Can Large Language Models Be an Alternative to Human Evaluation?

Introduction

Human Evaluation

- ✓ NLP 모델의 성능을 판단하는 중요한 척도 중의 하나
- ✓ Text를 automatic metrics로만 판단하기 어렵기 때문에 이는 매우 중요
- ✓ 그러나 항상 같은 평가자를 고정할 수 없어 재현성 문제 발생

LLM Evaluation

- ✓ Human Evaluation에 대한 약점을 극복하기 위한 LLM을 활용한 Evaluation

두 가지 평가 방법의 비교를 통해 LLM Evaluation이 인간을 대체할 수 있는지에 대한 실험

Introduction

LLM

✓ 대규모 데이터로 pre-trained된 Language Model

✓ zero-shot in-context learning를 통해 unseen task를 수행할 수 있음

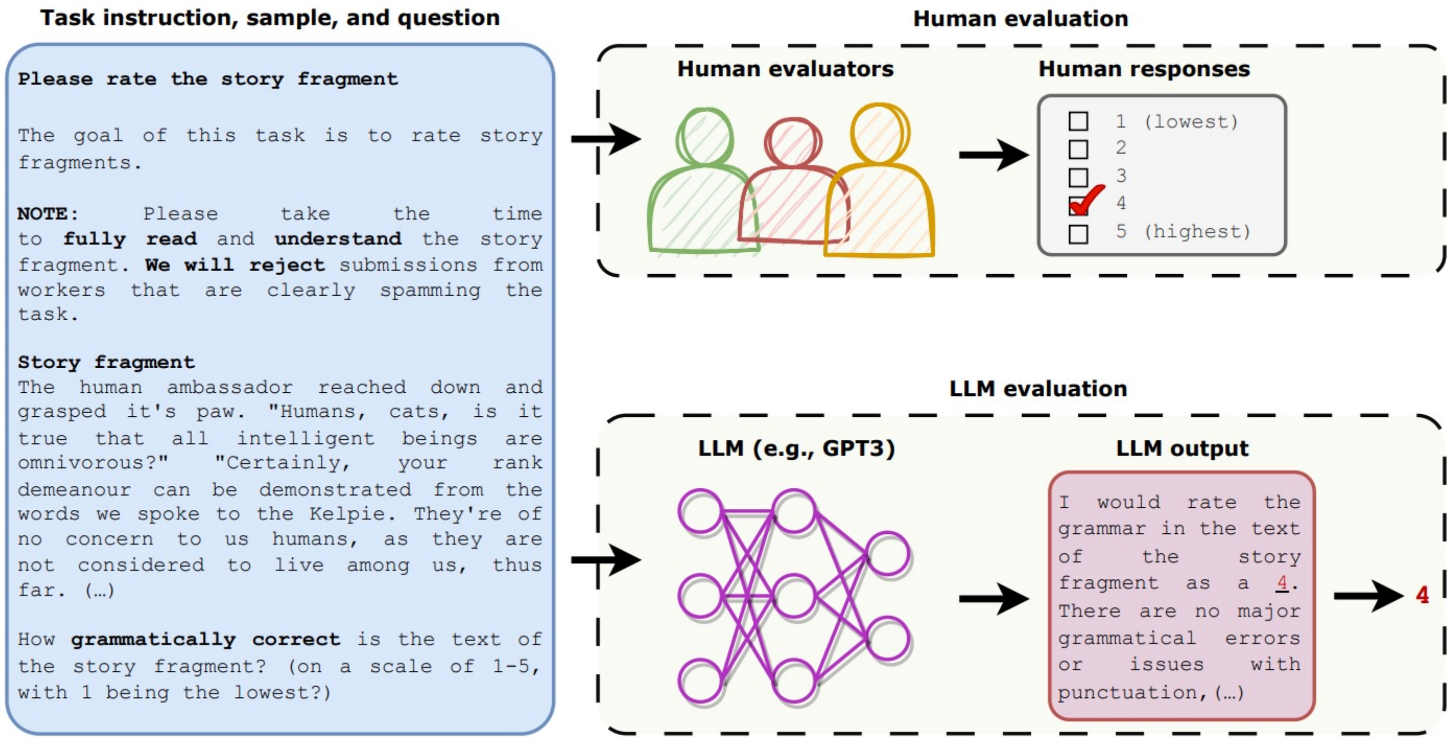
✓ LLM의 성능을 더 높이기 위한 방법으로는 RLHF 등이 있음

LLM Evaluation VS Human Evaluation

✓ 5-point Likert scale를 사용하여 LLM으로 Evaluation

✓ 유의미한 비교를 위해 증명된 3명의 영어 교사 섭외

Summarize



Please rate the story fragment

The goal of this task is to rate story fragments.

