Hallucination Detection

2024 겨울방학세미나

NLP&AI 강명훈

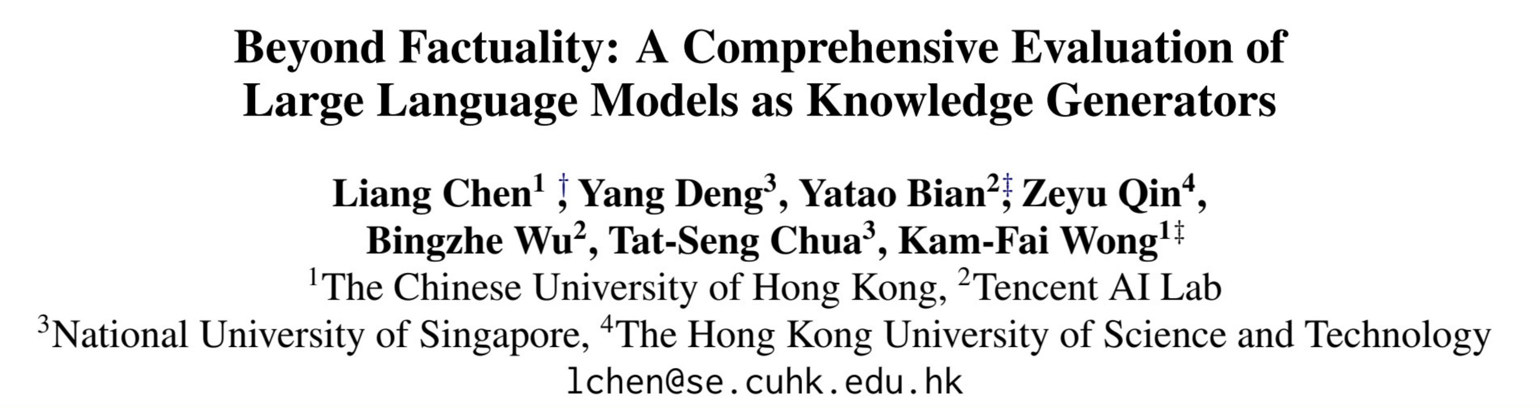
Theme of the seminar

- How can we assess the factuality of the generation output?
  - LLM 생성 결과물의 사실성 정도를 어떻게 정확하게 측정할 수 있는가?

- Does factuality matter in terms of the performance?
  - Factuality와 성능의 관계는 어떻게 되는가? Factuality가 보장되면 task별 성능이 높아지는가?

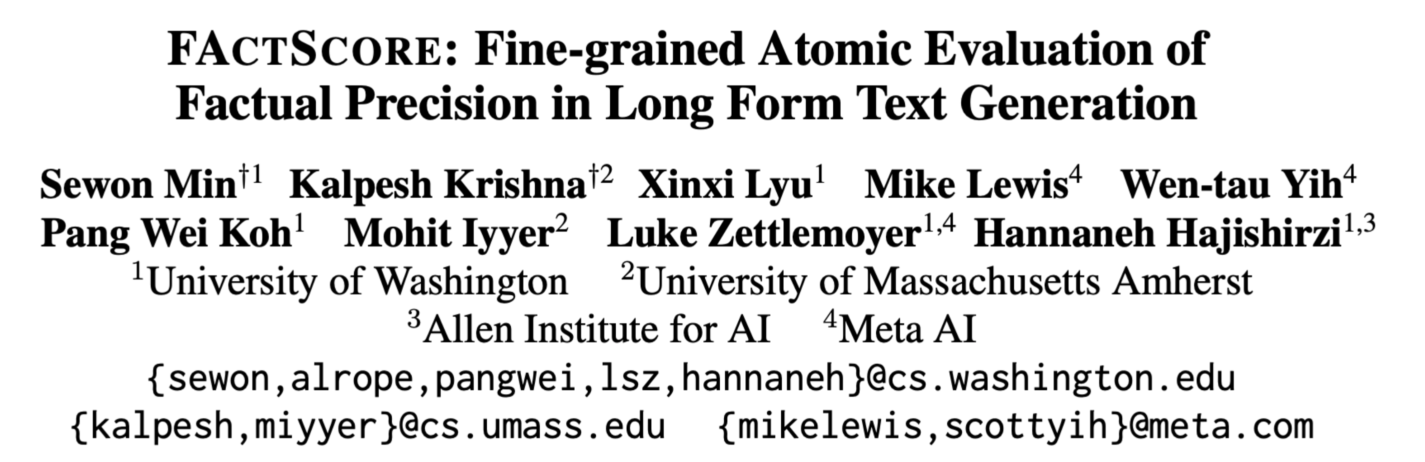
- How can we utilize LLM to assess the factuality of the LLM generation output?
  - LLM 생성 결과물을 평가하는데 다른 LLM을 사용하는 방법은 없을까?

Papers



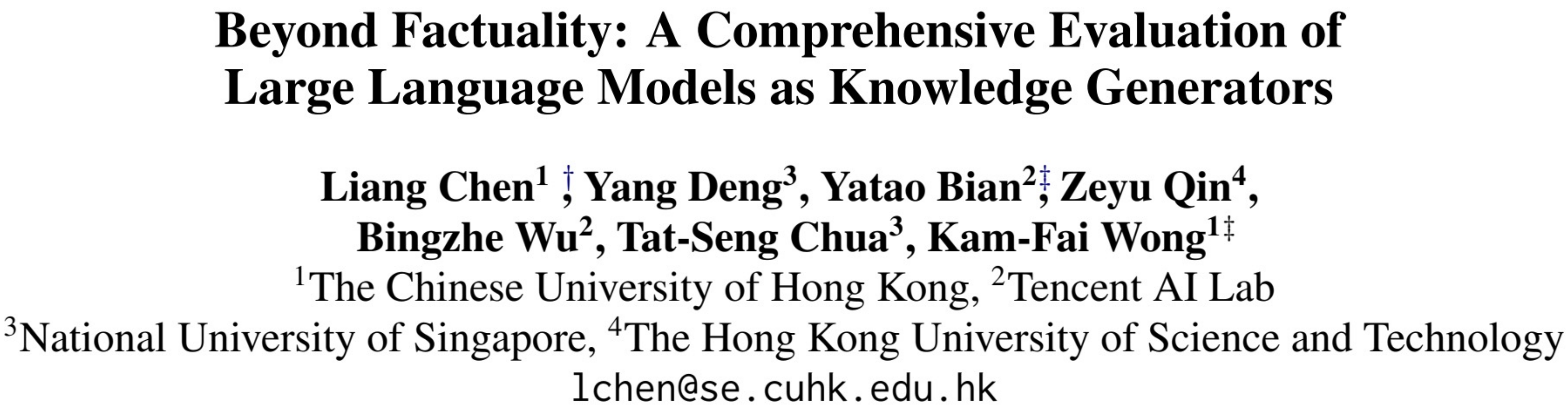
Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators

Liang Chen[1] †, Yang Deng[3], Yatao Bian[2]‡, Zeyu Qin[4],
Bingzhe Wu[2], Tat-Seng Chua[3], Kam-Fai Wong[1]‡
[1]The Chinese University of Hong Kong, [2]Tencent AI Lab
[3]National University of Singapore, [4]The Hong Kong University of Science and Technology
`lchen@se.cuhk.edu.hk`



FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation

Sewon Min[†1] **Kalpesh Krishna**[†2] **Xinxi Lyu**[1] **Mike Lewis**[4] **Wen-tau Yih**[4]
Pang Wei Koh[1] **Mohit Iyyer**[2] **Luke Zettlemoyer**[1,4] **Hannaneh Hajishirzi**[1,3]
[1]University of Washington [2]University of Massachusetts Amherst
[3]Allen Institute for AI [4]Meta AI
`{sewon,alrope,pangwei,lsz,hannaneh}@cs.washington.edu`
`{kalpesh,miyyer}@cs.umass.edu` `{mikelewis,scottyih}@meta.com`



Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators

Liang Chen[1] †, Yang Deng[3], Yatao Bian[2]‡, Zeyu Qin[4], Bingzhe Wu[2], Tat-Seng Chua[3], Kam-Fai Wong[1]‡

[1]The Chinese University of Hong Kong, [2]Tencent AI Lab

[3]National University of Singapore, [4]The Hong Kong University of Science and Technology

lchen@se.cuhk.edu.hk

EMNLP 2023 main accepted

Problem Setting

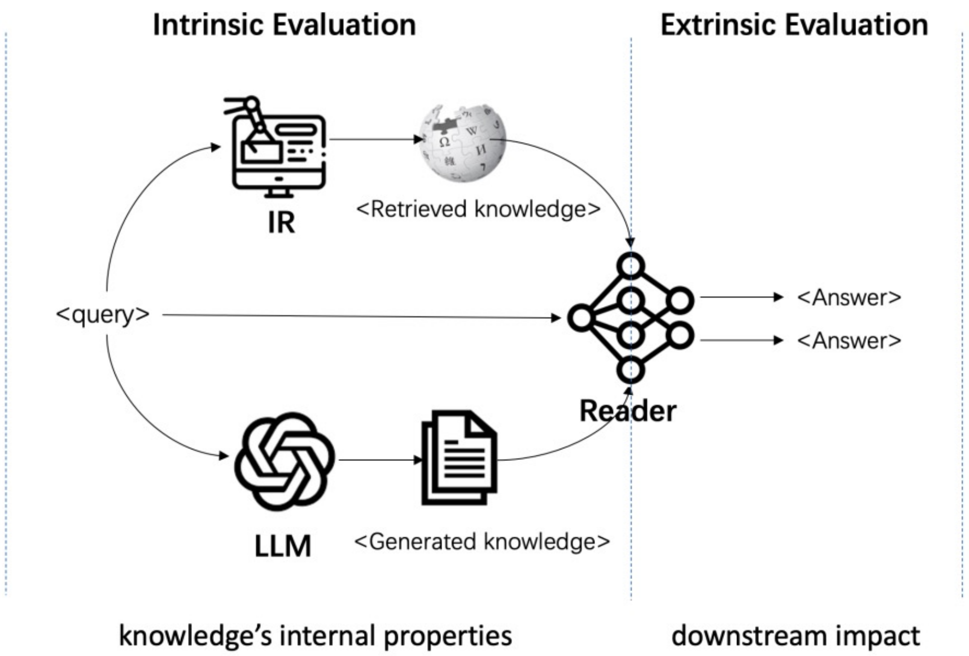


Figure 1: The CONNER Framework: Intrinsic evaluations probe the internal properties of acquired knowledge, while extrinsic evaluations assess its downstream impacts. This framework applies universally to two-stage processes in knowledge-intensive tasks.

LLM은 Knowledge-intensive task에서 기존 IR system 대비 우수한 성능을 보임

- open-domain question answering
- knowledge-grounded dialogue

Knowledge-intensive task를 수행하기 위해서 LLM은
world-knowledge를 생성하고 그 지식을 바탕으로 downstream task를 수행

여기서 사용된 world-knowledge의 품질은 어떻게 측정할 수 있는가?

