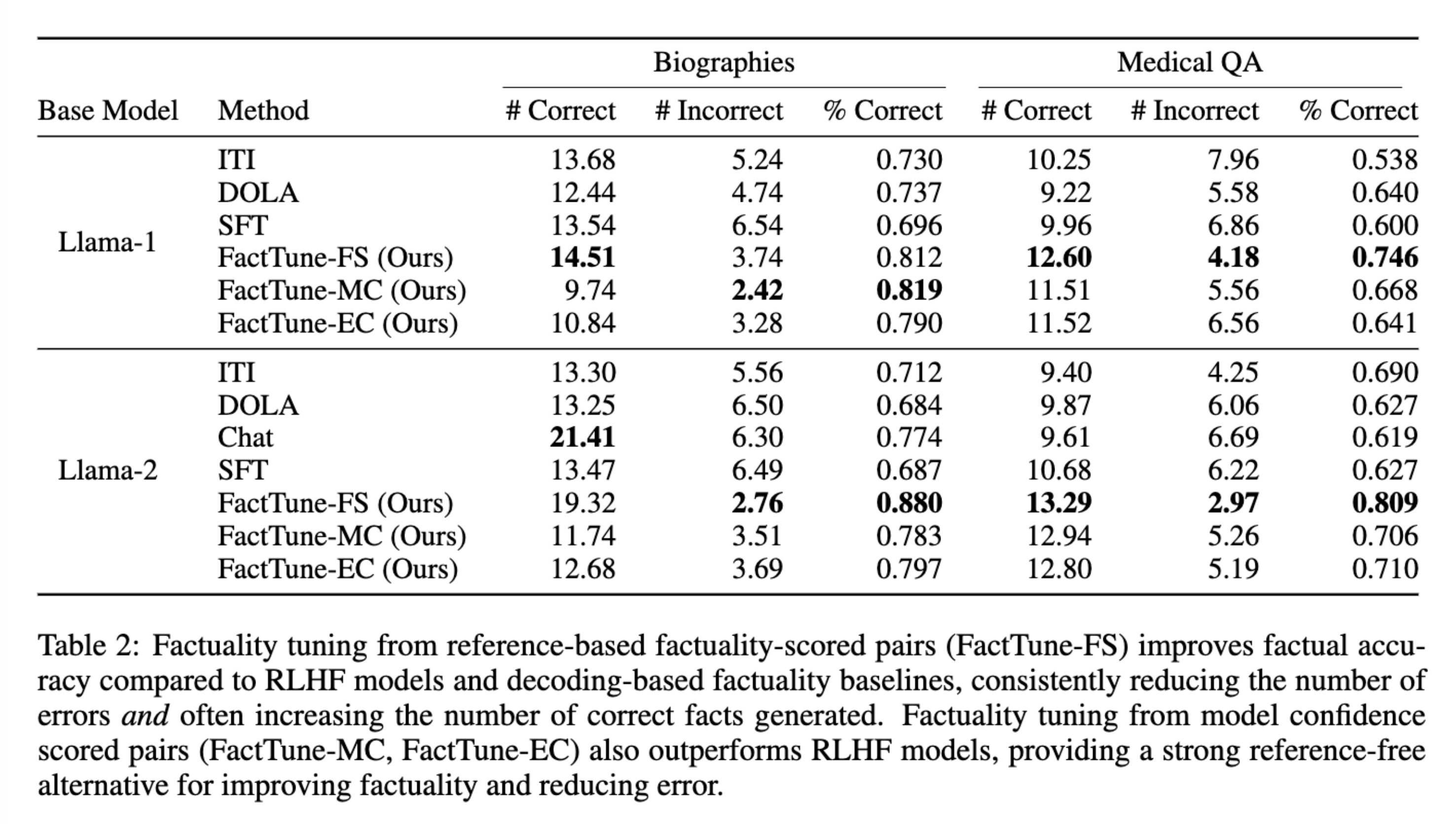


Figure 3: A sketch of the computation on the last token of a transformer with inference-time intervention (ITI) highlighted.

LM의 Truthfulness에 관여하는 일부 layer의 Representation에 Multi-head attention head representation activation 값을 추가하여 inference를 진행하는 방법

Main result

- Quantitative Result



		Biographies			Medical QA		
Base Model	Method	# Correct	# Incorrect	% Correct	# Correct	# Incorrect	% Correct
Llama-1	ITI	13.68	5.24	0.730	10.25	7.96	0.538
	DOLA	12.44	4.74	0.737	9.22	5.58	0.640
	SFT	13.54	6.54	0.696	9.96	6.86	0.600
	FactTune-FS (Ours)	**14.51**	3.74	0.812	**12.60**	**4.18**	**0.746**
	FactTune-MC (Ours)	9.74	**2.42**	**0.819**	11.51	5.56	0.668
	FactTune-EC (Ours)	10.84	3.28	0.790	11.52	6.56	0.641
Llama-2	ITI	13.30	5.56	0.712	9.40	4.25	0.690
	DOLA	13.25	6.50	0.684	9.87	6.06	0.627
	Chat	**21.41**	6.30	0.774	9.61	6.69	0.619
	SFT	13.47	6.49	0.687	10.68	6.22	0.627
	FactTune-FS (Ours)	19.32	**2.76**	**0.880**	**13.29**	**2.97**	**0.809**
	FactTune-MC (Ours)	11.74	3.51	0.783	12.94	5.26	0.706
	FactTune-EC (Ours)	12.68	3.69	0.797	12.80	5.19	0.710

Table 2: Factuality tuning from reference-based factuality-scored pairs (FactTune-FS) improves factual accuracy compared to RLHF models and decoding-based factuality baselines, consistently reducing the number of errors *and* often increasing the number of correct facts generated. Factuality tuning from model confidence scored pairs (FactTune-MC, FactTune-EC) also outperforms RLHF models, providing a strong reference-free alternative for improving factuality and reducing error.

