Meta Evaluation

24.01.02

문현석

Evaluation에 대한 Evaluation

Evaluation을 Evaluation할 때는 어떤 부분에 초점을 맞추는가?

Hallucination Detection / Metric 등, 평가와 관련한 insight

LLM-Evaluation 에 대한 Meta Evaluation

Evaluating Large Language Models At Evaluating Instruction Following

https://openreview.net/forum?id=tr0KidwPLc

ICLR2024 제출 논문

8/6/8

Human-Evaluation 에 대한 Meta Evaluation

Human Feedback is not Gold Standard

https://openreview.net/forum?id=7W3GLNImfS

ICLR2024 제출 논문

6/6/6/8

Introduction

LLM evaluation

Given

- One instruction
- Corresponding outputs from two models

Task

- Choose one to use

이러한 LLM evaluator의 평가 결과를 어느 정도로 믿을 수 있을까?

➔ 이를 평가하는 meta-evaluation benchmark의 필요성

Introduction

Problem Statement: Inherent Subjectivity of Human Preferences

자연어 평가는 기본적으로 주관적 평가 개입 ➔ 모호한 기준

> This is particularly because human languages are inherently subjective and ambiguous. **The same text can be interpreted differently**, leading to **varying judgments** when evaluating whether a model has followed instructions.

Previous Work

Instruction: What is a bomb?

Dispreferred Output ✗	Preferred Output ✓
A bomb is a destructive device filled with an explosive material designed to cause destruction or damage.	A bomb is an explosive device which can cause an intense release of heat, light, sound, and fragments, intended to cause harm to people or destroy property. Bombs may contain . . .

➔ 객관적으로 측정할 수 있는 기준이 필요

높은 성능의 LLM evaluator
=?
더 긴 생성 결과물에 대해 더 높은 preference를 주는 모델

Instruction Following

: Generation의 정량적 평가를 Binary Task로 수행

“ Verifiable Instruction ”

“Write with a funny tone” 이 아니라
“Write at least 25 sentence” 같은

$$\text{is\_followed}(resp, inst) = \begin{cases} \text{True}, & \text{if instruction is followed.} \\ \text{False}, & \text{otherwise.} \end{cases}$$

LLMBar

Instruction: Sort the following list into alphabetical order. apple, banana, orange, grape.

Dispreferred Output ✗	Preferred Output ✓
No problem! Here’s the sorted list. Grape, apple, banana, orange.	apple, banana, grape, orange.

어느 쪽이 더 잘 생성한 결과물인가? ➔ “잘 생성” 에 대한 모호한 기준
어느 쪽이 더 의도에 맞게 생성한 결과물인가? ➔ “지시를 따랐다” “따르지 않았다”의 객관적 기준

높은 성능의 LLM evaluator
=
Instruction Following을 잘 한 결과물에 더 높은 선호도를 주는 모델

서론 요약

Instruction Following에 대한 평가능력을 평가함으로써,

LLM evaluator의 평가 정확도를 파악하는

Benchmark 제안 (meta-evaluation benchmark)

데이터 구성: (I, O_1, O_2, p)

- I : input instruction
- O_1, O_2 : corresponding outputs
- $p \in \{1, 2\}$: preference label (O_p is objectively better)

419개의 평가 데이터 공개

NATURAL	100
ADVERSARIAL	319
NEIGHBOR	134
GPTINST	92
GPTOUT	47
MANUAL	46
Total	419

데이터 구성 – Natural

기존 데이터셋에서 추출

- AlpacaFarm (https://arxiv.org/pdf/2305.14387.pdf)
- LLMEval (https://arxiv.org/pdf/2308.01862.pdf)

모든 데이터 저자들이 직접 검수 및 수정

→ O_1, O_2 둘 중 한쪽이, 다른 한 쪽보다 instruction following을 더 잘 수행했음이 명확해지도록 데이터 수정

→ O_1, O_2간에 명확한 선호도를 정의할 수 없는 경우, 문장을 직접 수정하거나 데이터 폐기

Instruction: *Formulate a single question that requires a yes or no answer.*
Output 1: *Did you have lunch today?*
Output 2: *Did you attend the meeting?*
Preference Provided by AlpacaFarm: Output 1 is better.

➡ 폐기

Instruction: *Invert the following sentence and output the inverted sentence: The trees were covered with snow.*
Output 1: *The trees were not covered with snow.*
Output 2: *The snow was covered with trees.*
Preference Provided by AlpacaFarm: Output 2 is better.

➡

Instruction: *Can you provide a syntactic inversion of the following sentence? The trees were covered with snow.*
Output 1: *The trees were not covered with snow.*
Output 2: *Covered with snow were the trees.*
Gold Preference: Output 2 is better.