←Human labeling 없는 접근 방법 필요
←Gold world-knowledge가 필요 없는 reference-free 방법 필요
←단순 Factuality 측정 외에 종합적인 판단 지표 필요

Problem Setting

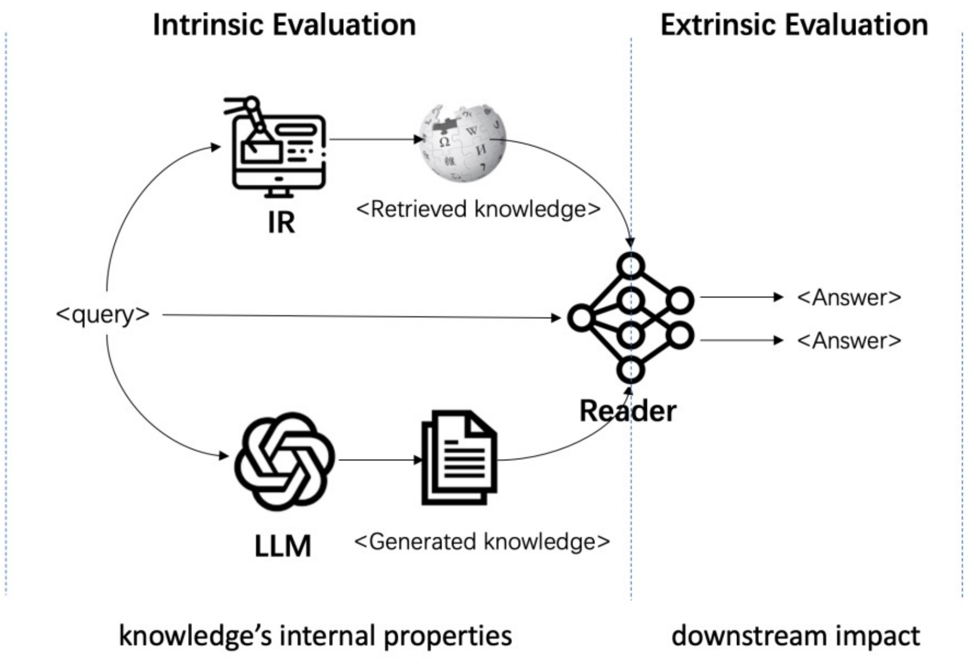


Figure 1: The CONNER Framework: Intrinsic evaluations probe the internal properties of acquired knowledge, while extrinsic evaluations assess its downstream impacts. This framework applies universally to two-stage processes in knowledge-intensive tasks.

LLM이 생성한 결과물안에 담긴 world-knowledge의 품질은 어떻게 측정할 수 있는가?

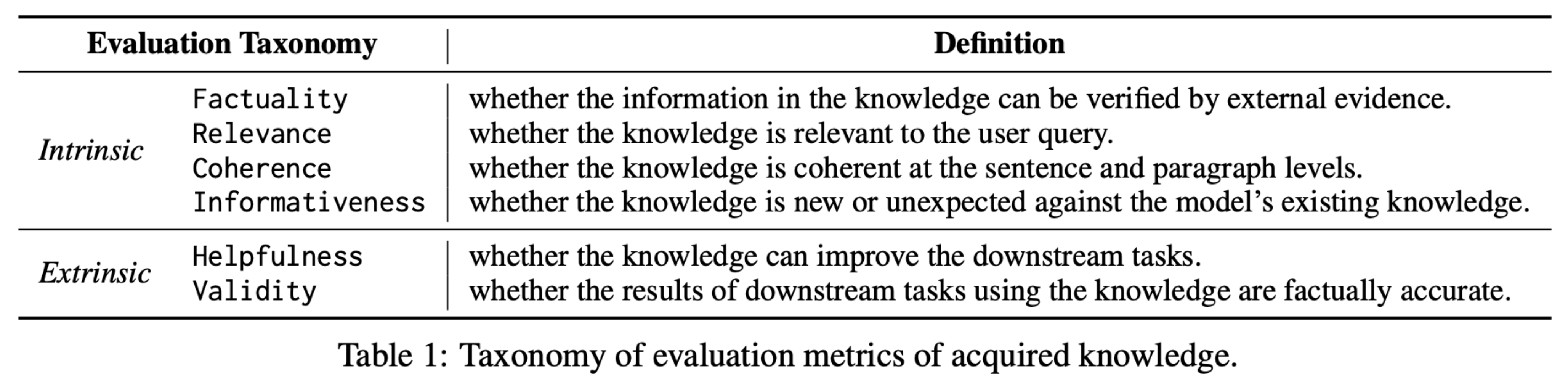
조건
- LLM이 Knowledge가 담긴 생성물을 생성하기 쉬운 환경

2가지 기준으로 평가 진행

Intrinsic: 모델이 생성한 **지식 그 자체**에 대한 평가
Extrinsic: 모델이 생성한 지식이 **downstream task성능 향상**에 영향을 주는지 평가

☞reference-free framework CONNER 제안

CONNER



Evaluation Taxonomy		Definition
Intrinsic	Factuality	whether the information in the knowledge can be verified by external evidence.
	Relevance	whether the knowledge is relevant to the user query.
	Coherence	whether the knowledge is coherent at the sentence and paragraph levels.
	Informativeness	whether the knowledge is new or unexpected against the model's existing knowledge.
Extrinsic	Helpfulness	whether the knowledge can improve the downstream tasks.
	Validity	whether the results of downstream tasks using the knowledge are factually accurate.

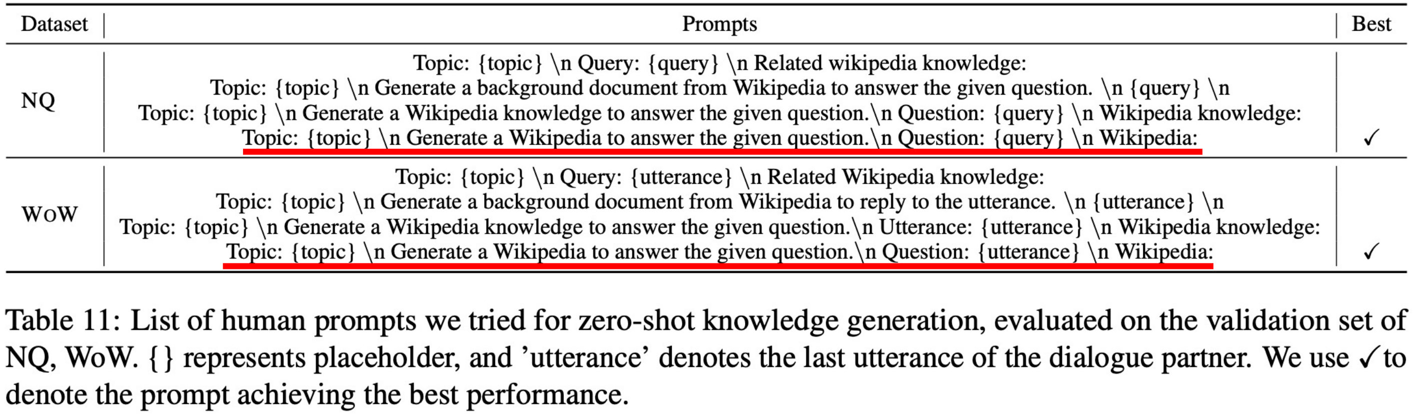
Table 1: Taxonomy of evaluation metrics of acquired knowledge.

Open-domain QA: NQ와 Knowledge-grounded dialogue: WoW의 160개 sample을 기반으로
LLaMA inference error case를 바탕으로 평가항목을 산출

Intrinsic Quality: 모델 생성 world-knowledge 그 자체에 대한 품질 평가
Extrinsic Impact: 모델 생성 world-knowledge가 downstream task에 주는 영향력 평가

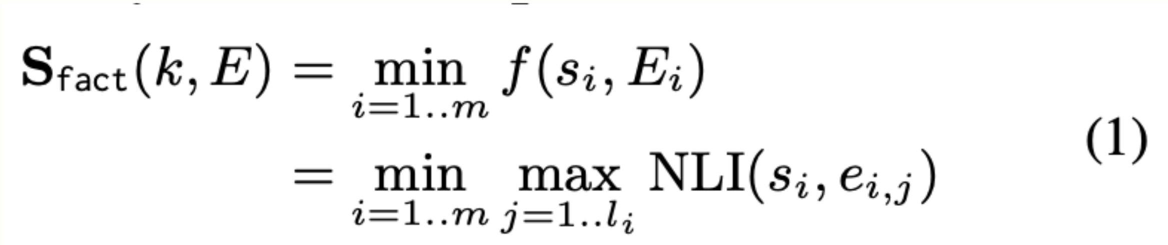
기본 평가 setting: LLM이 query와 관련된 지식을 생성하고 생성된 지식에 대한 Intrinsic & Extrinsic 평가 진행

CONNER: Intrinsic Quality



Dataset	Prompts	Best
NQ	Topic: {topic} \n Query: {query} \n Related wikipedia knowledge: Topic: {topic} \n Generate a background document from Wikipedia to answer the given question. \n {query} \n Topic: {topic} \n Generate a Wikipedia knowledge to answer the given question.\n Question: {query} \n Wikipedia knowledge: Topic: {topic} \n Generate a Wikipedia to answer the given question.\n Question: {query} \n Wikipedia:	✓
WoW	Topic: {topic} \n Query: {utterance} \n Related Wikipedia knowledge: Topic: {topic} \n Generate a background document from Wikipedia to reply to the utterance. \n {utterance} \n Topic: {topic} \n Generate a Wikipedia knowledge to answer the given question.\n Utterance: {utterance} \n Wikipedia knowledge: Topic: {topic} \n Generate a Wikipedia to answer the given question.\n Question: {utterance} \n Wikipedia:	✓

Table 11: List of human prompts we tried for zero-shot knowledge generation, evaluated on the validation set of NQ, WoW. {} represents placeholder, and 'utterance' denotes the last utterance of the dialogue partner. We use ✓to denote the prompt achieving the best performance.



$$\begin{aligned} \mathbf{S}_{\mathsf{fact}}(k, E) &= \min_{i=1..m} f(s_i, E_i) \\ &= \min_{i=1..m} \max_{j=1..l_i} \mathrm{NLI}(s_i, e_{i,j}) \end{aligned} \qquad (1)$$

- **Factuality**

whether the information in the knowledge can be verified by external evidence

Input: $k = \{s_1, \ldots, s_m\}$, $E_i = \{e_{1,l_i}, \ldots, e_{i,l_i}\}$,

output: S_{fact}

- LLM이 생성한 query와 관련된 m개의 문장으로 구성된 지식 k
- Retriever (ColBERTv2 모델)이 k와 관련된 Wikipedia 검색 지식 E_i (Gold-knowledge가 아님)
- NLI-RoBERTa-large을 이용하여 생성된 지식 문장 s_i가 관련 지식 $e_{i,j}$ 에 의해 수반되는지를 검증
- 3차원 벡터로 구성된 S_{fact} 산출

  factual-consistent (entail)

  non-verified (neutral)

  factual-inconsistent (refute)