FactTune의 Factuality 향상 확인

- # Correct, % Correct는 높을수록, # Incorrect는 낮을수록 좋음

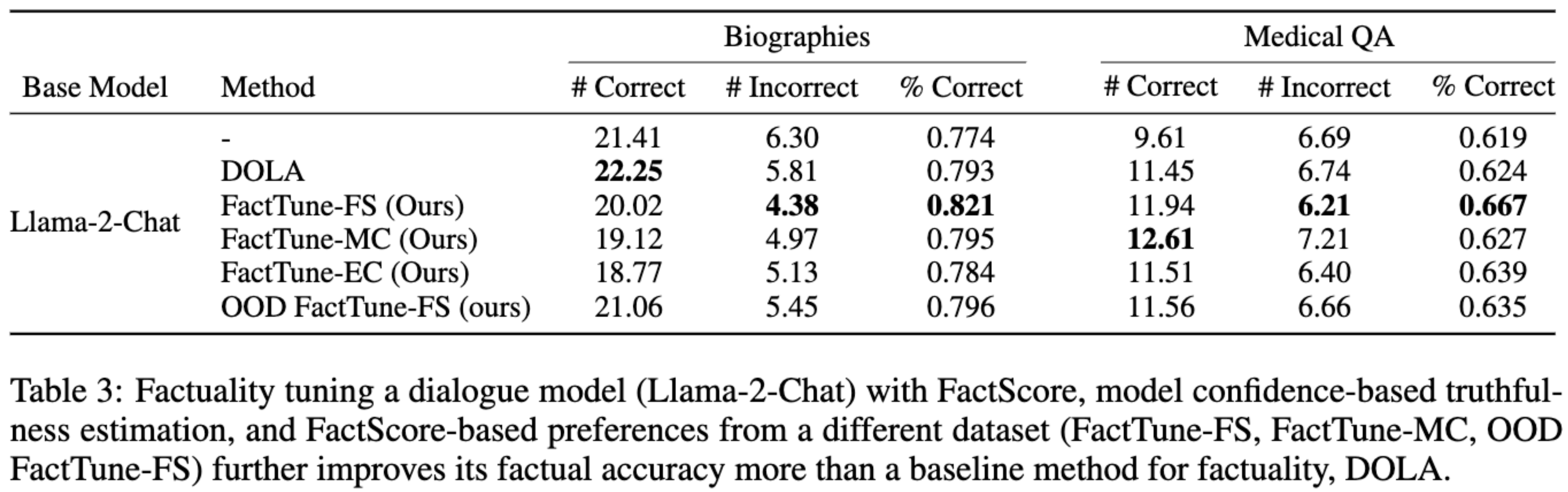
- SFT 방법 대비 # Correct, % Correct는 증가시키면서 # Incorrect는 하락

- 직접적인 비교대상인 RLHF학습 기반 LLama-2 Chat 모델 대비 #Incorrect의 수를 줄이면서 %Correct를 늘리는 것을 볼 수 있음

- ITI, DOLA는 기존 SFT대비 Domain에 따라 Factuality 향상의 효과가 상이한 것을 확인

Main result

- Quantitative Result



		Biographies			Medical QA		
Base Model	Method	# Correct	# Incorrect	% Correct	# Correct	# Incorrect	% Correct
Llama-2-Chat	-	21.41	6.30	0.774	9.61	6.69	0.619
	DOLA	**22.25**	5.81	0.793	11.45	6.74	0.624
	FactTune-FS (Ours)	20.02	**4.38**	**0.821**	11.94	**6.21**	**0.667**
	FactTune-MC (Ours)	19.12	4.97	0.795	**12.61**	7.21	0.627
	FactTune-EC (Ours)	18.77	5.13	0.784	11.51	6.40	0.639
	OOD FactTune-FS (ours)	21.06	5.45	0.796	11.56	6.66	0.635

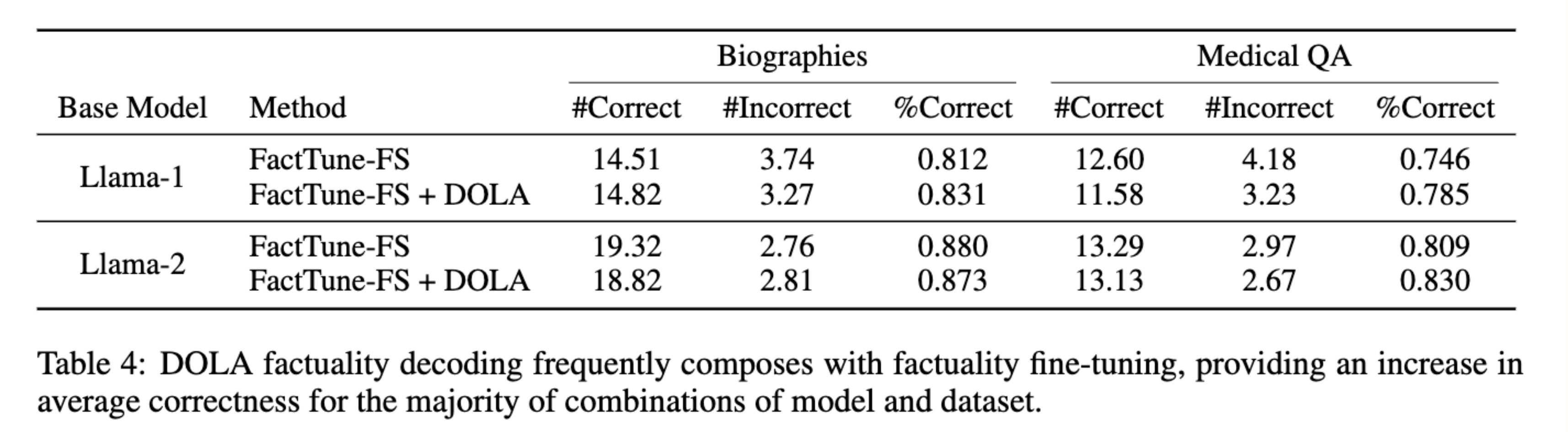
Table 3: Factuality tuning a dialogue model (Llama-2-Chat) with FactScore, model confidence-based truthfulness estimation, and FactScore-based preferences from a different dataset (FactTune-FS, FactTune-MC, OOD FactTune-FS) further improves its factual accuracy more than a baseline method for factuality, DOLA.

- RLHF학습 기반 LLaMA2-Chat 모델에 FactTune을 적용했을 때의Factuality 향상을 측정

- FactTune은 기존 RLHF대비 Correct생성물의 수를 크게 줄이지 않으면서 Incorrect의 수는 크게 줄여 %Correct의 향상이 있는 것을 볼 수 있음

Further Analysis

- DOLA+FactTune의 효과 측정



		Biographies			Medical QA		
Base Model	Method	#Correct	#Incorrect	%Correct	#Correct	#Incorrect	%Correct
Llama-1	FactTune-FS	14.51	3.74	0.812	12.60	4.18	0.746
	FactTune-FS + DOLA	14.82	3.27	0.831	11.58	3.23	0.785
Llama-2	FactTune-FS	19.32	2.76	0.880	13.29	2.97	0.809
	FactTune-FS + DOLA	18.82	2.81	0.873	13.13	2.67	0.830