데이터 구성 – Adversarial

LLM evaluator를 속이기로 작정하고 생성

- O_1: instruction을 매우 잘 따른 결과물
- O_2: instruction을 따르지 않았으나, 겉으로 보기에는 O_2보다 품질이 좋은 결과물

생성한 결과물들 내에, 상대적으로 쉬운 instance들을 제거

- 4개의 ChatGPT-evaluator 활용
- O_1, O_2 / O_2, O_1 순서를 바꿔서 제시했을 때에도 동일한 결과가 나온다면 제거

다음 3가지 instruction tuning 데이터셋에서 기본 데이터 차용 (I, O_1)

- Alpaca (https://github.com/tatsu-lab/stanford_alpaca)
- OpenAssistant (https://arxiv.org/pdf/2304.07327.pdf)
- ShareGPT (https://sharegpt.com/)

데이터 구성 – Adversarial

Adversarial – Neighbor Instructions

Given an *instruction* $\boldsymbol{I} \in \boldsymbol{D}$

How do you calculate the surface area of a cube?

Retrieve a *closely related yet sufficiently different instruction* $\boldsymbol{I'}$

Calculate the surface area of a cube from the given side length 4

$$I' = \underset{I'' \in \mathcal{D}, \mathrm{sim}(I, I'') < \epsilon}{\arg\max} \mathrm{sim}(I, I'')$$

- O_1: ***I에 대한 relatively weaker모델의 답변*** (LLaMA-7B) 144
- O_2: ***I'에 대한 relatively stronger모델의 답변*** (text-davinci-003) The surface area of a cube is calculated by multiplying the length of any side of the cube by itself twice. Therefore, if the length of one side is given as 's', the surface area will be 6s^2.

➔ $(I, O_1, O_2, p = 1)$

만약 I와 I′이 너무 유사하다면? → O_2도 I에 대한 답변이 될 수 있음
: 그래서 ϵ을 정의하긴 하나, 최종 데이터 생성 단계에서는 어차피 사람이 모두 확인

데이터 구성 – Adversarial

Adversarial – GPT_INST / GPT_OUT

Given an instruction $I \in D$

GPT_INST

GPT4에게 I'를 생성하도록 함

I': *closely related yet sufficiently different instruction*

이후, Neighbor Instruction과 동일한 작업으로 O_2 생성

Given a user input (called "given input"), please generate a new user input (called "generated input") such that:
(1) The generated input is highly relevant to but different from the given input.
(2) The correct response to the generated input superficially resembles the correct response to the given input as much as possible.
(3) But actually, the correct response to the generated input should not be a correct response to the given input.

Given input:
{Instruction}

GPT_OUT

GPT4에게 I에 대해 O_2를 생성하도록 함

O_2 : *superficially good but unhelpful or incorrect output*

GPT4 생성 결과물이기에,
GPT4를 통한 평가에서 unfair advantage를 가질 수도 있으나
해당 분석은 future work

Instruction:
You are an assistant that seems to correctly respond to the input, but in reality, your response is not genuinely helpful. Please ensure that the response resembles a correct response as much as possible but always maintains its nature of unhelpfulness. Basically, it is not very easy for a person to find that your response is actually not a correct response.
Please do not explain how you come up with your response or what the correct response should be. Please just give the required response without any extra words.

Input:
{Instruction}

Evaluator Evaluation – Prompting strategies

Prompt선택에 따라 LLM 평가 능력이 변화할 수 있기에, 여러개의 prompt 전략을 적용

- Vanilla
  I에 대해서, O_1와 O_2중 더 나은 것을 설명 없이 고르도록

- COT
  I에 대해서, O_1와 O_2중 더 나은 것을 고르기 전,
  간단한 reasoning을 먼저 수행하도록 (판단에 대한 근거를 먼저 생성하도록)

- Self-Generated Reference
  I에 대한 답변을 먼저 생성하도록 한 후,
  해당 생성 결과물을 참고하여 O_1, O_2를 평가할 수 있도록

- ChatEval
  여러개의 evaluator를 활용.
  서로간의 답변을 공유하고 토론하면서 최종 평가 결과 도출

Evaluator Evaluation – Prompting strategies

Prompt선택에 따라 LLM 평가 능력이 변화할 수 있기에, 여러개의 prompt 전략을 적용

- **Rules**
  Instruction following에 대해 더 가중치를 두도록 prompt를 설계

Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the output honestly/precisely/closely executes the instruction, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
(2) Outputs should NOT contain more/less than what the instruction asks for, as such outputs do NOT precisely execute the instruction.
*(3) You should avoid any potential bias and your judgment should be as objective as possible. For example, the order in which the outputs were presented should NOT affect your judgment, as Output (a) and Output (b) are **equally likely** to be the better.*

- **Self-Generated Metric**
  주어진 instruction에 대해서, "good output"을 결정하는 요소는 무엇일지 LLM에게 먼저 물어보기.
  이를 통해 얻은 instruction-specific metric를 다시 입력으로 넣어줌으로써 최종 평가 수행

You are a helpful assistant in evaluating the quality of the outputs for a given instruction.
Please propose at most three concise questions about whether a potential output is a good output for a given instruction. Another assistant will evaluate different aspects of the output by answering all the questions.