CONNER: Intrinsic Quality

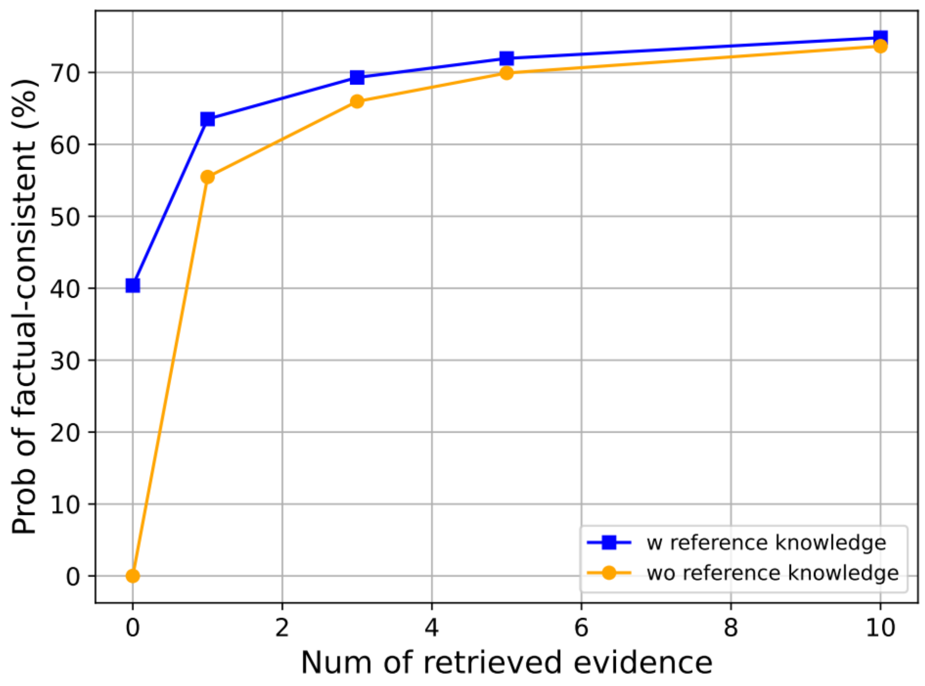


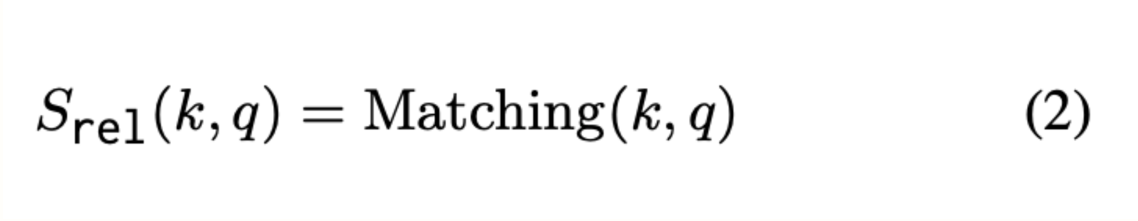
Figure 4: The influence of reference knowledge (e.g., the annotated Wiki. document in WoW dataset) in factuality evaluation weakens as the amount of retrieved evidence increases.

- Why no gold knowledge?

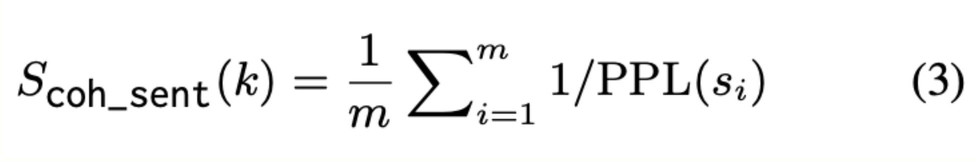
- Reference-free metric을 제안하기 위한 당위성을 검증하는 사전 실험 진행
- 쟁점: Factuality 평가에 Gold knowledge 사용이 필수인가?
- WoW 데이터셋의 dialogue별 gold knowledge와 검색된 retrieved knowledge를 evidence pool로 설정
- evidence pool에서 factuality 측정에 사용하기 위한 참조 evidence 개수 변경에 따른 Factual-consistent 비율 변화 측정
- Evidence 개수가 10개 이상일 경우 Prob가 converge되는 것을 확인할 수 있음

☞ 충분한 개수의 evidence만 확보된다면 retrieved evidence를 사용해도 적절하게 Factuality를 측정할 수 있다는 반증

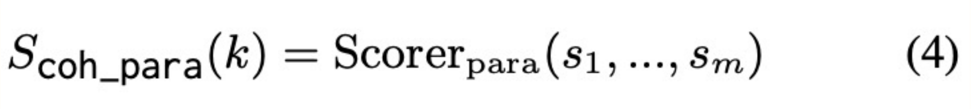
CONNER: Intrinsic Quality



$$S_{\mathtt{rel}}(k,q) = \mathrm{Matching}(k,q) \qquad (2)$$



$$S_{\mathsf{coh\_sent}}(k) = \frac{1}{m}\sum_{i=1}^{m} 1/\mathrm{PPL}(s_i) \qquad (3)$$



$$S_{\mathsf{coh\_para}}(k) = \mathrm{Scorer}_{\mathrm{para}}(s_1, \ldots, s_m) \qquad (4)$$

- ## Relevance

whether the knowledge is relevant to the user query

Input: q, $k = \{s_1, \ldots., s_m\}$

output: $\mathrm{S}_{\mathrm{rel}}$

- LLM이 생성한 지식 k와 query 와의 관련성 측정
- BERT-ranking-large 모델을 이용하여 관련성 측정

- ## Coherence

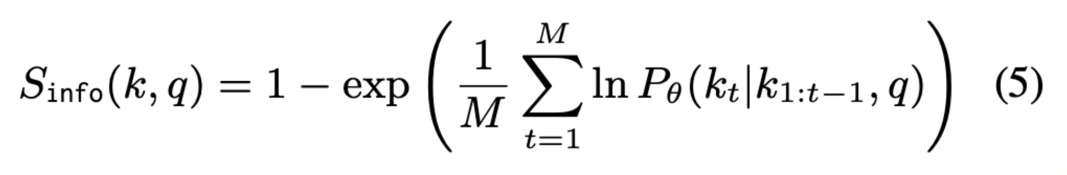
whether the knowledge is coherent at the sentence and paragraph levels

Input: $k = \{s_1, \ldots., s_m\}$,

output: $\mathrm{S}_{\mathrm{coh\_sent}}$, $\mathrm{S}_{\mathrm{coh\_para}}$

- LLM이 생성한 지식 k 자체 및 문장 간 일관성 측정
- $\mathrm{S}_{\mathrm{coh\_sent}}$ 의 PPL은 GPT-neo-2.7B 모델 출력을 바탕으로 산출
- $\mathrm{S}_{\mathrm{coh_para}}$는 Coherence-Momentum 모델의 score를 이용

CONNER: Intrinsic Quality



$$S_{\mathtt{info}}(k,q) = 1 - \exp\left(\frac{1}{M}\sum_{t=1}^{M} \ln P_\theta(k_t|k_{1:t-1},q)\right) \quad (5)$$

- Informativeness

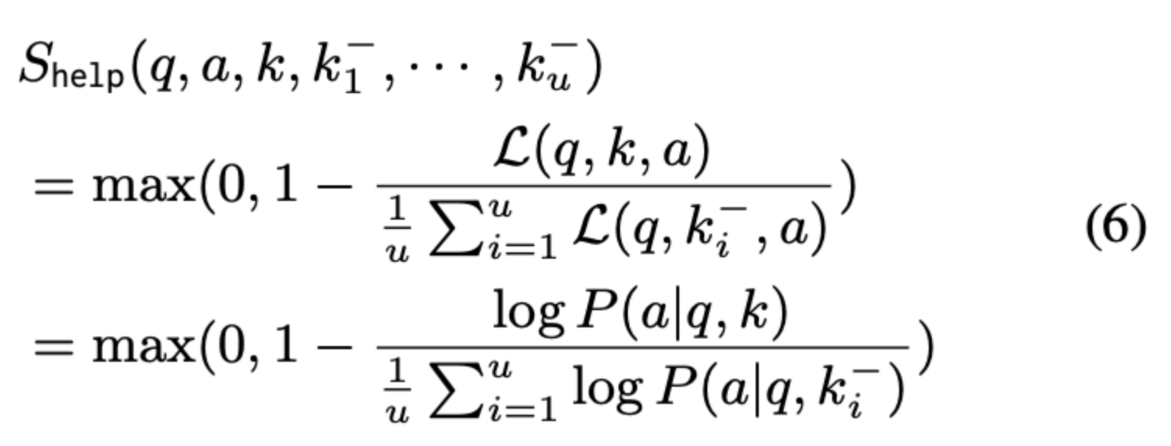
whether the knowledge is new or unexpected against the model's existing knowledge

Input:q, $k = \{s_1, \ldots, s_m\}$

output: $\mathrm{S_{info}}$

- Query를 GPT-neo-2.7B에 입력으로 주고 생성을 했을 때 얼마나 새로운 지식, 즉 정보를 담고 있는지를 측정
- 만약 GPT-neo-2.7B가 pre-train 단계에서 본 지식이면 생성 확률이 높아져서 전체 score는 0으로 수렴할 것이고
- 반대로 pre-train 단계에서 보지 못한 지식은 생성 확률이 낮아지므로 전체 score는 1로 수렴

CONNER: Extrinsic Quality



$$
\begin{aligned}
&S_{\text{help}}(q, a, k, k_1^-, \cdots, k_u^-) \\
&= \max(0, 1 - \frac{\mathcal{L}(q, k, a)}{\frac{1}{u}\sum_{i=1}^{u} \mathcal{L}(q, k_i^-, a)}) \qquad (6) \\
&= \max(0, 1 - \frac{\log P(a|q, k)}{\frac{1}{u}\sum_{i=1}^{u} \log P(a|q, k_i^-)})
\end{aligned}
$$

- Helpfulness

whether the knowledge can improve the downstream tasks.

Input:q, a, k, $\{k_1^-, \ldots, k_u^-\}$

output: S_{help}

- 생성된 지식이 answer 생성에 얼마나 도움이 되는지를 측정
- LLaMA-65B 모델을 이용하여 query, 생성 knowledge, negative knowledge가 주어질 때 negative knowledge loss 합 대비 생성 knowledge loss 비율
- 만약 생성 knowledge가 주어질 때의 generation loss가 높다면 score가 0으로 수렴될 것

CONNER: Extrinsic Quality

$$S_{\mathtt{val}}(q, a^*, a) = \mathrm{NLI}_{\mathrm{fact}}((q, a), (q, a^*)) \qquad (7)$$

$$S_{\mathtt{val}}(a, E) = f(a, E) = \max_{i=1..l} \mathrm{NLI}_{\mathrm{fact}}(a, e_i) \qquad (8)$$

- Validity

whether the results of downstream tasks using the knowledge are factually accurate

Input:q, a, a^*

output: S_{val}

- 모델이 생성한 answer와 ground-truth answer 혹은 evidence와의 수반 정도를 바탕으로 validity 측정

Experiment Setup

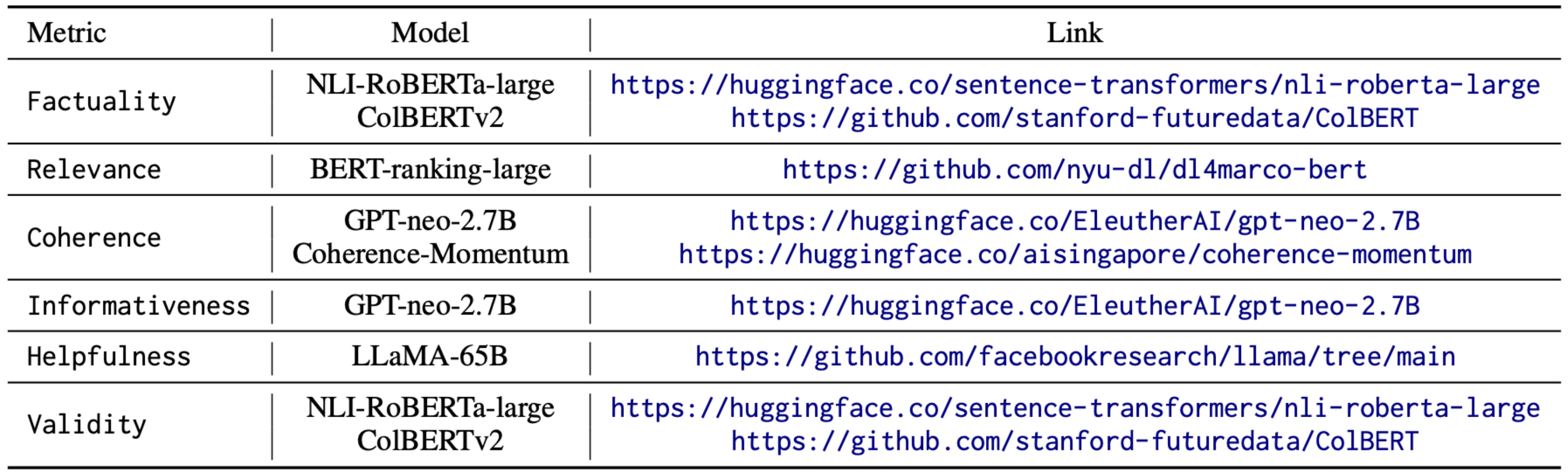
- 사용 Model & Evaluation setting

Baseline

- DPR: query와 관련된 지식을 검색하는 모델
- FLANT-T5, LLaMA, ChatGPT: query와 관련된 지식을 생성하는 모델

Dataset

- Natural Questions (NQ)
- Wizard of Wikipedia (WoW)