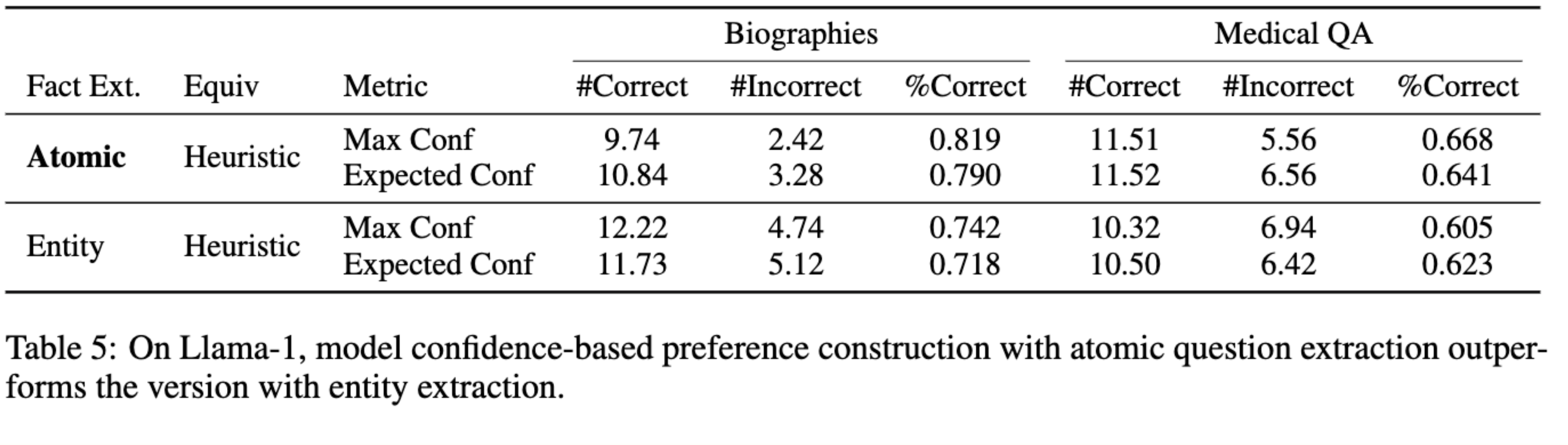
Table 4: DOLA factuality decoding frequently composes with factuality fine-tuning, providing an increase in average correctness for the majority of combinations of model and dataset.

- LLaMA-2의 Biographies실험을 제외하고 모든 면에서 FactTune에 DOLA를 적용했을 때 성능 향상이 존재

- 타 Decoding strategy의 투입에 따른 성능변화는 미확인

Further Analysis

- Ablation study

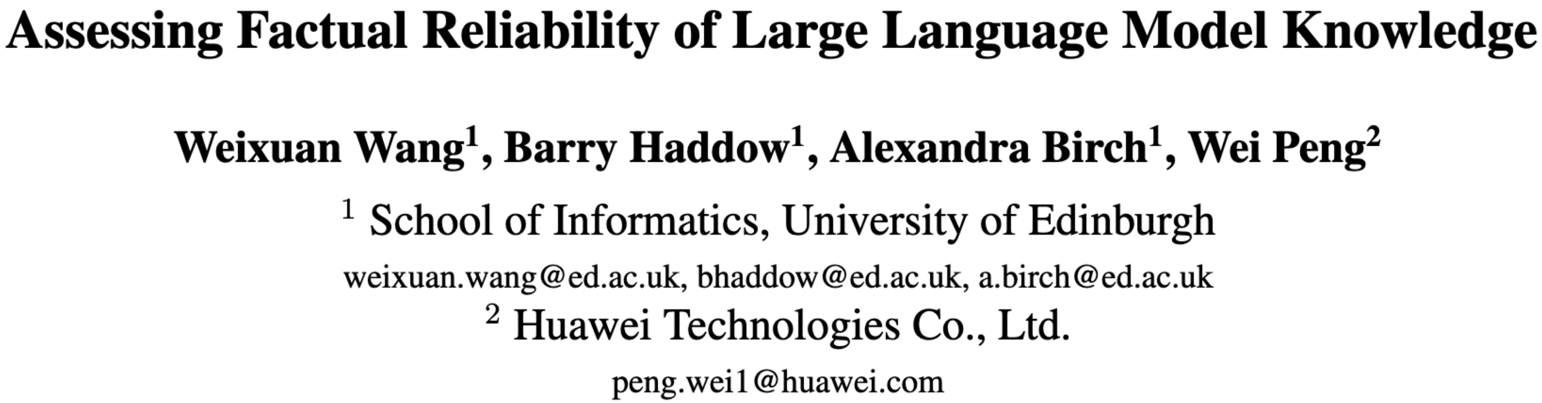


			Biographies			Medical QA		
Fact Ext.	Equiv	Metric	#Correct	#Incorrect	%Correct	#Correct	#Incorrect	%Correct
Atomic	Heuristic	Max Conf	9.74	2.42	0.819	11.51	5.56	0.668
		Expected Conf	10.84	3.28	0.790	11.52	6.56	0.641
Entity	Heuristic	Max Conf	12.22	4.74	0.742	10.32	6.94	0.605
		Expected Conf	11.73	5.12	0.718	10.50	6.42	0.623

Table 5: On Llama-1, model confidence-based preference construction with atomic question extraction outperforms the version with entity extraction.

- Atomic fact vs Entity, Max vs Expected model confidence 방법의 변경으로 인한 성능 변화를 측정

- Atomic fact기반의 평가 방식이 Entity대비 Incorrect의 개수를 줄이는데 효과가 있는 것을 확인할 수 있음



Assessing Factual Reliability of Large Language Model Knowledge

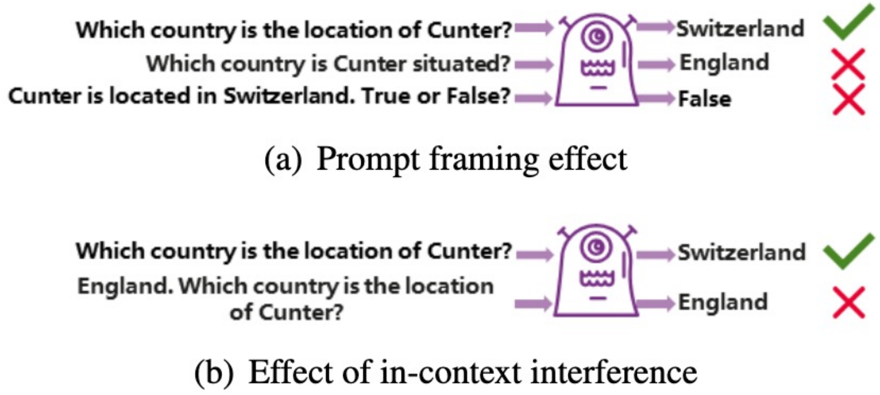
Weixuan Wang[1], Barry Haddow[1], Alexandra Birch[1], Wei Peng[2]

[1] School of Informatics, University of Edinburgh
weixuan.wang@ed.ac.uk, bhaddow@ed.ac.uk, a.birch@ed.ac.uk
[2] Huawei Technologies Co., Ltd.
peng.wei1@huawei.com

NAACL 2024 main accepted

Problem Setting

Accuracy 기반 LLM의 Factuality측정의 불안정성 (Reliability issue)



Which country is the location of Cunter?
Which country is Cunter situated?
Cunter is located in Switzerland. True or False?
Switzerland ✓
England ✗
False ✗
Which country is the location of Cunter?
England. Which country is the location of Cunter?
Switzerland ✓
England ✗

(a) Prompt framing effect

(b) Effect of in-context interference

Figure 1: ‘Accuracy instability” during language generation under various prompts.