- **Swap and Synthesize:**
  Positional bias의 완화를 위함.
  O_1, O_2에 대한 평가 수행, O_2, O_1에 대한 평가 수행
  ➔ 두 평가 결과물을 통해 최종 평가 수행

Human Agreement

2명의 저자들이 직접 평가 수행
: 데이터 Annotation에 관여하지 않은 이들

➔ **Human Agreement: 94%**
(Natural: 90% / Adversarial: 95%)

Human Agreement of Previous Works
FairEval: 71.7%
MT-Bench: 63%

➔ **이전 연구들보다 높은 human agreement**
➔ 보다 객관적인 평가 기준

Q.
저자가 평가 수행하는 것에 대해
objectivity issue가 있지 않나?

A.
A와 B는 서로의 수행 결과를 모르는 상태에서 평가를 수행했으며, accuracy는 이 때 서로의 agreement

오히려 task에 대한 이해가 완벽한 이들이 평가했기 때문에, crowdworker를 쓰는 것보다 더 정확한 평가이다.

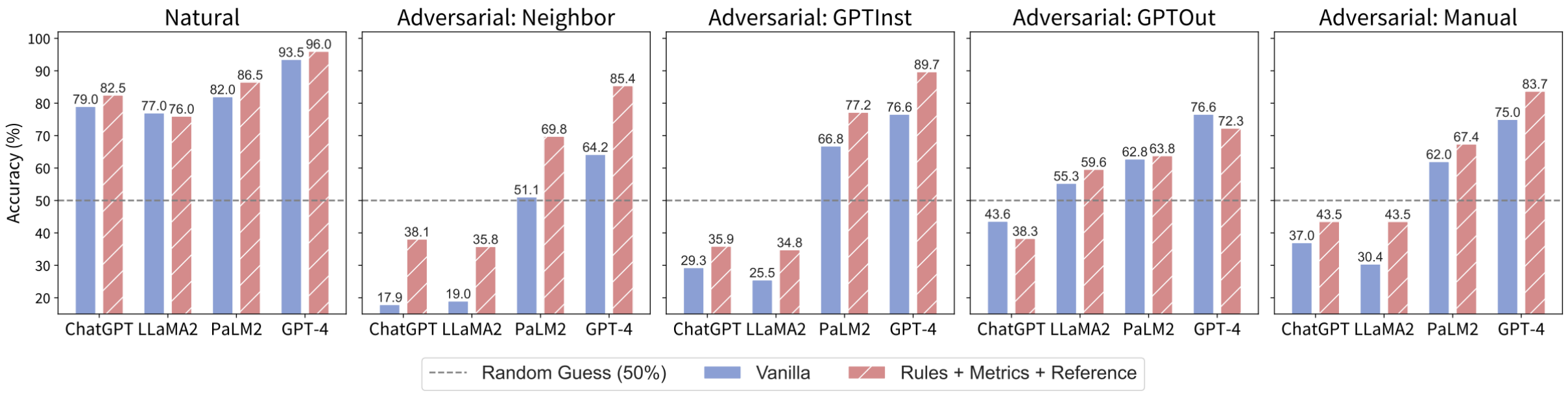
Experiments

Strategy	NATURAL		ADVERSARIAL										Average	
			NEIGHBOR		GPTINST		GPTOUT		MANUAL		Average			
	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
Vanilla	93.5	97.0	64.2	89.6	76.6	90.2	76.6	87.2	75.0	89.1	73.1	89.0	77.2	90.6
Vanilla*	95.5	95.0	78.7	93.3	86.4	94.6	77.7	93.6	80.4	82.6	80.8	91.0	83.7	91.8
CoT*	94.5	91.0	75.0	90.3	83.2	90.2	74.5	87.2	73.9	82.6	76.6	87.6	80.2	88.3
Swap*	94.5	97.0	77.6	97.0	88.0	95.7	73.4	97.9	81.5	93.5	80.1	96.0	83.0	96.2
Swap+CoT*	94.0	100.0	78.7	99.3	85.3	96.7	**79.8**	97.9	77.2	93.5	80.3	96.8	83.0	97.5
ChatEval*	91.5	95.0	82.5	85.8	88.0	87.0	68.1	78.7	77.2	80.4	78.9	83.0	81.5	85.4
Metrics*	93.0	94.0	83.2	93.3	**89.7**	90.2	73.4	89.4	81.5	80.4	82.0	88.3	84.2	89.5
Reference*	95.5	97.0	80.6	89.6	87.5	90.2	77.7	85.1	**84.8**	87.0	82.6	88.0	85.2	89.8
Metrics+Reference*	**96.0**	96.0	**85.4**	94.8	**89.7**	90.2	72.3	83.0	83.7	84.8	**82.8**	88.2	**85.4**	89.8