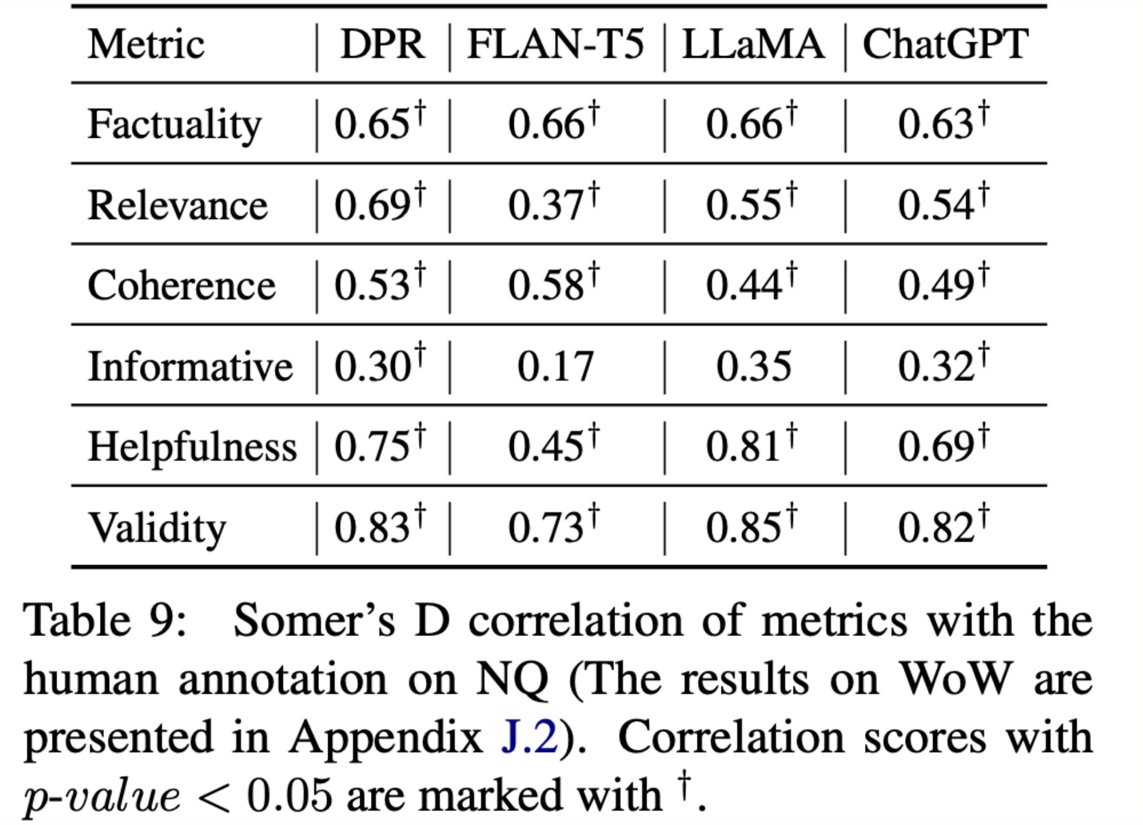


Metric	Model	Link
Factuality	NLI-RoBERTa-large ColBERTv2	https://huggingface.co/sentence-transformers/nli-roberta-large https://github.com/stanford-futuredata/ColBERT
Relevance	BERT-ranking-large	https://github.com/nyu-dl/dl4marco-bert
Coherence	GPT-neo-2.7B Coherence-Momentum	https://huggingface.co/EleutherAI/gpt-neo-2.7B https://huggingface.co/aisingapore/coherence-momentum
Informativeness	GPT-neo-2.7B	https://huggingface.co/EleutherAI/gpt-neo-2.7B
Helpfulness	LLaMA-65B	https://github.com/facebookresearch/llama/tree/main
Validity	NLI-RoBERTa-large ColBERTv2	https://huggingface.co/sentence-transformers/nli-roberta-large https://github.com/stanford-futuredata/ColBERT

Table 14: List of all models that we use in designing our framework.

Main result

- Metric validity



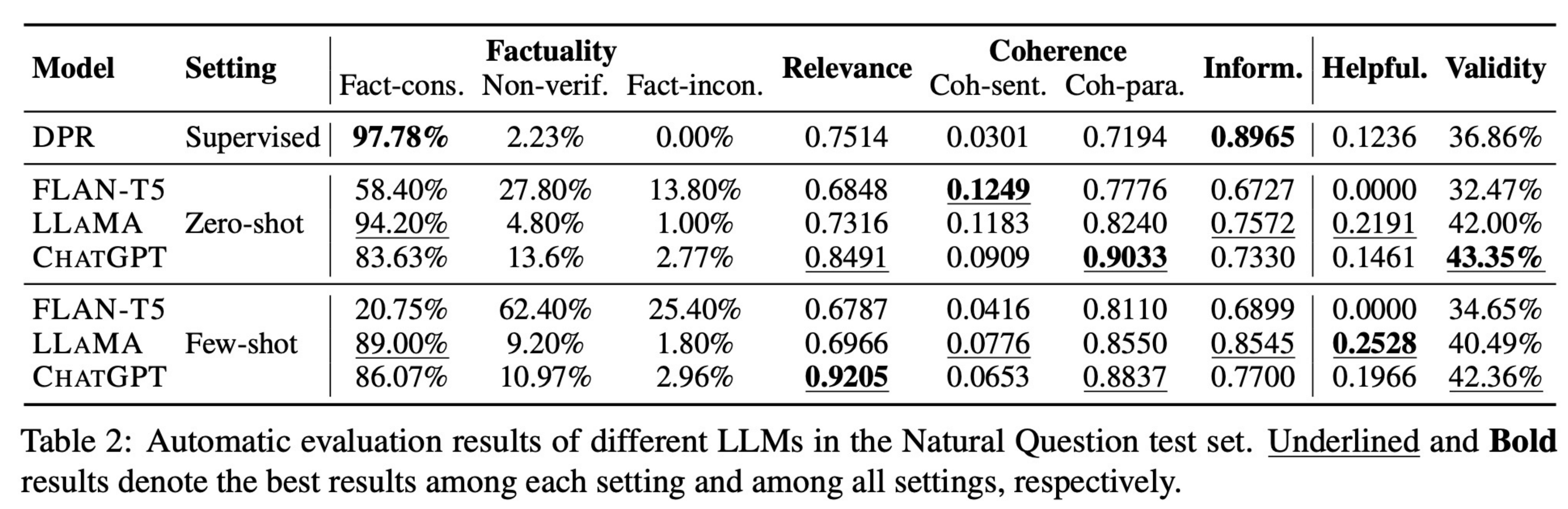
Metric	DPR	FLAN-T5	LLaMA	ChatGPT
Factuality	$0.65^\dagger$	$0.66^\dagger$	$0.66^\dagger$	$0.63^\dagger$
Relevance	$0.69^\dagger$	$0.37^\dagger$	$0.55^\dagger$	$0.54^\dagger$
Coherence	$0.53^\dagger$	$0.58^\dagger$	$0.44^\dagger$	$0.49^\dagger$
Informative	$0.30^\dagger$	0.17	0.35	$0.32^\dagger$
Helpfulness	$0.75^\dagger$	$0.45^\dagger$	$0.81^\dagger$	$0.69^\dagger$
Validity	$0.83^\dagger$	$0.73^\dagger$	$0.85^\dagger$	$0.82^\dagger$

Table 9: Somer's D correlation of metrics with the human annotation on NQ (The results on WoW are presented in Appendix J.2). Correlation scores with $p\text{-}value < 0.05$ are marked with †.

- NQ, WoW test set에서 추출한 320개 sample에 대하여 DPR, FLAN-T5, LLaMA, ChatGPT 모델로 80개씩 Knowledge & Answer Generation을 수행한 뒤
- 3명의 annotator들이 Intrinsic, Extrinsic scoring 진행
  → 0, 1, 2 3가지 점수로 labeling
- 해당 annotation을 기준으로 제안한 CONNER framework의 예측 점수와 annotation 점수 간의 상관관계를 측정
- 4개의 모델의 대부분의 지표에서 통계적으로 유의미한 양의 상관관계가 나타나는 것을 볼 수 있음

Main result

- Quantitative Result



Model	Setting	Factuality			Relevance	Coherence		Inform.	Helpful.	Validity
		Fact-cons.	Non-verif.	Fact-incon.		Coh-sent.	Coh-para.			
DPR	Supervised	**97.78%**	2.23%	0.00%	0.7514	0.0301	0.7194	**0.8965**	0.1236	36.86%
FLAN-T5	Zero-shot	58.40%	27.80%	13.80%	0.6848	**0.1249**	0.7776	0.6727	0.0000	32.47%
LLAMA		94.20%	4.80%	1.00%	0.7316	0.1183	0.8240	0.7572	0.2191	42.00%
CHATGPT		83.63%	13.6%	2.77%	0.8491	0.0909	**0.9033**	0.7330	0.1461	**43.35%**
FLAN-T5	Few-shot	20.75%	62.40%	25.40%	0.6787	0.0416	0.8110	0.6899	0.0000	34.65%
LLAMA		89.00%	9.20%	1.80%	0.6966	0.0776	0.8550	0.8545	**0.2528**	40.49%
CHATGPT		86.07%	10.97%	2.96%	**0.9205**	0.0653	0.8837	0.7700	0.1966	42.36%

Table 2: Automatic evaluation results of different LLMs in the Natural Question test set. Underlined and **Bold** results denote the best results among each setting and among all settings, respectively.