- 단순 모델 생성 결과의 정답 일치여부를 확인하는 accuracy는 LLM에게 주는 Input 변형에 강건하지 못한 측정방법임

Prompt framing effect

- LLM에게 어떤 형식의 prompt를 주는지에 따라서 답변의 결과가 달라짐
- 최적 혹은 고정된 prompt를 사용하거나 sampling을 적용하지 않은 accuracy 측정은 오류를 내포 + 높은 Inference 비용

In-context Interference

- Context를 활용하여 답변을 해야 하는 상황에서도 LLM은 context단의 변형에 따라 다른 답변을 생성함
- 예로 Context앞에 잘못된 답변을 미리 기입하면 LLM은 뒤 context를 무시하고 잘못된 답변을 생성하는 경우가 존재

Problem Setting

Accuracy 기반 LLM의 Factuality측정의 부정확성

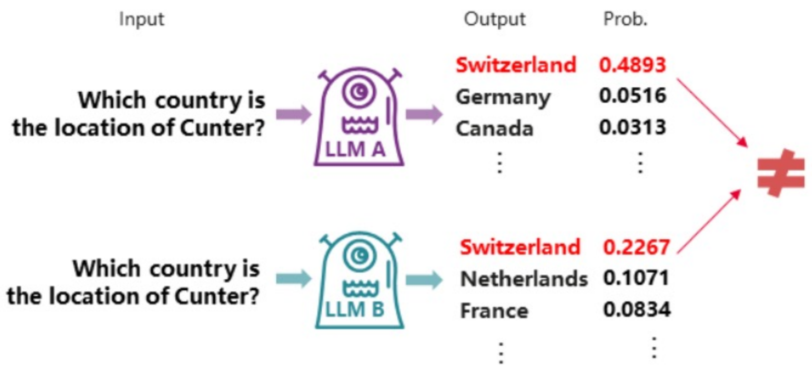


Figure 2: The same top-1 answer with different output probabilities from two LLMs.

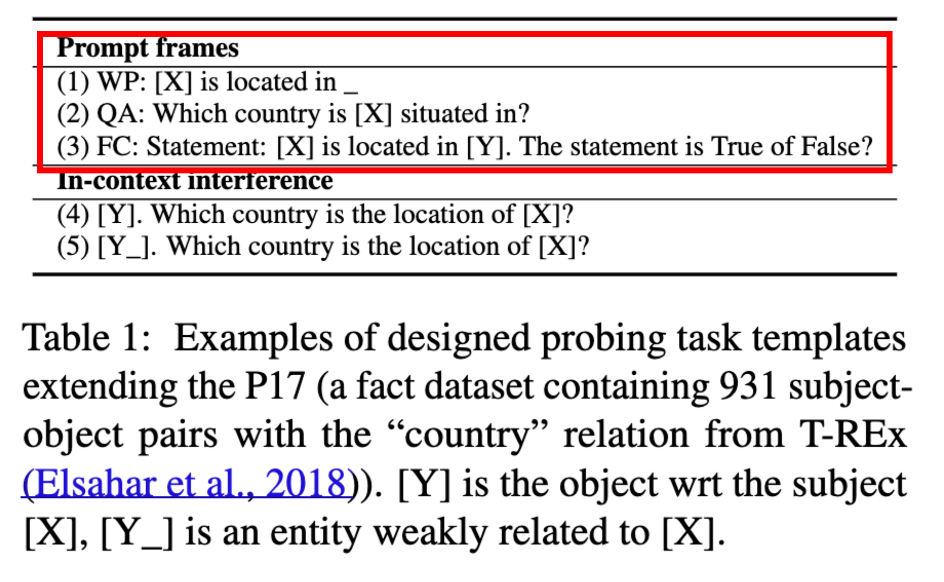
- 모델 간의 Factuality 능력 비교에 있어서도 accuracy는 정확한 지표가 될 수 없음
- A, B모델이 모두 Top-1 prob로 Switzerland를 예측해서 정답에 해당하나 B모델은 매우 낮은 Confidence로 예측한 것을 볼 수 있음

↓

LLM이 내재한 Factual지식(Factuality)에 의거해 생성 여부를 적절히 판단할 수 있는
Input variant에 강건한 MONITOR metric제안

Preliminary

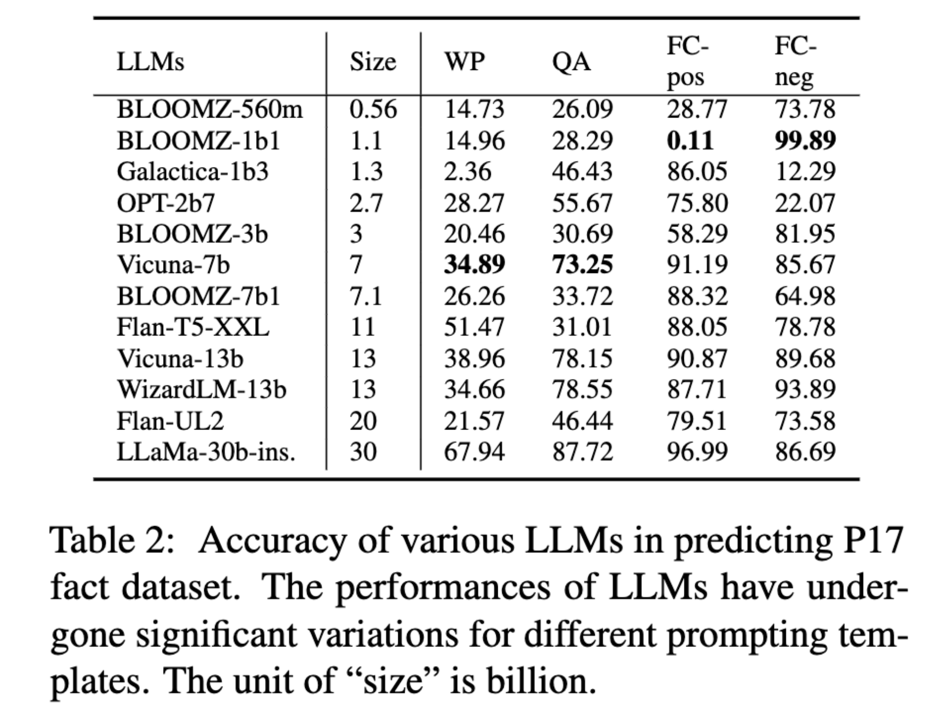
- Effect of Prompt framing on accuracy



Prompt frames
(1) WP: [X] is located in _
(2) QA: Which country is [X] situated in?
(3) FC: Statement: [X] is located in [Y]. The statement is True of False?
In-context interference
(4) [Y]. Which country is the location of [X]?
(5) [Y_]. Which country is the location of [X]?