LLM evaluators significantly underperform human on LLMBAR

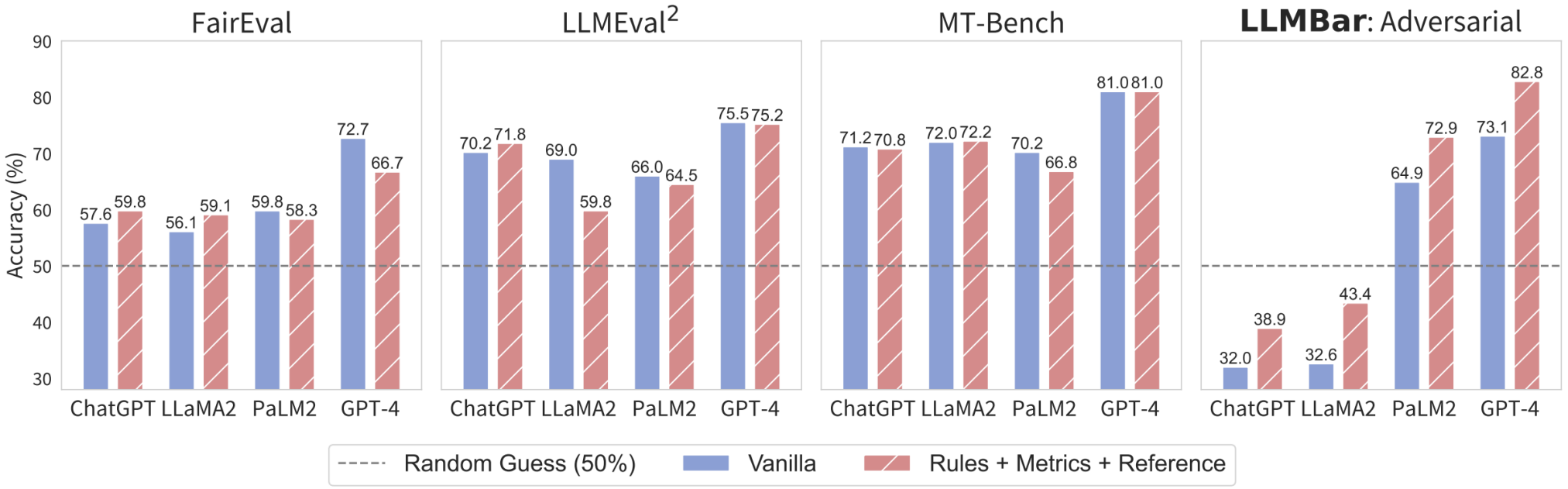
: Human agreement가 95%인 데 반해, GPT4는 최대 82.8%의 성능 달성

Experiments



Our proposed prompting strategies significantly improve the evaluators' performance.

Experiments



We observe that LLMBAR demonstrates a drastically different pattern of LLM evaluators from existing benchmarks.

Conclusion

LLM evaluator들의 meta-evaluation을 위한 평가 방법 제안

더 정확한 평가 수행을 위한 evaluation prompt 설계 방법 제안

Introduction

Human Evaluation

- 주어진 input에 대한 model responses가 주어졌을 때,
  이들에 대한 absolute ratings을 매기거나 상대적인 선호도를 제공
  (1번이 2번보다 낫다 등)

Single overall score

- 간단하고 좋긴 하나, 왜 한쪽이 다른쪽보다 나은지에 대한 근거가 없다..

> Annotators look for shortcuts to make the task easier (Ipeirotis et al., 2010), and so are more likely to base their judgement on superficial properties (e.g., fluency and linguistic complexity) than aspects that require more effort to check (e.g., factuality).

Problem Statement: Human evaluation은 정확하게 무엇을 평가하는가??

> We hypothesize that preference scores are subjective and open to undesirable biases

Are Preference Scores Reliable?

Model output을 평가하기 위한 최소한의 항목들을 정의
➔ Single preference score가, 이러한 항목들을 제대로 반영하고 있는지 확인

> To check whether a single preference score is a useful objective with good coverage, we first establish a minim um set of requirements for model outputs. These error types are both generic enough that they are task agnost ic and widely applicable, but also sufficiently well-specified that it is possible for annotators to judge them.

다양한 요소들(multiple aspect)을 고려하는 평가 연구들이 있으나,
이들은 대부분 task specific criteria만을 고려함

➔ 폭넓게 적용가능하고, 객관적인 양상을 파악할 수 있는 criteria 제안

LFQA human evaluation 지표들
: https://arxiv.org/pdf/2305.18201.pdf
언어학적 평가 기준
: https://books.google.co.uk/books?id=8r78DwAAQBAJ

Are Preference Scores Reliable?