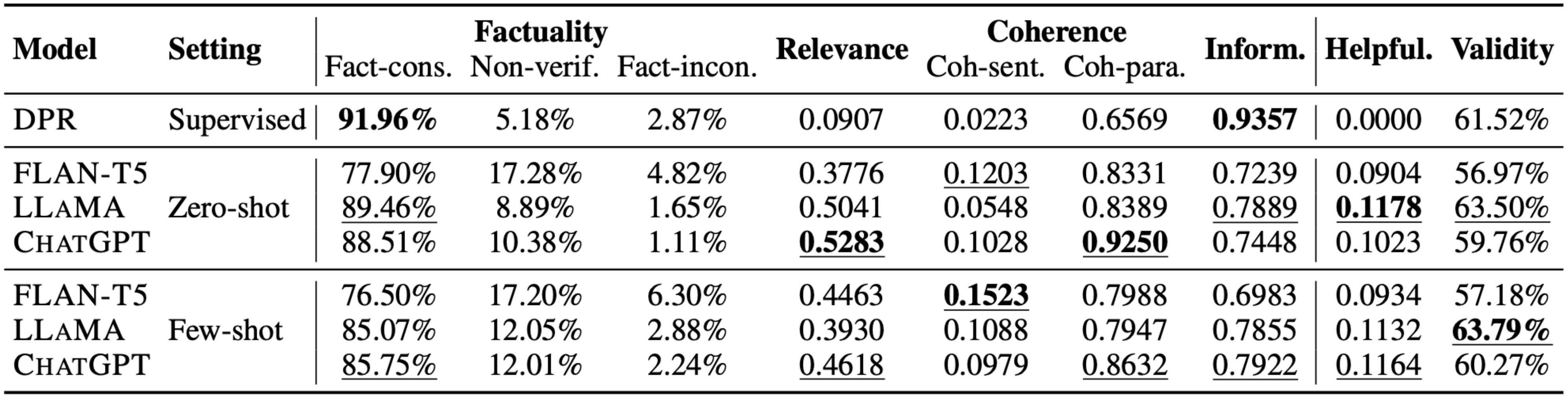
NQ에서의 DPR 검색 지식과 LLM 생성 지식에 대한 CONNER 점수 측정 결과

- DPR은 LLM 대비 Factuality, Informativeness에서 우수한 점수를 보임

- LLM이 DPR 대비 낮은 Factuality를 보이나 실제 downstream task에 미치는 영향을 보는 지표인 Helpfulness, Validity 부분에서는 앞서는 모습도 보임

Main result

- Quantitative Result



Model	Setting	Factuality			Relevance	Coherence		Inform.	Helpful.	Validity
		Fact-cons.	Non-verif.	Fact-incon.		Coh-sent.	Coh-para.			
DPR	Supervised	**91.96%**	5.18%	2.87%	0.0907	0.0223	0.6569	**0.9357**	0.0000	61.52%
FLAN-T5	Zero-shot	77.90%	17.28%	4.82%	0.3776	0.1203	0.8331	0.7239	0.0904	56.97%
LLAMA		89.46%	8.89%	1.65%	0.5041	0.0548	0.8389	0.7889	**0.1178**	63.50%
CHATGPT		88.51%	10.38%	1.11%	**0.5283**	0.1028	**0.9250**	0.7448	0.1023	59.76%
FLAN-T5	Few-shot	76.50%	17.20%	6.30%	0.4463	**0.1523**	0.7988	0.6983	0.0934	57.18%
LLAMA		85.07%	12.05%	2.88%	0.3930	0.1088	0.7947	0.7855	0.1132	**63.79%**
CHATGPT		85.75%	12.01%	2.24%	0.4618	0.0979	0.8632	0.7922	0.1164	60.27%

Table 3: Automatic evaluation results of different LLMs in the Wizard of Wikipedia test set.

- 앞서 NQ의 결과와 비슷하게 DPR은 LLM대비 Factuality, Inform에서 상대적 우수한 점수를 보임
- 그러나 Relevance, Helpfulness 부분에서는 LLM대비 미흡한 점수를 보이는 것을 알 수 있음

☞ 모델별 Intrinsic 지표와 Extrinsic 지표의 Trade off 관계를 파악해볼 수 있음

☞ Intrinsic 지표 향상을 위한 Retriever 관점의 접근과 Extrinsic 지표 향상을 위한 LLM 관점의 접근이 조화 되어야 함을 알 수 있음

Further Analysis

- Impact of Long-tail knowledge & Model size

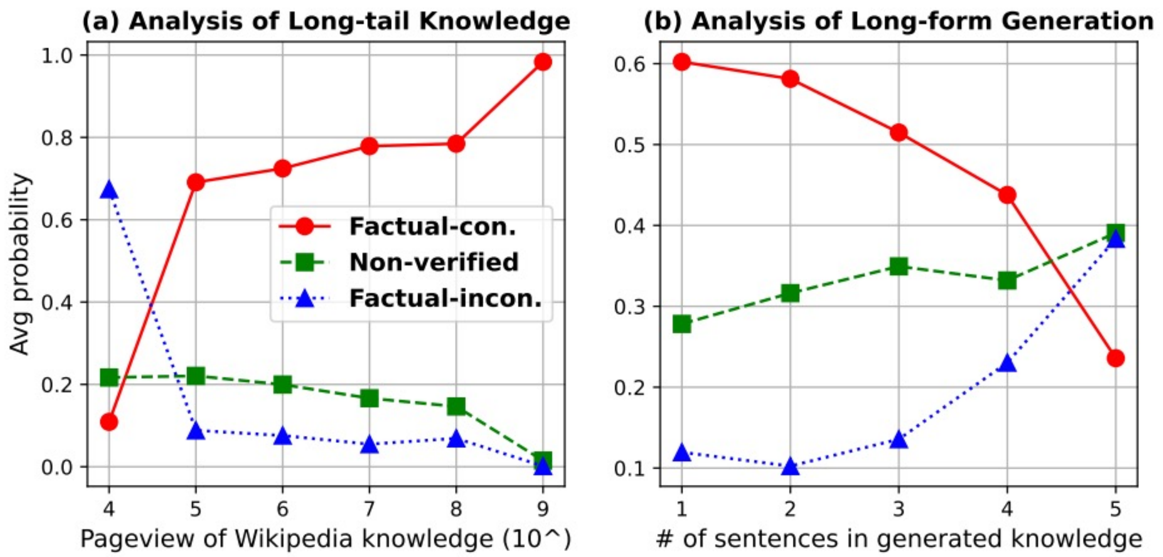


Figure 2: The impact of knowledge frequency and length on the factuality of the generated knowledge.

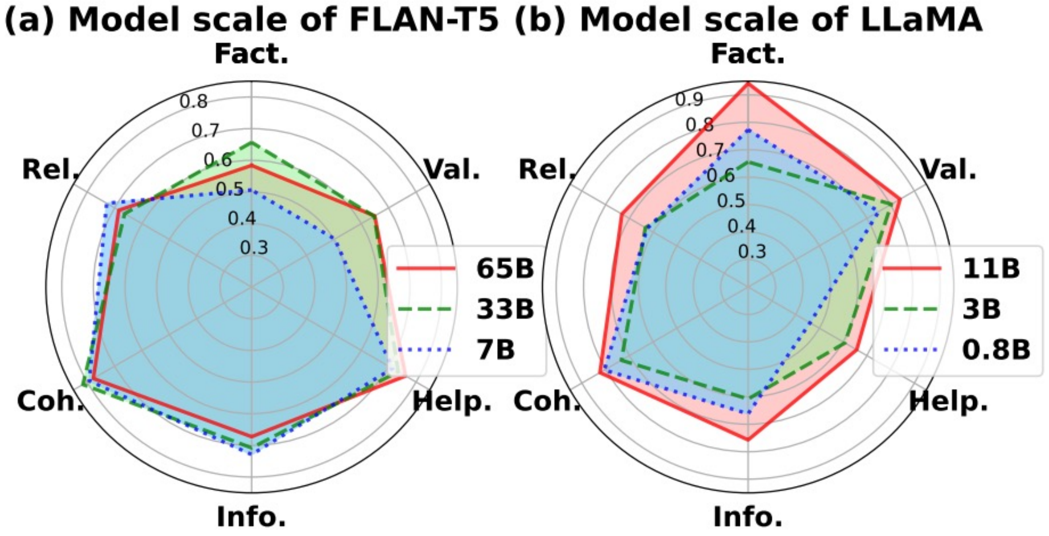
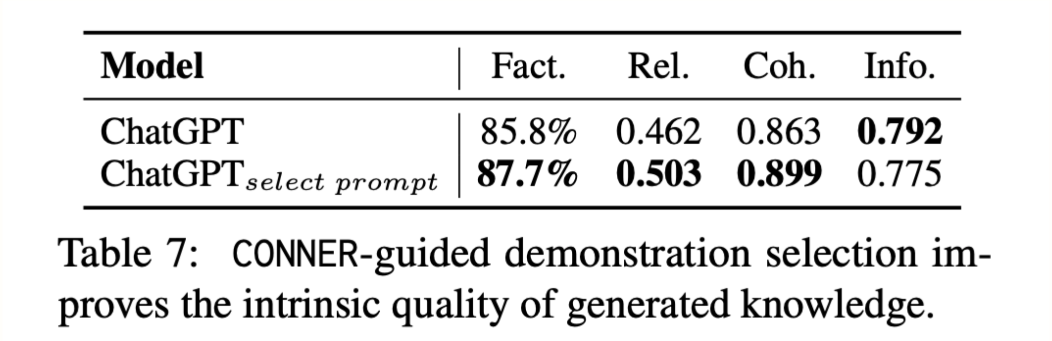


Figure 3: Performance on NQ with different sizes of FLAN-T5 and LLaMA as the knowledge generator (Help. and Val. scores are linearly scaled).

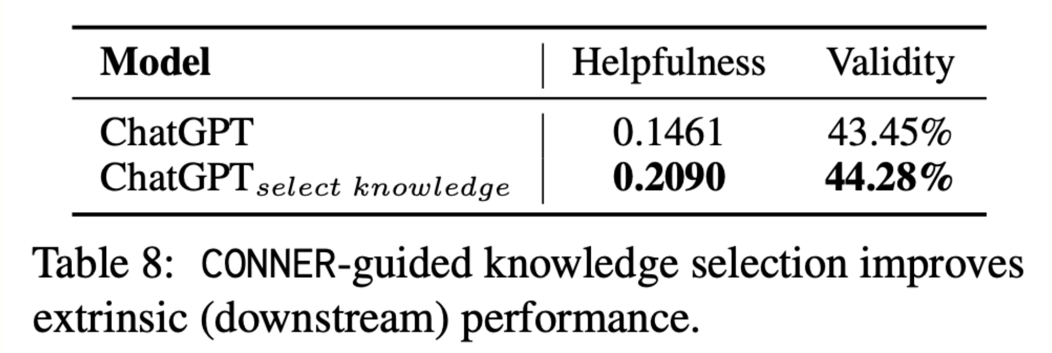
Practical use case

- Use case for Prompt engineering & Knowledge selection



Model	Fact.	Rel.	Coh.	Info.
ChatGPT	85.8%	0.462	0.863	**0.792**
ChatGPT$_{select\ prompt}$	**87.7%**	**0.503**	**0.899**	0.775

Table 7: CONNER-guided demonstration selection improves the intrinsic quality of generated knowledge.