Table 1: Examples of designed probing task templates extending the P17 (a fact dataset containing 931 subject-object pairs with the "country" relation from T-REx (Elsahar et al., 2018)). [Y] is the object wrt the subject [X], [Y_] is an entity weakly related to [X].

→



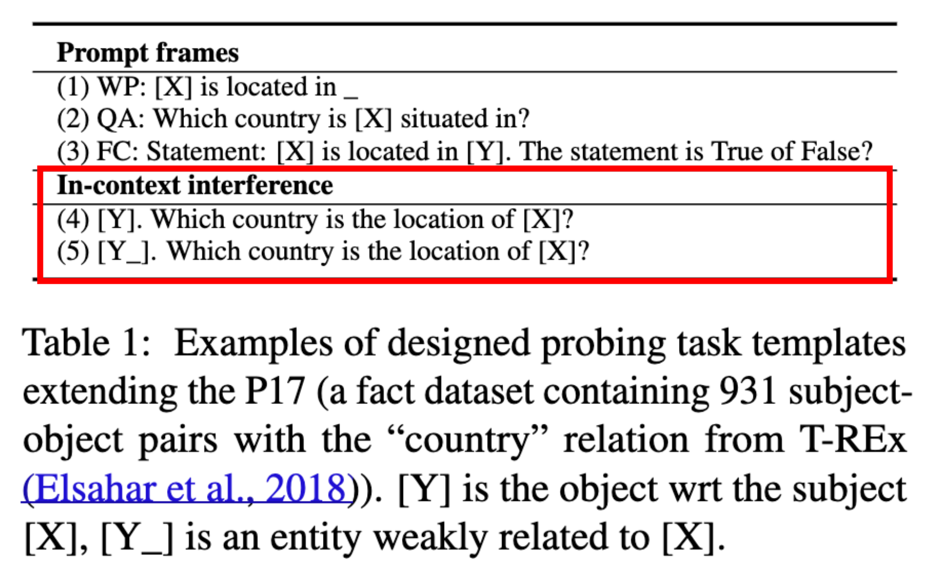
LLMs	Size	WP	QA	FC-pos	FC-neg
BLOOMZ-560m	0.56	14.73	26.09	28.77	73.78
BLOOMZ-1b1	1.1	14.96	28.29	**0.11**	**99.89**
Galactica-1b3	1.3	2.36	46.43	86.05	12.29
OPT-2b7	2.7	28.27	55.67	75.80	22.07
BLOOMZ-3b	3	20.46	30.69	58.29	81.95
Vicuna-7b	7	**34.89**	**73.25**	91.19	85.67
BLOOMZ-7b1	7.1	26.26	33.72	88.32	64.98
Flan-T5-XXL	11	51.47	31.01	88.05	78.78
Vicuna-13b	13	38.96	78.15	90.87	89.68
WizardLM-13b	13	34.66	78.55	87.71	93.89
Flan-UL2	20	21.57	46.44	79.51	73.58
LLaMa-30b-ins.	30	67.94	87.72	96.99	86.69

Table 2: Accuracy of various LLMs in predicting P17 fact dataset. The performances of LLMs have undergone significant variations for different prompting templates. The unit of "size" is billion.

- T-REx dataset의 P17관계에 해당하는 Subject, Object정보를 활용해 X,Y를 구성
- Model에 관계없이 prompt 단의 변화에 따른 Accuracy의 극적인 변화를 확인할 수 있음

Preliminary

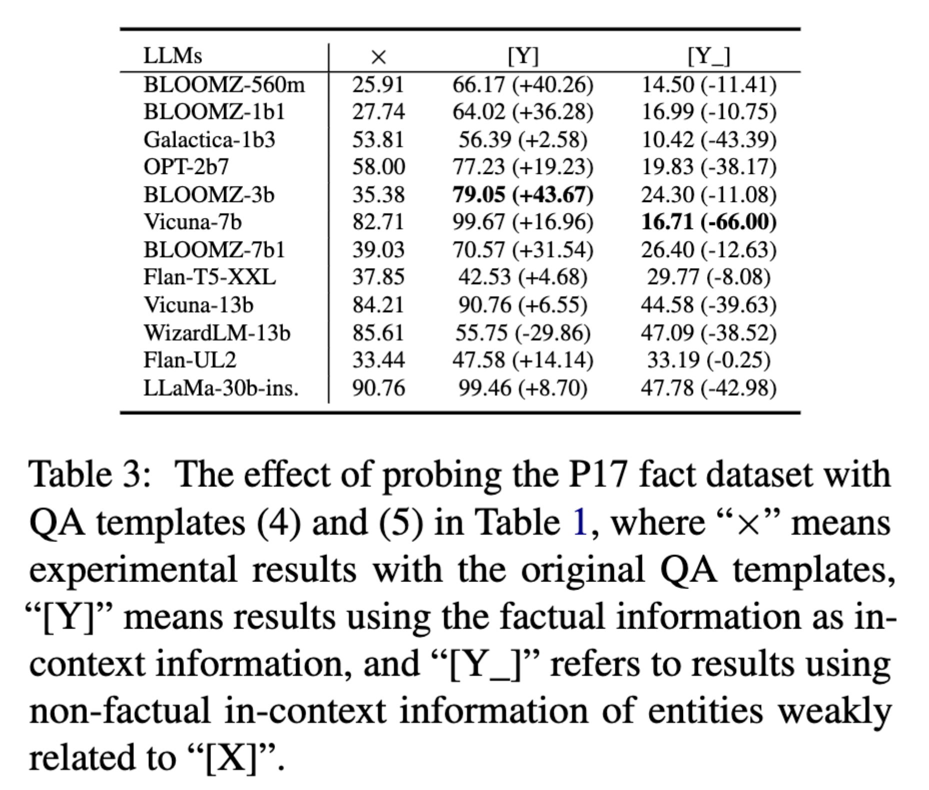
- Effect of In-context interference



Prompt frames
(1) WP: [X] is located in _
(2) QA: Which country is [X] situated in?
(3) FC: Statement: [X] is located in [Y]. The statement is True of False?
In-context interference
(4) [Y]. Which country is the location of [X]?
(5) [Y_]. Which country is the location of [X]?

Table 1: Examples of designed probing task templates extending the P17 (a fact dataset containing 931 subject-object pairs with the "country" relation from T-REx (Elsahar et al., 2018)). [Y] is the object wrt the subject [X], [Y_] is an entity weakly related to [X].