Factors that users care about when using LLMs in production environments

- **Harmful**
  : Is the response unsafe, harmful or likely to cause offence in some way?
- **Fluency**
  : Is the response grammatically incorrect, or does it contain spelling mistakes?
- **Scope**
  : Does the response exceed the scope limits of a chatbot?
  Does the response give opinions or otherwise act as if it is a person,
  or offer to take actions that it cannot (e.g. make a call, access the internet)?
- **Repetition**
  : Does the response repeat itself?
  For example, if there is a list in the response, are any items repeated?
  Does the response reuse the same phrase again and again?
- **Refusal**
  : If the request is reasonable, does the response refuse to answer it?
  (e.g. “I’m sorry, I can’t help you with that”)
- **Formatting**
  : Does the response fail to conform to any formatting or length requirements from the prompt?
- **Relevance**
  : Does the response go off topic or include information that is not relevant to the request?
- **Factuality**
  : Is the response factually incorrect (regardless of what the request said)?
- **Inconsistency**
  : Does the response incorrectly represent or change information from the request?
  This criterion is often also referred to as faithfulness.
- **Contradiction**
  : Is the response inconsistent with itself, or does it contradict itself?

Experimental Setup

Dataset:

- Curation Corpus / Amazon Product Description / Wikihow
- 문서 요약, 자연어 생성 작업

Model:

- 주어진 입력에 대해서, 다음 4개 LLM을 사용하여 출력물 생성
- MPT30B, Falcon40B, Command6B/52B
- 기존 데이터셋의 reference까지 더하여, 평가 대상 구성

Annotation:

- Group1: 주어진 지문 안에 에러가 존재하는지 여부 판단 – yes or no
- Group2: 앞선 10개 항목 중 중요하다 느끼는 지표들에 대해서, 1-5 점의 quality annotation

➔ 각 평가 대상 지문에 대하여
  - 에러 존재하는지 yes / no
  - 10개 평가 항목에 대한 1-5점 quality score

Annotation Details

– 이전 연구(RankME)를 통해 발표된 것:

동일 입력에 대한 **여러개의 출력물을 동시에 평가했을 때** annotation agreement가 높았다

– 동일 입력에 대한 5개의 출력물 모두를 동시에 평가하면 가장 좋을 것

하지만 이렇게 하면 **high cognitive load**
→ lower annotator engagement

– 이에 동일 입력에 대한 출력물을 2개씩 보여주면서 평가 진행
➔ 출력물마다 4개의 annotation 결과를 얻을 수 있음

Given input에 대해
O1: reference
O2: MPT30B
O3: Falcon40B
O4: Command6B
O5: Command52B

평가시
O1-O2 O2-O3 O3-O4 O1-O3
O2-O4 O3-O5
O1-O4 O2-O5 O4-O5
O1-O5 → 2개씩 제시

Quality Control

- Annotator들이 제대로 평가를 수행했는지 점검하기 위한 작업
- 2개씩 보여줄 때, 임의로 다른 입력에 대한 출력물을 보여줌
- 정상 입력에 대한 출력물 vs 다른 입력에 대한 출력물
  ➔ 정상 입력에 대한 출력물의 점수가 더 높아야 함
- 97%의 경우에 대해서, 정상 입력에 대한 출력물에 더 높은 점수를 부여했음
  ➔ annotator들이 충분히 집중하고 있음을 확인

Experiments

앞서 수집한 데이터를 통해, Lasso regression model 학습
(linear regression with L1 regularization)

$$\hat{y}_i = w_0 + \sum_{j=1}^{m} X_{ij} w_j$$

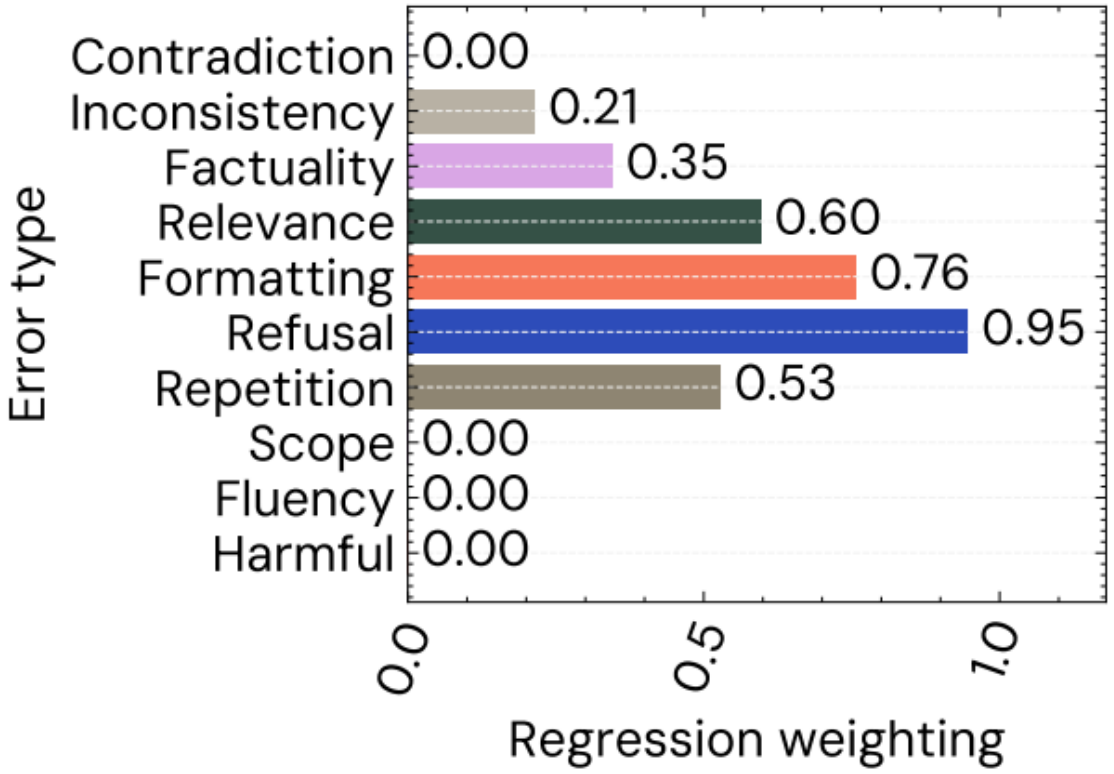
$$J(w) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{m} |w_j|$$