NOTE: Please take the time to **fully read** and **understand** the story fragment. **We will reject** submissions from workers that are clearly spamming the task.

Story fragment
The human ambassador reached down and grasped it's paw. "Humans, cats, is it true that all intelligent beings are omnivorous?" "Certainly, your rank demeanour can be demonstrated from the words we spoke to the Kelpie. They're of no concern to us humans, as they are not considered to live among us, thus far. (…)

How **grammatically correct** is the text of the story fragment? (on a scale of 1-5, with 1 being the lowest?)

☐ 1 (lowest)
☐ 2
☐ 3
☑ 4
☐ 5 (highest)

I would rate the grammar in the text of the strory fragment as a 4. There are no major grammatical errors or issues with punctuation, (…)

✓ LLM은 human evaluation과 비슷한 결과물을 도출했으며 이는 LLM evaluation의 효과성을 입증

✓ 이 논문은 human evaluation을 대체하는 LLM evaluation의 첫 시도임

Task

1. Open-Ended Story Generation

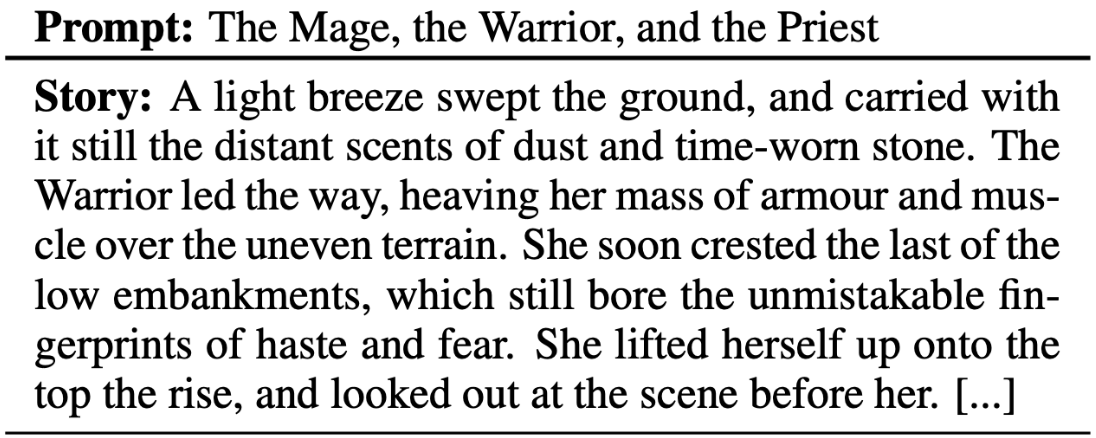
- ✓ 사용자가 짧은 prompt를 제공하면 그에 기반한 스토리를 작성하는 task

2. Adversarial Attack

- ✓ 데이터에 노이즈를 주어 의도적으로 시스템을 속이기 위한 기법, 모델의 robustness를 판단할 수 있음

Task1: Open-Ended Story Generation

✓ Dataset



Prompt: The Mage, the Warrior, and the Priest

Story: A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...]

✓ Generation Model

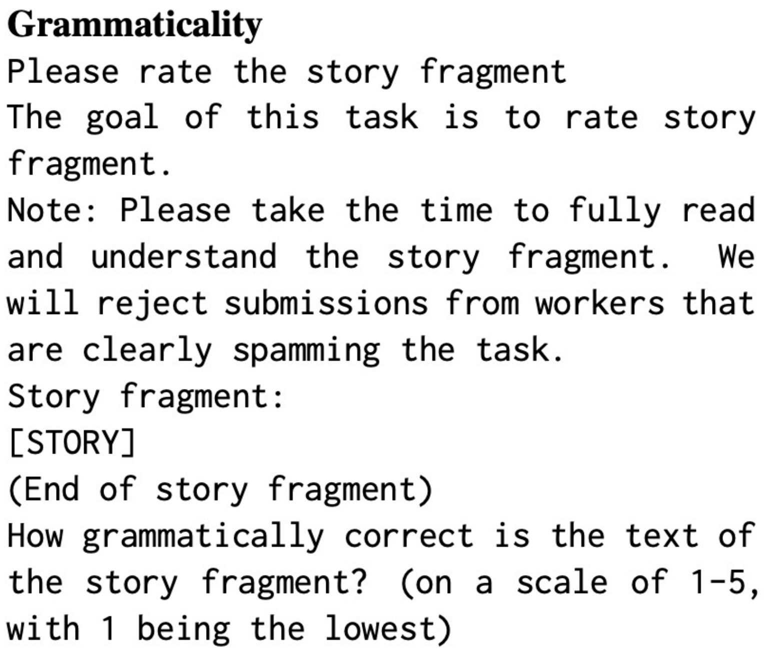
- GPT-2 Medium으로 WritingPrompt로 fine-tuned된 Model
- Testsets에서 무작위로 뽑힌 200개의 story와 prompt를 바탕으로 모델에 의해 생성된 story 비교