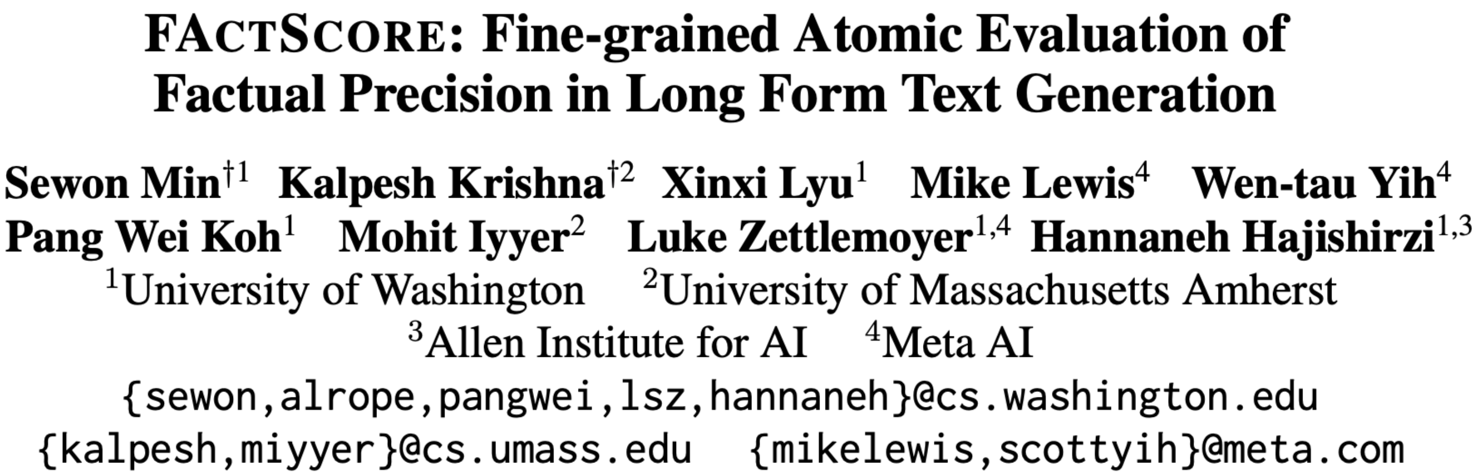
Model	Helpfulness	Validity
ChatGPT	0.1461	43.45%
ChatGPT$_{select\ knowledge}$	**0.2090**	**44.28%**

Table 8: CONNER-guided knowledge selection improves extrinsic (downstream) performance.

CONNER 점수를 활용한 LLM의 Knowledge generation 향상 방안

- Prompt 변경에 따른 CONNER 점수 변동을 하나의 지표로 삼아서 knowledge generation 고도화를 위한 prompt engineering을 시도할 수 있음

- 생성한 Knowledge의 품질을 CONNER 점수로 산출하여서 실제 downstream task 성능 향상을 도모할 수 있음

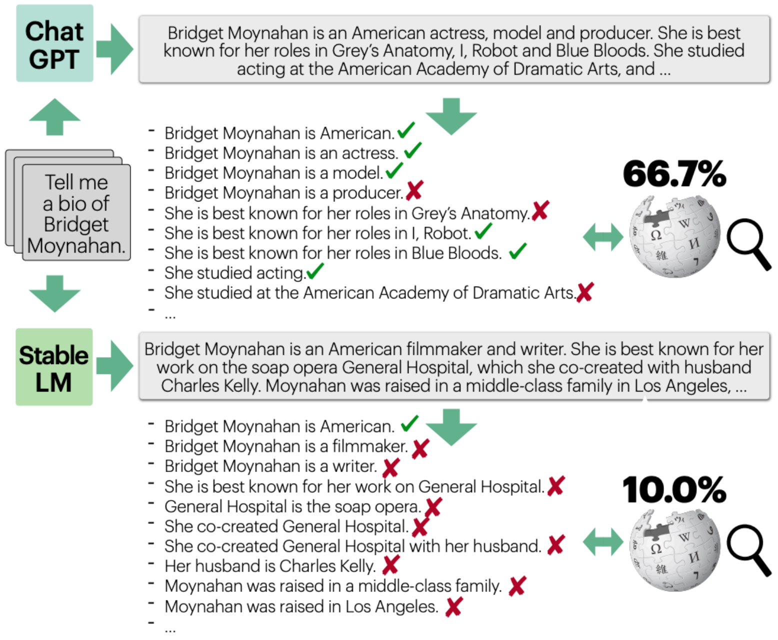


FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation

Sewon Min[†1] Kalpesh Krishna[†2] Xinxi Lyu[1] Mike Lewis[4] Wen-tau Yih[4]
Pang Wei Koh[1] Mohit Iyyer[2] Luke Zettlemoyer[1,4] Hannaneh Hajishirzi[1,3]
[1]University of Washington [2]University of Massachusetts Amherst
[3]Allen Institute for AI [4]Meta AI
{sewon,alrope,pangwei,lsz,hannaneh}@cs.washington.edu
{kalpesh,miyyer}@cs.umass.edu {mikelewis,scottyih}@meta.com

EMNLP 2023 main accepted

Problem Setting



Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

- Bridget Moynahan is American. ✓
- Bridget Moynahan is a filmmaker. ✗
- Bridget Moynahan is a writer. ✗
- She is best known for her work on General Hospital. ✗
- General Hospital is the soap opera. ✗
- She co-created General Hospital. ✗
- She co-created General Hospital with her husband. ✗
- Her husband is Charles Kelly. ✗
- Moynahan was raised in a middle-class family. ✗
- Moynahan was raised in Los Angeles. ✗
- ...

Figure 1: An overview of FACTSCORE, a fraction of *atomic facts* (pieces of information) supported by a given knowledge source. FACTSCORE allows a more fine-grained evaluation of factual precision, e.g., in the figure, the top model gets a score of 66.7% and the bottom model gets 10.0%, whereas prior work would assign 0.0 to both. FACTSCORE can either be based on human evaluation, or be automated, which allows evaluation of a large set of LMs with no human efforts.

LLM generation output의 factual precision을 어떻게 측정할 것인가?

factual precision: the **truthfulness of a statement** should depend on a **particular knowledge source** that end users consider to be trustworthy and reliable

Generation output의 Factual Precision을 측정하기 어려운 이유

1) 하나의 Generation안에는 여러 facts들이 혼재
2) 따라서 혼재된 fact를 개별 atomic fact로 분해하여 검증할 수 있으나
3) 이러한 작업은 인간 수작업으로 진행할 때 매우 큰 시간과 비용 소요

←Factual precision을 정확히 측정할 수 있는 framework와 자동화된 접근 방법 필요

FActScore

- Intuition

가정: Factual Generation은 generation 내부에 담긴 내용이 matching되는 gold knowledge로 검증될 수 있다

↓

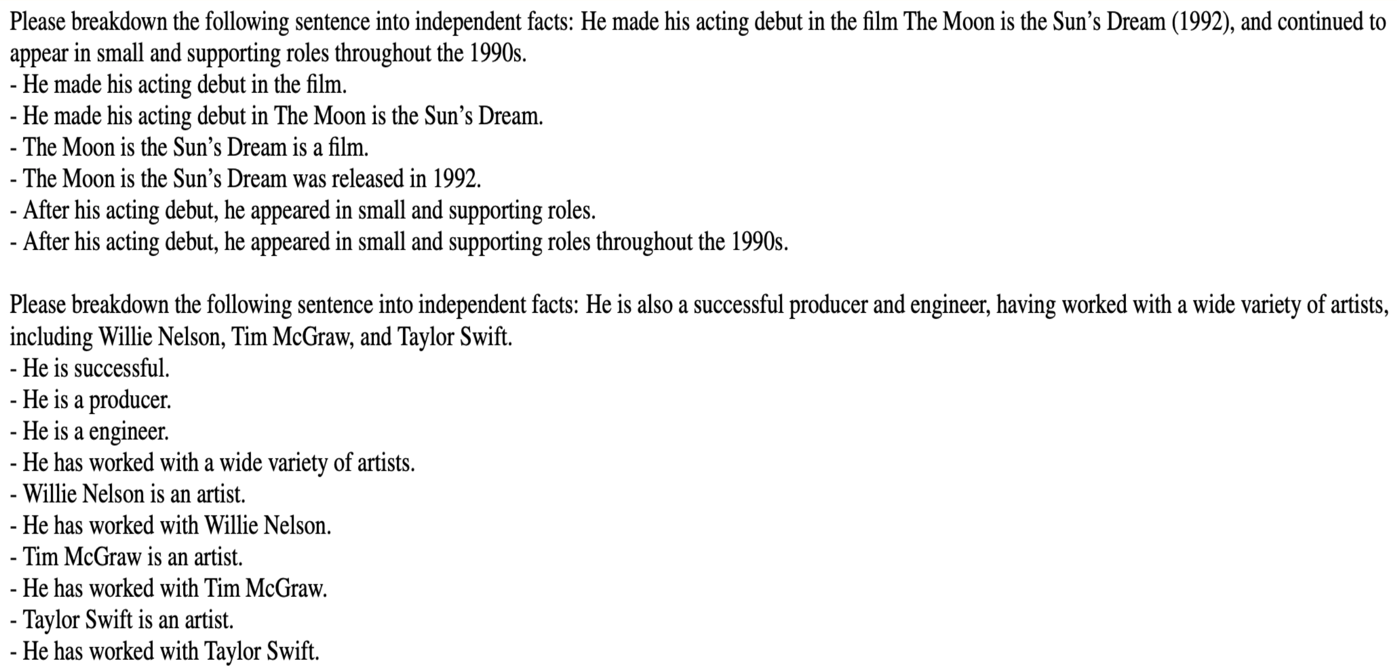
LLM generation에 존재하는 개별 **atomic facts**가 gold knowledge로 증명되는지 측정

이를 위해서는 다음의 선결조건이 충족되어야 함

1) Gold knowledge matching이 이뤄질 수 있는 **objective한 domain**이어야 하며
2) 해당 domain내의 내용들은 다양한 **world-knowledge를 포괄**할 수 있어야 함
3) Gold knowledge를 얻을 수 있는 **knowledge source**가 존재해야 함

Setting: LLM에게 임의의 person entity가 주어질 때, 해당 person의 biography를 서술하도록 하는 상황 설정

FActScore



Please breakdown the following sentence into independent facts: He made his acting debut in the film The Moon is the Sun's Dream (1992), and continued to appear in small and supporting roles throughout the 1990s.
- He made his acting debut in the film.
- He made his acting debut in The Moon is the Sun's Dream.
- The Moon is the Sun's Dream is a film.
- The Moon is the Sun's Dream was released in 1992.
- After his acting debut, he appeared in small and supporting roles.
- After his acting debut, he appeared in small and supporting roles throughout the 1990s.

Please breakdown the following sentence into independent facts: He is also a successful producer and engineer, having worked with a wide variety of artists, including Willie Nelson, Tim McGraw, and Taylor Swift.
- He is successful.
- He is a producer.
- He is a engineer.
- He has worked with a wide variety of artists.
- Willie Nelson is an artist.
- He has worked with Willie Nelson.
- Tim McGraw is an artist.
- He has worked with Tim McGraw.
- Taylor Swift is an artist.
- He has worked with Taylor Swift.