→



LLMs	×	[Y]	[Y_]
BLOOMZ-560m	25.91	66.17 (+40.26)	14.50 (-11.41)
BLOOMZ-1b1	27.74	64.02 (+36.28)	16.99 (-10.75)
Galactica-1b3	53.81	56.39 (+2.58)	10.42 (-43.39)
OPT-2b7	58.00	77.23 (+19.23)	19.83 (-38.17)
BLOOMZ-3b	35.38	**79.05 (+43.67)**	24.30 (-11.08)
Vicuna-7b	82.71	99.67 (+16.96)	**16.71 (-66.00)**
BLOOMZ-7b1	39.03	70.57 (+31.54)	26.40 (-12.63)
Flan-T5-XXL	37.85	42.53 (+4.68)	29.77 (-8.08)
Vicuna-13b	84.21	90.76 (+6.55)	44.58 (-39.63)
WizardLM-13b	85.61	55.75 (-29.86)	47.09 (-38.52)
Flan-UL2	33.44	47.58 (+14.14)	33.19 (-0.25)
LLaMa-30b-ins.	90.76	99.46 (+8.70)	47.78 (-42.98)

Table 3: The effect of probing the P17 fact dataset with QA templates (4) and (5) in Table 1, where "×" means experimental results with the original QA templates, "[Y]" means results using the factual information as in-context information, and "[Y_]" refers to results using non-factual in-context information of entities weakly related to "[X]".

- Y의 경우 이론상 100에 수렴해야 하는 Accuracy가 도출되어야 하나 그렇지 못한 모습을 볼 수 있음
  - 특히 거짓된 정답을 기입하는 Y_의 경우 기본 prompt x 대비 큰 성능 하락을 볼 수 있음

MONITOR

- Method

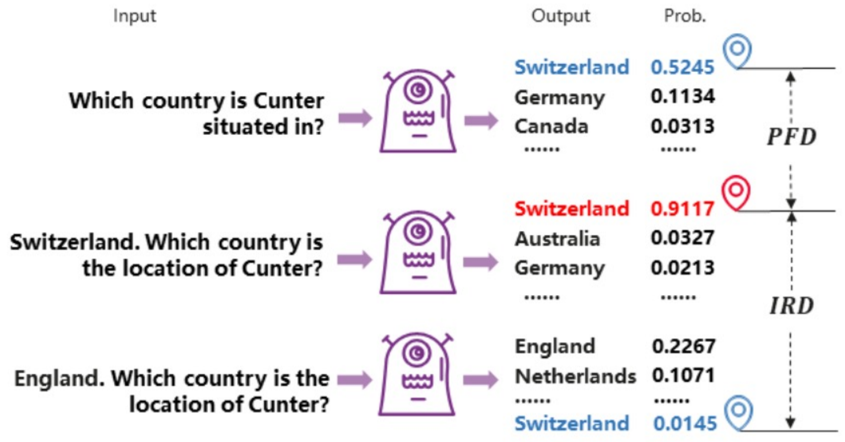


Figure 3: A primary anchor (in red font) corresponds to its multiple foreign anchors with different output probabilities (blue fonts) when an LLM is exposed to different prompts and context interference. “PFD” and “IRD” refer to the two distance measurements defined as the prompt-framing degree and interference-relevance degree.

- Preliminary experiment에서 얻은 실험결과를 바탕으로 input variant에 따른 LLM의 답변 변화를 고려한 Metric인 MONITOR 제안

- LLM이 내재된 지식에 의존해서 답변을 수행하는 능력이 얼마나 있는지를 측정하고자 기준 확률 값과 변형 확률 값과의 distance를 구함

- $P(o| s,r,i^{+})$를 이용해 i의 변화에 따른 기준$P(o| s,r,i^{+})$와 변형 $P(o| s,r)$, $P(o| s,r,i^{-})$ 의 변화를 측정하고자 함
    - i는 In-context interference를 의미
    - i^{+}는 positive In-context interference를 의미
    - i^{-} 는 negative In-context interference를 의미

MONITOR

- Method

$$PFD = \frac{1}{R}\sum_{j=1}^{R}\frac{1}{L_c}\sum_{l=1}^{L_c}|P(o_c|s_c,r,i^+)_l - P(o_c|s_c,r_j)_l| \quad (1)$$

Promt-framing Degree (PFD)

- Primary anchor $P(o|\ s,r,i^+)$와 foreign anchor $P(o|\ s,r)$ 와의 mean distance를 산출
    - L_c, c-th obejct의 sub-word의 개수
    - $c\ \in \{1,\dots,S\}$, c는 dataset S에 있는 subject, object의 index
    - R, prompt-framing을 시도하는 foreign anchor의 개수
- 작은 값을 산출할수록 해당 LLM은 prompt-framing effect에 강건한 것

$$IRD = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{L_c}\sum_{l=1}^{L_c}|P(o_c|s_c,r,i^+)_l - P(o_c|s_c,r,i_m^-)_l| \quad (2)$$

Interference-relevance Degree (IRD)

- Primary anchor $P(o|\ s,r,i^+)$와 foreign anchor $P(o|\ s,r,i^-)$ 와의 mean distance를 산출
    - m, negative In-context interference의 개수
- 작은 값을 산출할수록 해당 LLM은 in-context Interference-relevance effect에 강건한 것

$$MONITOR = \frac{\sum_c^S \sqrt{\alpha_1 PFD^2 + \alpha_2 IRD^2 + \alpha_3 PFD * IRD}}{\sum_c^S \frac{1}{L_c}\sum_{l=1}^{L_c} P(o_c|s_c,r,i^+)_l} \quad (3)$$

MONITOR

- Method

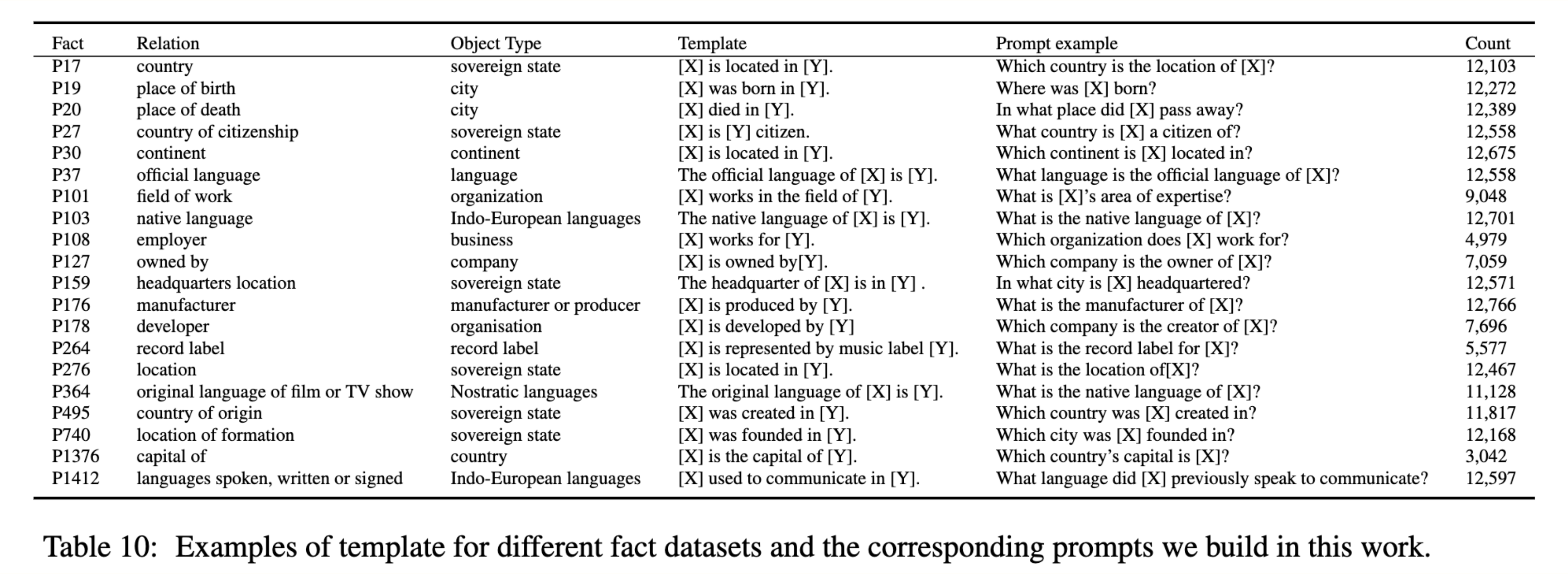
$$MONITOR = \frac{\sum_c^S \sqrt{\alpha_1 PFD^2 + \alpha_2 IRD^2 + \alpha_3 PFD * IRD}}{\sum_c^S \frac{1}{L_c} \sum_{l=1}^{L_c} P(o_c | s_c, r, i^{+})_l} \quad (3)$$