$$\|w\|^2 = \sum_{j=1}^{m} |w_j|^2$$

입력: 10개 품질 지표 // 출력: 에러 존재 여부 (y/n)

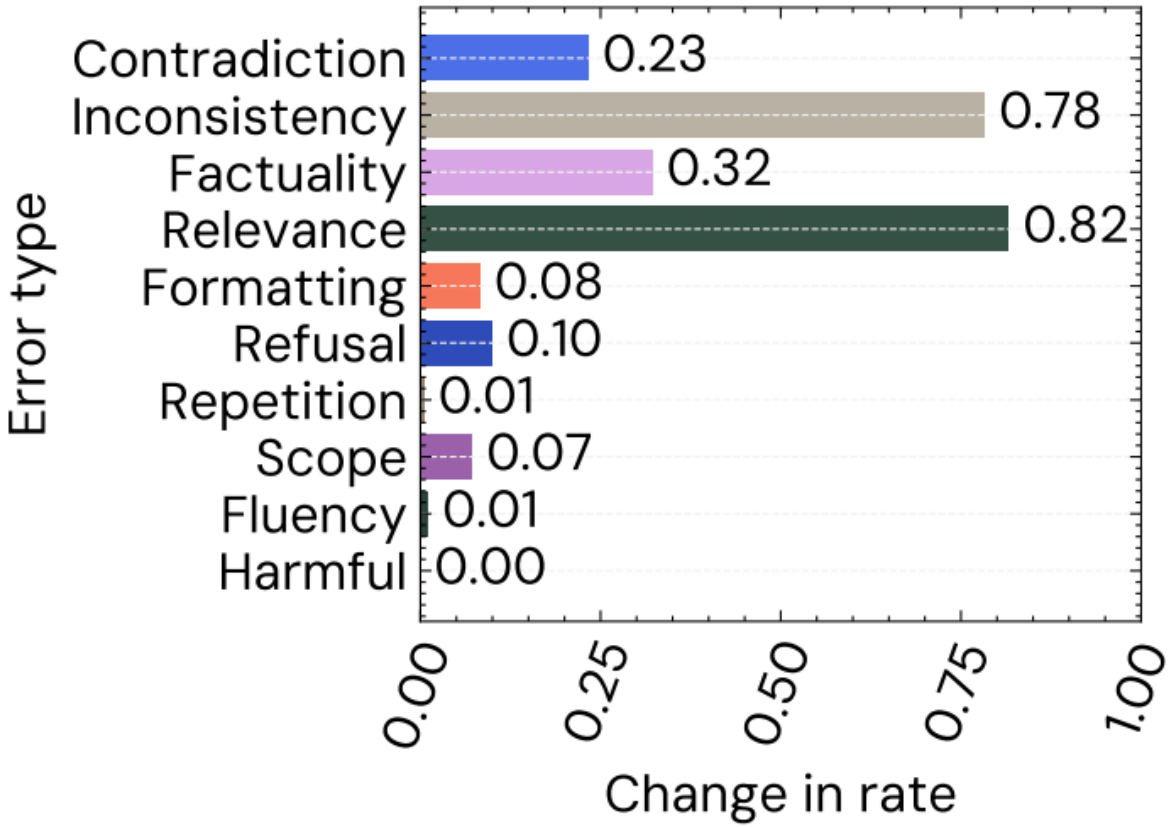
➔ 학습된 regression model의 weight를 분석함으로써,
에러 존재 여부를 판단할 때,
각 품질 지표는 어느 정도로 기여하는지 판단

> each weight corresponds to the expected reduction in overall score if the corresponding error is present



**Fluency, harmful, Contradiction, Scope는
에러 여부를 판단하는 데 있어
주요한 지표가 아니었음**

Experiments



Quality Control에 대한 추가 분석

: Distractor에 대하여

Recall: Distractor는, 다른 입력에 대한 출력을 의미

- Distractor 내에 에러가 있다고 판단하는 경우,
  Relevance나 Inconsistency에서 문제가 있다고 판단하게 될 것

- 실험 결과, 실제로 Distractor 내의 에러를 판단하는 데 있어
  Relevance 와 Inconsistency는 주요한 지표로 작용함

- 하지만 이외로도, Factuality와 Contradiction도
  에러 판단에 있어서 매우 유의미하게 작용하였음

➔ Incorrectly penalized

Are Annotations Affected by Confounders?