- **Data creation**

1. Sampling people entities

- wikidata에서 183명의 person entity를 랜덤하게 추출
- 추출된 person entity에 대하여 wikipedia 등장 빈도(frequency)와 국적(Nationality)를 labeling

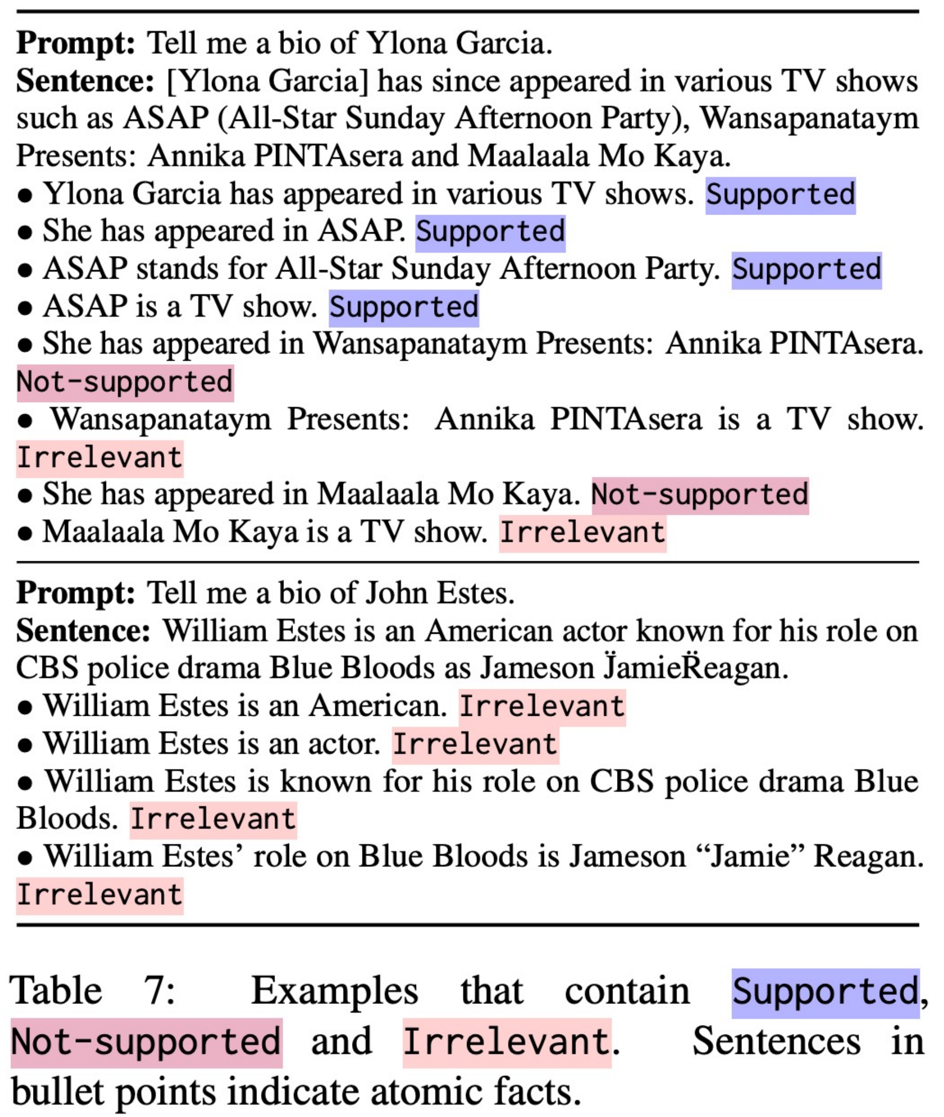
2. Atomic facts generation

- LLM이 생성한 biography에 대하여 human labeler들이 일부의 atomic facts generation을 수행
- 해당 케이스를 예시로 삼아서 InstructGPT(text-davinci-003)에게 ICL수행하여 자동화된 atomic fact generation 수행

3. Labeling factual precision & editing

- Annotator들에게 atomic fact에 대한 검증 및 labeling을 진행: Irrelevant, Supported, Not-supported

FActScore



Prompt: Tell me a bio of Ylona Garcia.
Sentence: [Ylona Garcia] has since appeared in various TV shows such as ASAP (All-Star Sunday Afternoon Party), Wansapanataym Presents: Annika PINTAsera and Maalaala Mo Kaya.

- Ylona Garcia has appeared in various TV shows. `Supported`
- She has appeared in ASAP. `Supported`
- ASAP stands for All-Star Sunday Afternoon Party. `Supported`
- ASAP is a TV show. `Supported`
- She has appeared in Wansapanataym Presents: Annika PINTAsera. `Not-supported`
- Wansapanataym Presents: Annika PINTAsera is a TV show. `Irrelevant`
- She has appeared in Maalaala Mo Kaya. `Not-supported`
- Maalaala Mo Kaya is a TV show. `Irrelevant`

Prompt: Tell me a bio of John Estes.
Sentence: William Estes is an American actor known for his role on CBS police drama Blue Bloods as Jameson J̈amieR̈eagan.

- William Estes is an American. `Irrelevant`
- William Estes is an actor. `Irrelevant`
- William Estes is known for his role on CBS police drama Blue Bloods. `Irrelevant`
- William Estes' role on Blue Bloods is Jameson "Jamie" Reagan. `Irrelevant`

Table 7: Examples that contain `Supported`, `Not-supported` and `Irrelevant`. Sentences in bullet points indicate atomic facts.

- **Data creation**

1. Sampling people entities

- wikidata에서 183명의 person entity를 랜덤하게 추출
- 추출된 person entity에 대하여 wikipedia 등장 빈도(frequency)와 국적(Nationality)를 labeling

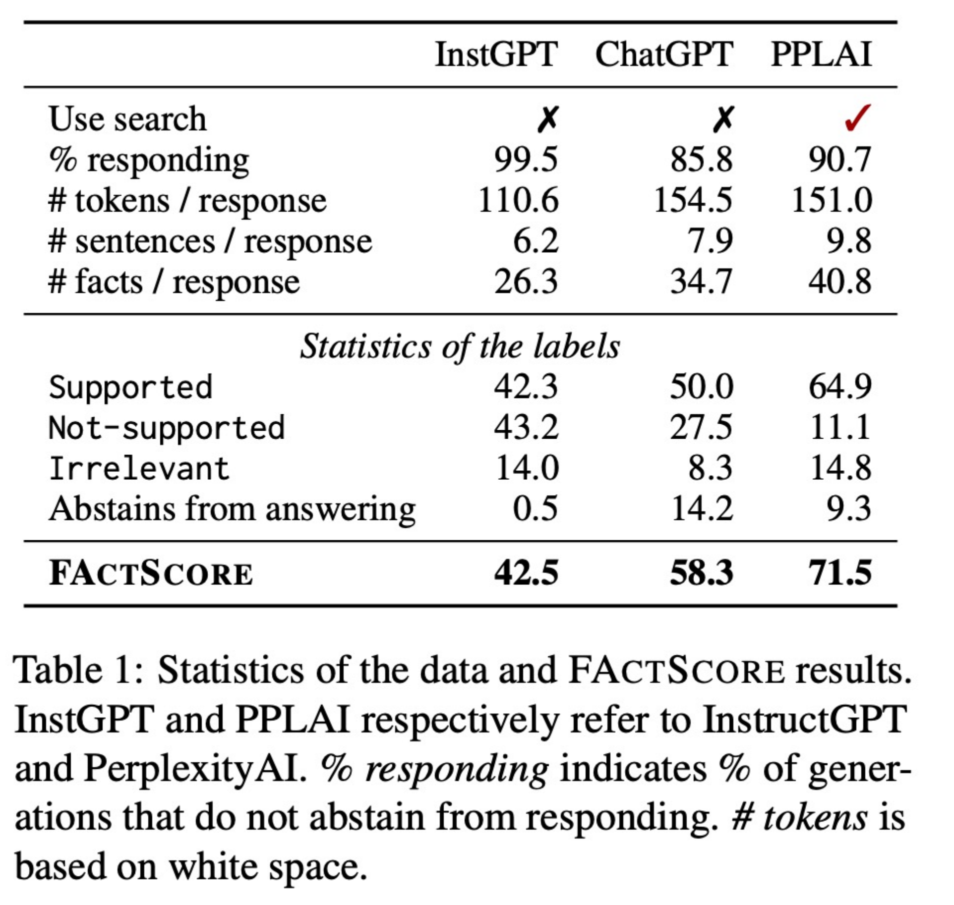
2. Atomic facts generation

- LLM이 생성한 biography에 대하여 human labeler들이 일부의 atomic facts generation을 수행
- 해당 케이스를 예시로 삼아서 InstructGPT(text-davinci-003)에게 ICL수행하여 자동화된 atomic fact generation 수행

3. Labeling factual precision & editing

- Annotator들에게 atomic fact에 대한 검증 및 labeling을 진행: Irrelevant, Supported, Not-supported

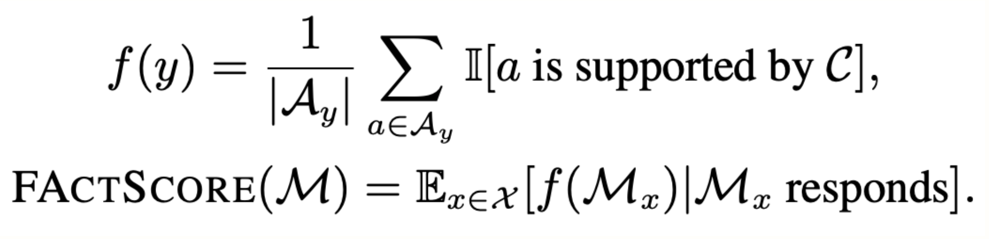
FActScore



	InstGPT	ChatGPT	PPLAI
Use search	✗	✗	✓
% responding	99.5	85.8	90.7
# tokens / response	110.6	154.5	151.0
# sentences / response	6.2	7.9	9.8
# facts / response	26.3	34.7	40.8
Statistics of the labels			
Supported	42.3	50.0	64.9
Not-supported	43.2	27.5	11.1
Irrelevant	14.0	8.3	14.8
Abstains from answering	0.5	14.2	9.3
FACTSCORE	**42.5**	**58.3**	**71.5**

Table 1: Statistics of the data and FACTSCORE results. InstGPT and PPLAI respectively refer to InstructGPT and PerplexityAI. *% responding* indicates % of generations that do not abstain from responding. *# tokens* is based on white space.

- Measuring Factscore



$$f(y) = \frac{1}{|\mathcal{A}_y|} \sum_{a \in \mathcal{A}_y} \mathbb{I}[a \text{ is supported by } \mathcal{C}],$$

$$\text{FACTSCORE}(\mathcal{M}) = \mathbb{E}_{x \in \mathcal{X}}[f(\mathcal{M}_x) | \mathcal{M}_x \text{ responds}].$$

Notation
$\mathcal{X}$: set of prompts, $\mathcal{C}$: knowledge source
$\mathcal{M}$: language model, $\mathcal{A}_y$: list of atomic facts

InstructGPT, ChatGPT, Perplexity AI의 183개 people entity에 대한 generation 결과에 대한 Factscore 측정

☞ 세 LLM모두 factual precision이 담보되지 못한 생성을 수행하는 것을 확인할 수 있음
(특히 Retriever가 없는 LLM의 경우)

FActScore

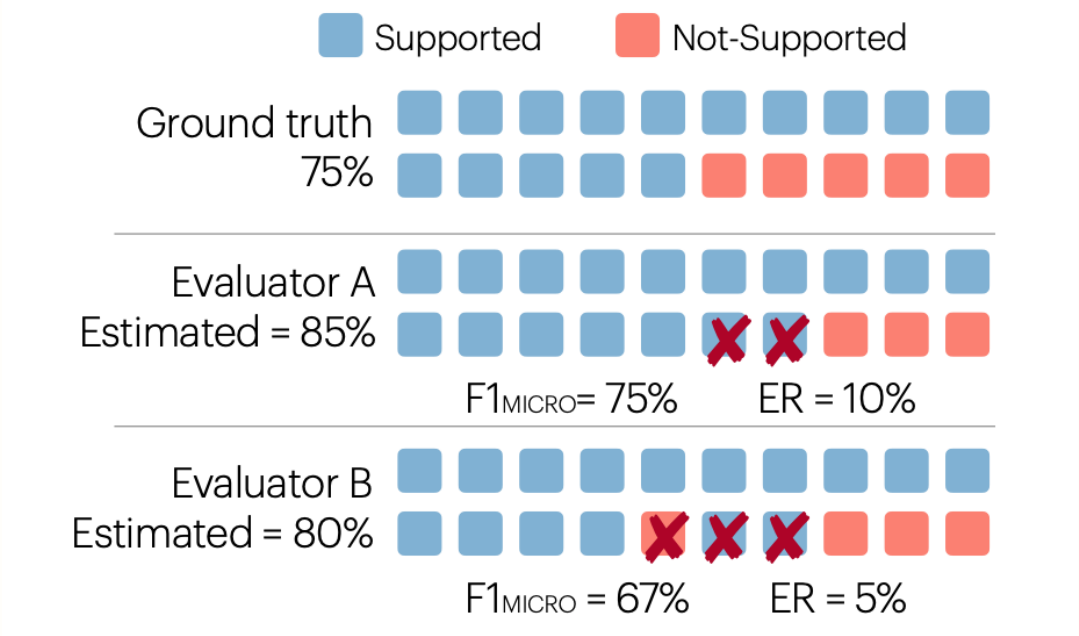


Figure 4: A case in which $F1_{MICRO}$ and Error Rate (ER) rank two evaluators differently. Evaluator A is better in $F1_{MICRO}$, and Evaluator B is better in ER.