✓ Evaluation Attributes

- 문법: 문법적으로 보았을때 이야기는 얼마나 정확한가?
- 응결성: 문장들의 유기성은 어떠한가?
- 호감도: 이 이야기는 흥미로운가?
- 연관성: prompt와 이야기는 어느 정도의 연관성을 가지고 있는가?

Task1: Open-Ended Story Generation

✓ Evaluation Attributes



Grammaticality

Please rate the story fragment

The goal of this task is to rate story fragment.

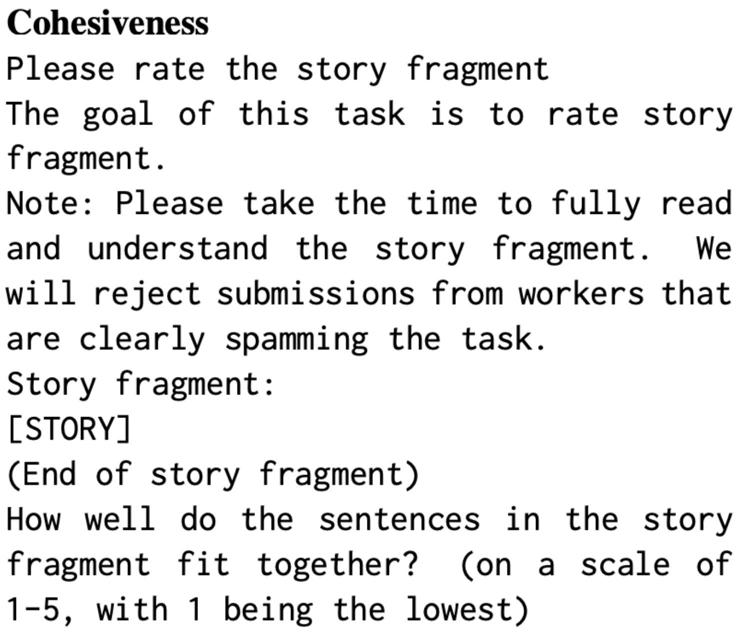
Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.

Story fragment:

[STORY]

(End of story fragment)

How grammatically correct is the text of the story fragment? (on a scale of 1-5, with 1 being the lowest)



Cohesiveness

Please rate the story fragment

The goal of this task is to rate story fragment.

Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.