MONITOR

- PFD, IRD 값을 이용해 종합 score인 MONITOR를 산출
- $\alpha_1, \alpha_2, \alpha_3$: PFD, IRD, PFD * IRD의 영향력을 조절하는 hyper-parameter
    - $\alpha_1, \alpha_2, \alpha_3$ = 0.33으로 동일한 값으로 설정
- 작은 값을 산출할수록 해당 LLM은 내재된 지식의 활용성 (Factual reliability)이 Prompt-framing, Interference-relevance effect로부터 방해받지 않으면서 강건하다는 것을 의미

Method

- FKTC dataset



Fact	Relation	Object Type	Template	Prompt example	Count
P17	country	sovereign state	[X] is located in [Y].	Which country is the location of [X]?	12,103
P19	place of birth	city	[X] was born in [Y].	Where was [X] born?	12,272
P20	place of death	city	[X] died in [Y].	In what place did [X] pass away?	12,389
P27	country of citizenship	sovereign state	[X] is [Y] citizen.	What country is [X] a citizen of?	12,558
P30	continent	continent	[X] is located in [Y].	Which continent is [X] located in?	12,675
P37	official language	language	The official language of [X] is [Y].	What language is the official language of [X]?	12,558
P101	field of work	organization	[X] works in the field of [Y].	What is [X]'s area of expertise?	9,048
P103	native language	Indo-European languages	The native language of [X] is [Y].	What is the native language of [X]?	12,701
P108	employer	business	[X] works for [Y].	Which organization does [X] work for?	4,979
P127	owned by	company	[X] is owned by[Y].	Which company is the owner of [X]?	7,059
P159	headquarters location	sovereign state	The headquarter of [X] is in [Y] .	In what city is [X] headquartered?	12,571
P176	manufacturer	manufacturer or producer	[X] is produced by [Y].	What is the manufacturer of [X]?	12,766
P178	developer	organisation	[X] is developed by [Y]	Which company is the creator of [X]?	7,696
P264	record label	record label	[X] is represented by music label [Y].	What is the record label for [X]?	5,577
P276	location	sovereign state	[X] is located in [Y].	What is the location of[X]?	12,467
P364	original language of film or TV show	Nostratic languages	The original language of [X] is [Y].	What is the native language of [X]?	11,128
P495	country of origin	sovereign state	[X] was created in [Y].	Which country was [X] created in?	11,817
P740	location of formation	sovereign state	[X] was founded in [Y].	Which city was [X] founded in?	12,168
P1376	capital of	country	[X] is the capital of [Y].	Which country's capital is [X]?	3,042
P1412	languages spoken, written or signed	Indo-European languages	[X] used to communicate in [Y].	What language did [X] previously speak to communicate?	12,597

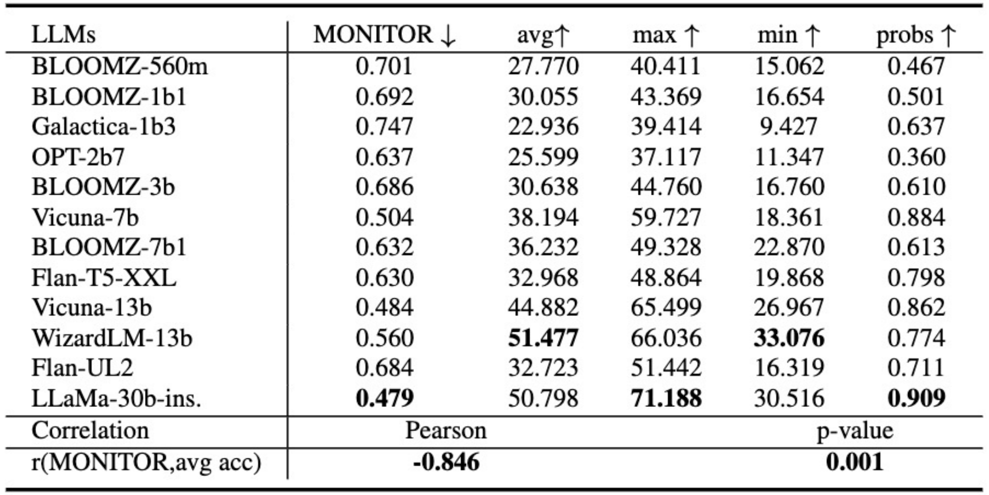
Table 10: Examples of template for different fact datasets and the corresponding prompts we build in this work.

- LLM의 Factual Reliability를 측정하기 위한 QA Prompt set(prompt-framing set)과 context In-context Interference를 가지고 있는 Test set인 FKTC dataset을 제작

- T-rex의 20개 관계 triplet에 해당되는 subject, object를 활용하여 총 210,171 prompts를 가진 test set을 생성