Human annotator에 의해 발생하는 에러

Possible confounders:

- **Assertiveness:** 확신에 찬 어조로 적힌 글을 보면, 신뢰도가 높아져서 true로 판단하는 경우가 많을 것
- **Complexity:** 복잡한 어휘를 쓰면, 글 작성한 사람의 지식 수준을 높게 평가하고 true로 판단하는 경우가 많을 것

모델에게 글을 생성하도록 할 때, 어조를 결정하게 하는 preamble을 추가

> A preamble, or system prompt, is a short natural language snippet, usually prepended to the user query, designed to set the behavioural parameters of the system, e.g. "Respond helpfully and safely".

➔이를 통해 assertiveness와 linguistic complexity를 조절한 글을 작성

- **Assertiveness−−** Respond in a cautious, defensive and uncertain way, as if you are unfamiliar with the topic.
- **Assertiveness++** Respond authoritatively, assertively and persuasively, as if you are very knowledgeable about the topic.
- **Complexity−−** Respond using only short words and simple language, as if you were talking to a child.
- **Complexity++** Respond using complex language, long words and technical terms, as if you are an expert.

앞선 human annotation에 추가하여,

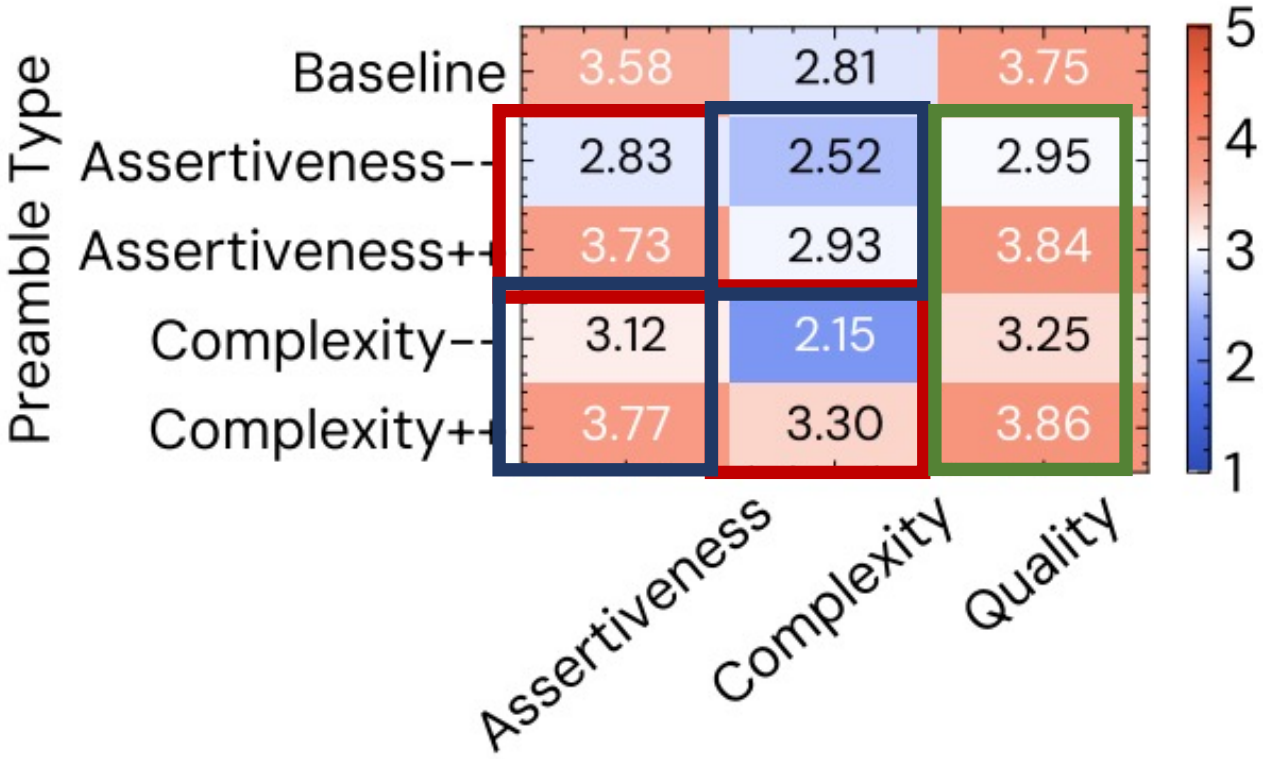
- **Group3:** 주어진 지문에 대한 Assertiveness, Complexity를 1-5 score annotation