- Estimating Factscore

Inst-LLaMA, ChatGPT를 이용하여서 자동화된 Factscore 측정을 시도

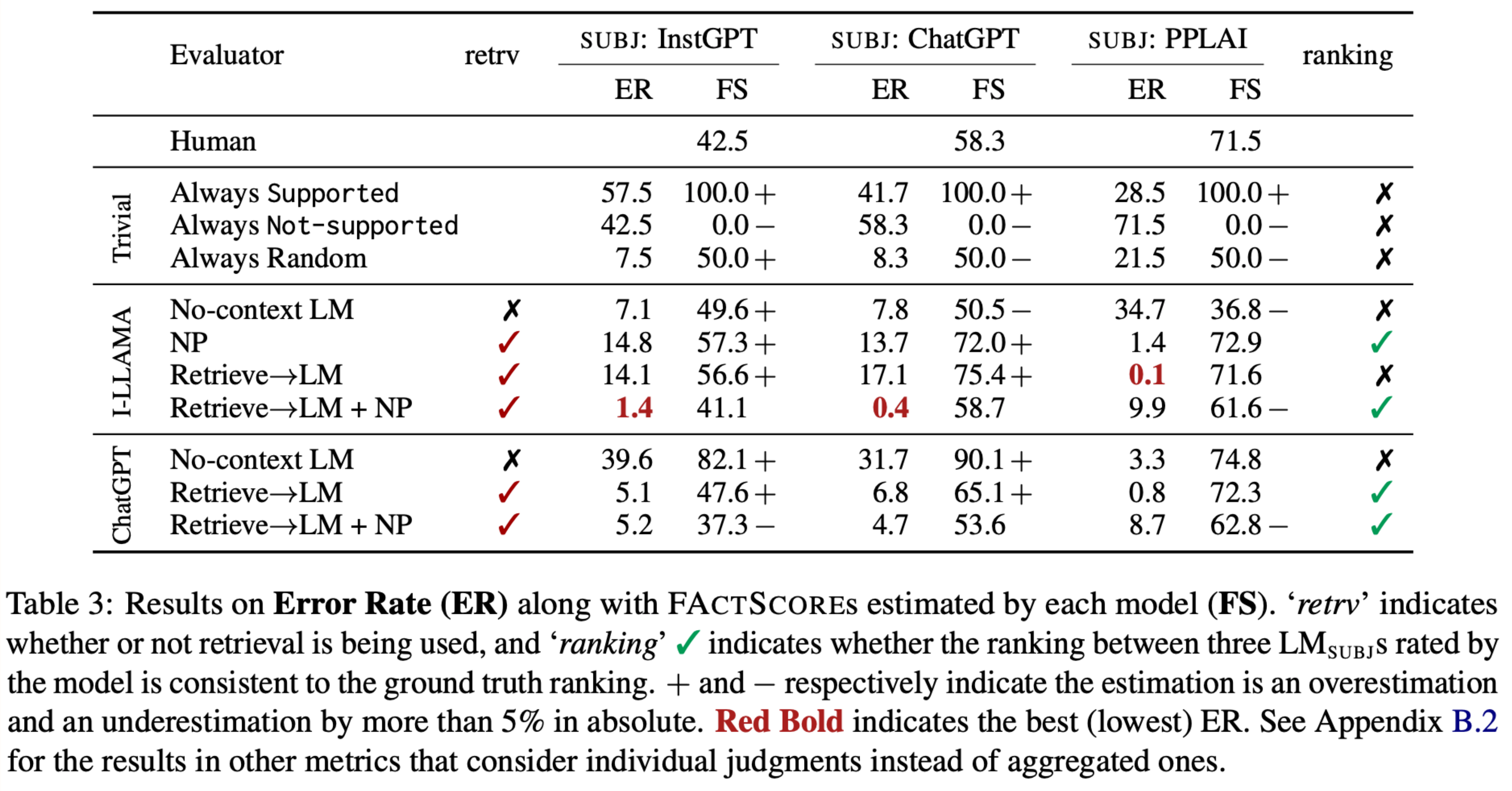
Ground truth Factscore와 Estimated Factscore의 차이 ER(error rate)을 기준으로 가장 타당한 방법 탐색

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

Main result

- Quantitative Result



	Evaluator	retrv	SUBJ: InstGPT		SUBJ: ChatGPT		SUBJ: PPLAI		ranking
			ER	FS	ER	FS	ER	FS	
	Human		42.5		58.3		71.5		
Trivial	Always Supported		57.5	100.0 +	41.7	100.0 +	28.5	100.0 +	✗
	Always Not-supported		42.5	0.0 –	58.3	0.0 –	71.5	0.0 –	✗
	Always Random		7.5	50.0 +	8.3	50.0 –	21.5	50.0 –	✗
I-LLAMA	No-context LM	✗	7.1	49.6 +	7.8	50.5 –	34.7	36.8 –	✗
	NP	✓	14.8	57.3 +	13.7	72.0 +	1.4	72.9	✓
	Retrieve→LM	✓	14.1	56.6 +	17.1	75.4 +	**0.1**	71.6	✗
	Retrieve→LM + NP	✓	**1.4**	41.1	**0.4**	58.7	9.9	61.6 –	✓
ChatGPT	No-context LM	✗	39.6	82.1 +	31.7	90.1 +	3.3	74.8	✗
	Retrieve→LM	✓	5.1	47.6 +	6.8	65.1 +	0.8	72.3	✓
	Retrieve→LM + NP	✓	5.2	37.3 –	4.7	53.6	8.7	62.8 –	✓

Table 3: Results on **Error Rate (ER)** along with FActScores estimated by each model (**FS**). '*retrv*' indicates whether or not retrieval is being used, and '*ranking*' ✓ indicates whether the ranking between three LM$_{\text{SUBJ}}$s rated by the model is consistent to the ground truth ranking. + and – respectively indicate the estimation is an overestimation and an underestimation by more than 5% in absolute. **Red Bold** indicates the best (lowest) ER. See Appendix B.2 for the results in other metrics that consider individual judgments instead of aggregated ones.

- Retriever를 적용할 때 error rate가 감소

- Retriever가 결합되었을 때 대체로 ChatGPT가 좋은 성능을 보이나

- 최고 성능 case는 모두 LLaMA에서 도출

Main result

- Case study: Evaluation of New LMs

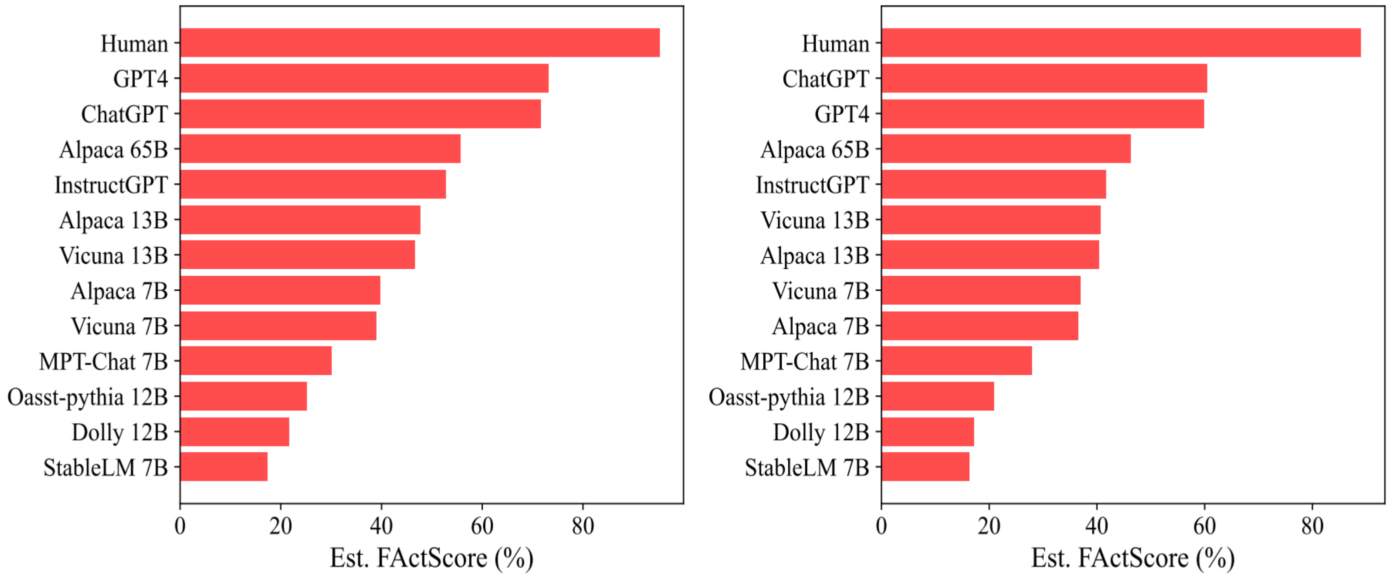


Figure 3: Ranking between 13 subjects (human and 12 LMs), rated by the two best variants of our estimator: ChatGPT (**left**) and LLAMA+NP (**right**), both with retrieval. Scores from two metrics have a Pearson's r of 0.99. See Table 5 for % of responding and # of atomic facts per response of each LM. The variance in estimation based on different subsets of prompts is reported in Figure 5 of Appendix B.4.

- GPT-4, Alpaca 등 논문 작성 당시 최신 10개의 LLM에 대하여 Factscore estimation 방법을 적용해 Factscore 산출

- 모든 LLM은 Human 대비 generation에서 Factual precision이 낮음

- Alpaca, Vicuna같은 Open-LLM들 간의 Factscore 격차가 큰 것을 알 수 있음

☞ 단 이 분석이 본 논문이 설정한 Biography generation에 한정되는 것을 유념해야 할 필요 있음

Discussion

- Pros
  - LLM generation output의 Factuality를 측정하고자 할 때 precision, recall 관점을 견지해서 분석할 수 있다
    - 특히 Factual Recall, 즉 generation output에 담긴 atomic fact가 world knowledge를 얼마나 포괄하는지에 관한 연구가 필요
  - 단순 generation output자체의 Factuality를 측정하는 것에 그치는 것이 아니라 실제 task 수행에 있어서 어떤 영향을 끼치는지 분석하는 것도 연구의 큰 의의를 가질 수 있음
- Cons
  - CONNER와 같이 모든 모듈마다 다른 model을 사용하는 비효율적이고 heuristic한 framework에서 벗어나 하나의 모델을 활용한 systematic한 방법이 필요
  - Generation factuality측정이 정확이 보장이 된 이후의 어떤 Mitigation 전략을 취할 것인가에 대한 연구는 아직 요원함

Thank you