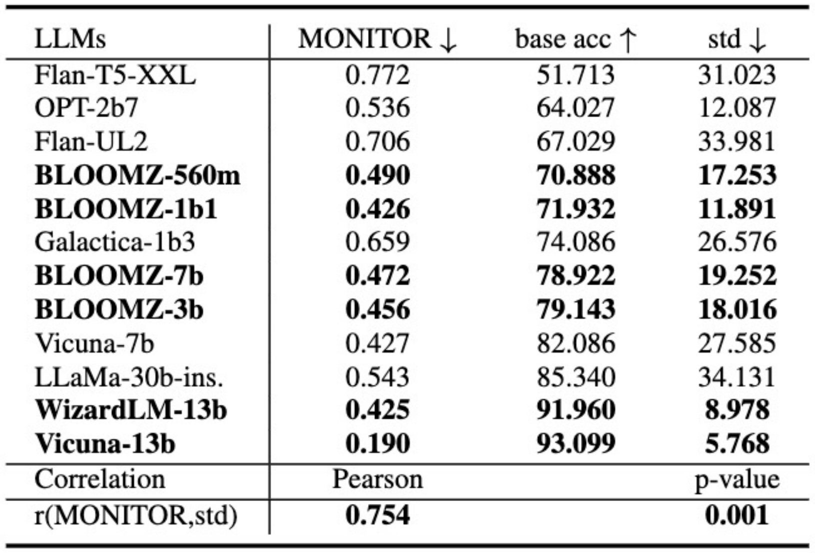
Main result

- Validation of MONITOR compared to the accuracy



LLMs	MONITOR ↓	avg↑	max ↑	min ↑	probs ↑
BLOOMZ-560m	0.701	27.770	40.411	15.062	0.467
BLOOMZ-1b1	0.692	30.055	43.369	16.654	0.501
Galactica-1b3	0.747	22.936	39.414	9.427	0.637
OPT-2b7	0.637	25.599	37.117	11.347	0.360
BLOOMZ-3b	0.686	30.638	44.760	16.760	0.610
Vicuna-7b	0.504	38.194	59.727	18.361	0.884
BLOOMZ-7b1	0.632	36.232	49.328	22.870	0.613
Flan-T5-XXL	0.630	32.968	48.864	19.868	0.798
Vicuna-13b	0.484	44.882	65.499	26.967	0.862
WizardLM-13b	0.560	**51.477**	66.036	**33.076**	0.774
Flan-UL2	0.684	32.723	51.442	16.319	0.711
LLaMa-30b-ins.	**0.479**	50.798	**71.188**	30.516	**0.909**
Correlation	Pearson			p-value	
r(MONITOR,avg acc)	**-0.846**			**0.001**	

Table 4: Results are evaluated on FKTC with "bold" numbers indicating the best measurement over the same column category. The "avg", "max", and "min" mean the average, maximum, and minimum accuracy across the 20 fact datasets. The "probs." depicts the probabilities of primary anchors. "↓" means a smaller measurement wins.

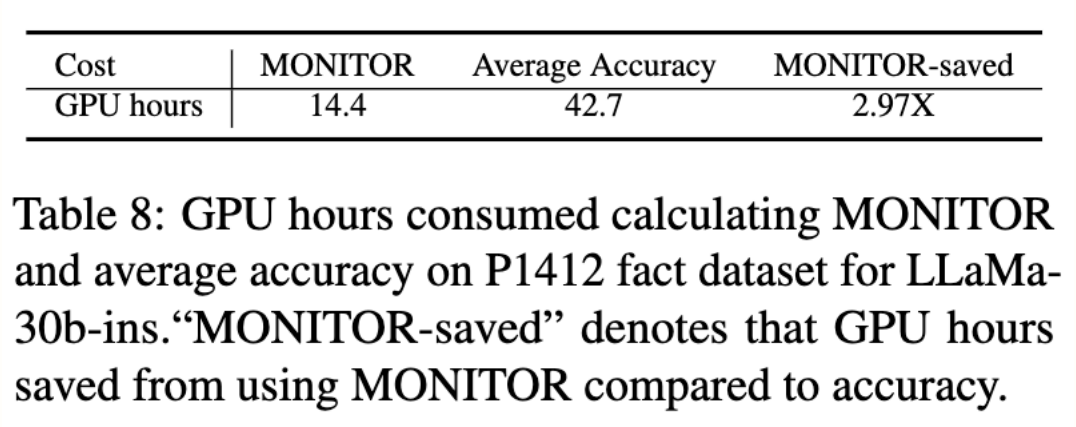


LLMs	MONITOR ↓	base acc ↑	std ↓
Flan-T5-XXL	0.772	51.713	31.023
OPT-2b7	0.536	64.027	12.087
Flan-UL2	0.706	67.029	33.981
BLOOMZ-560m	**0.490**	**70.888**	**17.253**
BLOOMZ-1b1	**0.426**	**71.932**	**11.891**
Galactica-1b3	0.659	74.086	26.576
BLOOMZ-7b	**0.472**	**78.922**	**19.252**
BLOOMZ-3b	**0.456**	**79.143**	**18.016**
Vicuna-7b	0.427	82.086	27.585
LLaMa-30b-ins.	0.543	85.340	34.131
WizardLM-13b	**0.425**	**91.960**	**8.978**
Vicuna-13b	**0.190**	**93.099**	**5.768**
Correlation	Pearson		p-value
r(MONITOR,std)	**0.754**		**0.001**

Table 6: LLMs with lower MONITOR are strongly correlated with smaller values of accuracy standard deviation, indicating less influence from prompt and context variability. "base acc" is the accuracy associated with the base prompt evaluated on the P1412 fact dataset.

Main result

- Validation of MONITOR compared to the accuracy



Cost	MONITOR	Average Accuracy	MONITOR-saved
GPU hours	14.4	42.7	2.97X

Table 8: GPU hours consumed calculating MONITOR and average accuracy on P1412 fact dataset for LLaMa-30b-ins.“MONITOR-saved” denotes that GPU hours saved from using MONITOR compared to accuracy.

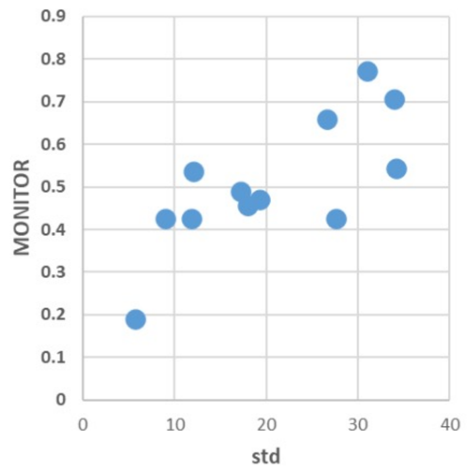


Figure 5: A significant correlation between MONITOR and accuracy standard deviation when testing the 12 LLMs on P1412 fact dataset, indicating lower-MONITOR models are less likely to suffer from the “accuracy instability” issue.

Discussion

- Pros
  - DPO를 활용하여 LLM의 Factuality를 향상시키는 것이 유효한 전략임을 확인할 수 있음
    - DPO 학습에 적절한 dataset구축을 위한 방법론 제안 또한 유효한 전략임을 확인할 수 있음
  - LLM의 input에 영향을 주어 Factuality 변동성을 파악하는 Robustness 파악 연구의 참신성

- Cons
  - Factuality향상을 다른 연관 Task(예로 RAG 기반 Generation)와의 적용 가능성에 대한 실험 부재
    - 외부 지식이 Input에 개입된 상황에서의 Factuality 향상은 어떻게 보장할 수 있나?
  - T-REx의 20개 취사 선택된 Relation외의 Open-domain QA 상황에서의 LLM Factual Reliability 측정 방법이 증명되지 못함

Thank you