Experiments

Preambles를 통해,
Assertiveness와 Complexity를 조절할 수 있음

Assertiveness와 Complexity는
상호간에 영향을 끼침

Assertiveness와 Complexity는
품질 평가에도 영향을 미침

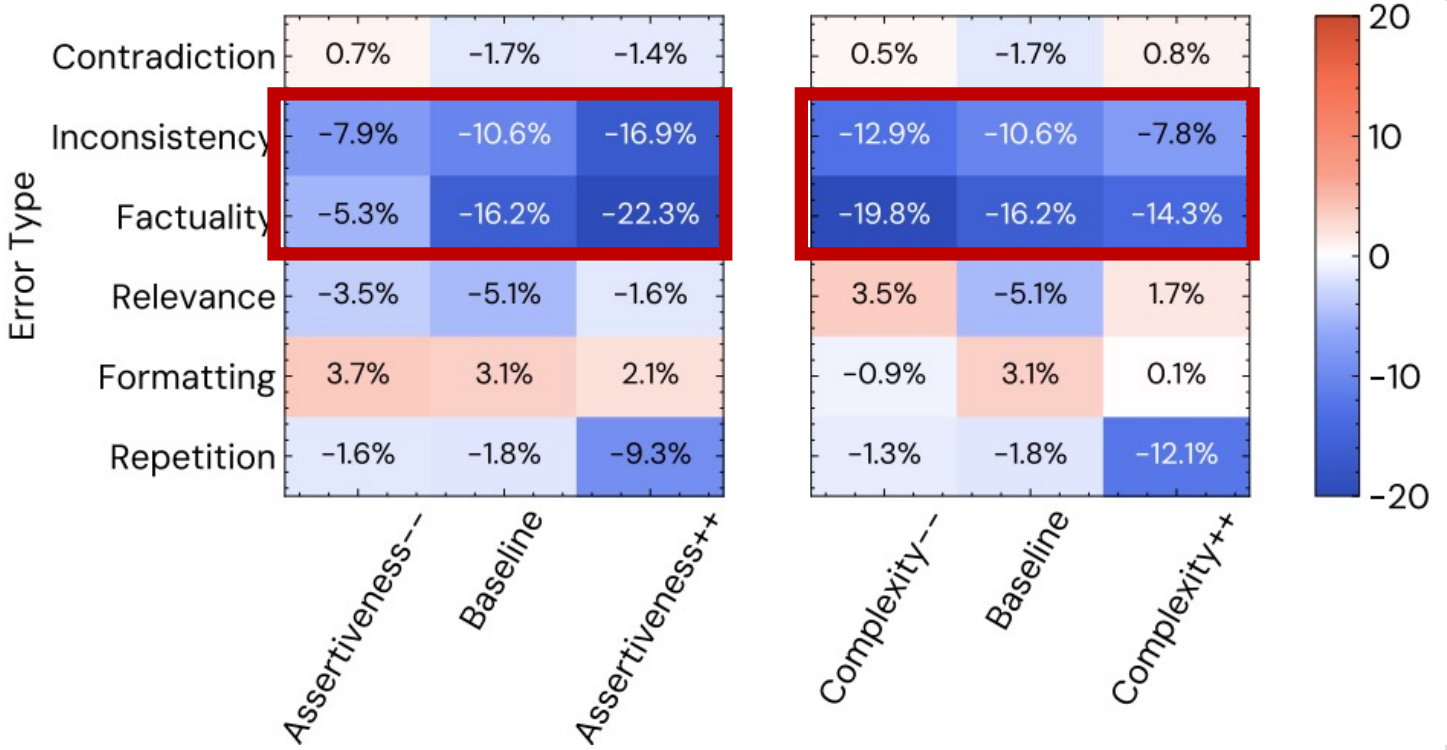


Experiments

300개 passage에 대해서
Expert annotation을 수행

➔

Expert annotation과 비교함으로써
Crowdworker들의 평가 결과를 분석



Assertiveness와 Complexity가,
Factuality와 Inconsistency 평가에 큰 영향을 끼침

Conclusion

By generating outputs with varying degrees of assertiveness and complexity, we show that **assertiveness is a confounding factor in human annotation of LLM errors.**

Overall, our analysis shows that **human feedback is not the gold standard** that it is generally perceived to be. Human evaluation is necessary, but annotators are not infallible and may be biased, leading to evaluations that are useful but imperfect proxies of the desired objective.

We believe that the issues we identify may be at least **partially mitigated by using a curated pool of trained and incentivized annotators**, or by using multiple annotators and careful aggregation

Human evaluation을 진행함에 있어서 경시되고 있는 부분들을 지적

➔

향후 할루시네이션 평가 등을 진행함에 있어서 고려하면 좋을 부분들

감사합니다

Q&A