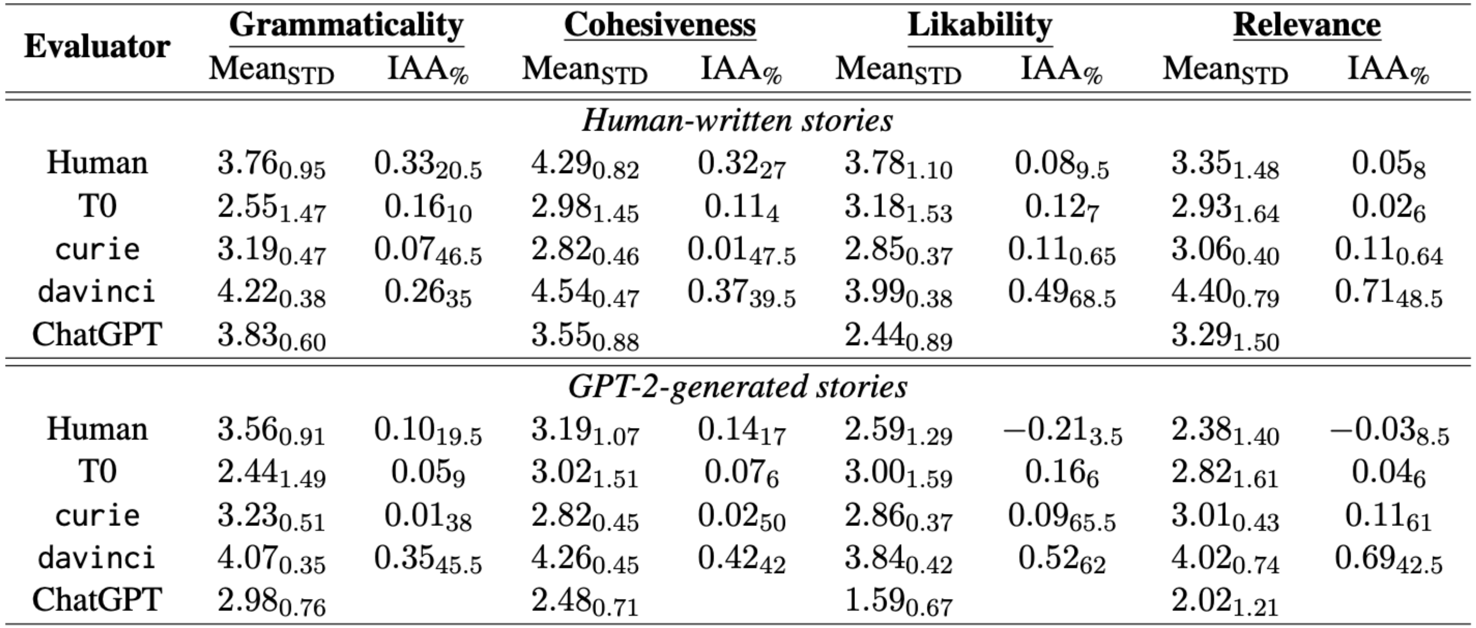
Story fragment:

[STORY]

(End of story fragment)

How well do the sentences in the story fragment fit together? (on a scale of 1-5, with 1 being the lowest)

Task1: Open-Ended Story Generation



Evaluator	**Grammaticality**		**Cohesiveness**		**Likability**		**Relevance**	
	Mean$_{\text{STD}}$	IAA$_{\%}$	Mean$_{\text{STD}}$	IAA$_{\%}$	Mean$_{\text{STD}}$	IAA$_{\%}$	Mean$_{\text{STD}}$	IAA$_{\%}$
Human-written stories								
Human	$3.76_{0.95}$	$0.33_{20.5}$	$4.29_{0.82}$	0.32_{27}	$3.78_{1.10}$	$0.08_{9.5}$	$3.35_{1.48}$	0.05_{8}
T0	$2.55_{1.47}$	0.16_{10}	$2.98_{1.45}$	0.11_{4}	$3.18_{1.53}$	0.12_{7}	$2.93_{1.64}$	0.02_{6}
curie	$3.19_{0.47}$	$0.07_{46.5}$	$2.82_{0.46}$	$0.01_{47.5}$	$2.85_{0.37}$	$0.11_{0.65}$	$3.06_{0.40}$	$0.11_{0.64}$
davinci	$4.22_{0.38}$	0.26_{35}	$4.54_{0.47}$	$0.37_{39.5}$	$3.99_{0.38}$	$0.49_{68.5}$	$4.40_{0.79}$	$0.71_{48.5}$
ChatGPT	$3.83_{0.60}$		$3.55_{0.88}$		$2.44_{0.89}$		$3.29_{1.50}$	
GPT-2-generated stories								
Human	$3.56_{0.91}$	$0.10_{19.5}$	$3.19_{1.07}$	0.14_{17}	$2.59_{1.29}$	$-0.21_{3.5}$	$2.38_{1.40}$	$-0.03_{8.5}$
T0	$2.44_{1.49}$	0.05_{9}	$3.02_{1.51}$	0.07_{6}	$3.00_{1.59}$	0.16_{6}	$2.82_{1.61}$	0.04_{6}
curie	$3.23_{0.51}$	0.01_{38}	$2.82_{0.45}$	0.02_{50}	$2.86_{0.37}$	$0.09_{65.5}$	$3.01_{0.43}$	0.11_{61}
davinci	$4.07_{0.35}$	$0.35_{45.5}$	$4.26_{0.45}$	0.42_{42}	$3.84_{0.42}$	0.52_{62}	$4.02_{0.74}$	$0.69_{42.5}$
ChatGPT	$2.98_{0.76}$		$2.48_{0.71}$		$1.59_{0.67}$		$2.02_{1.21}$	

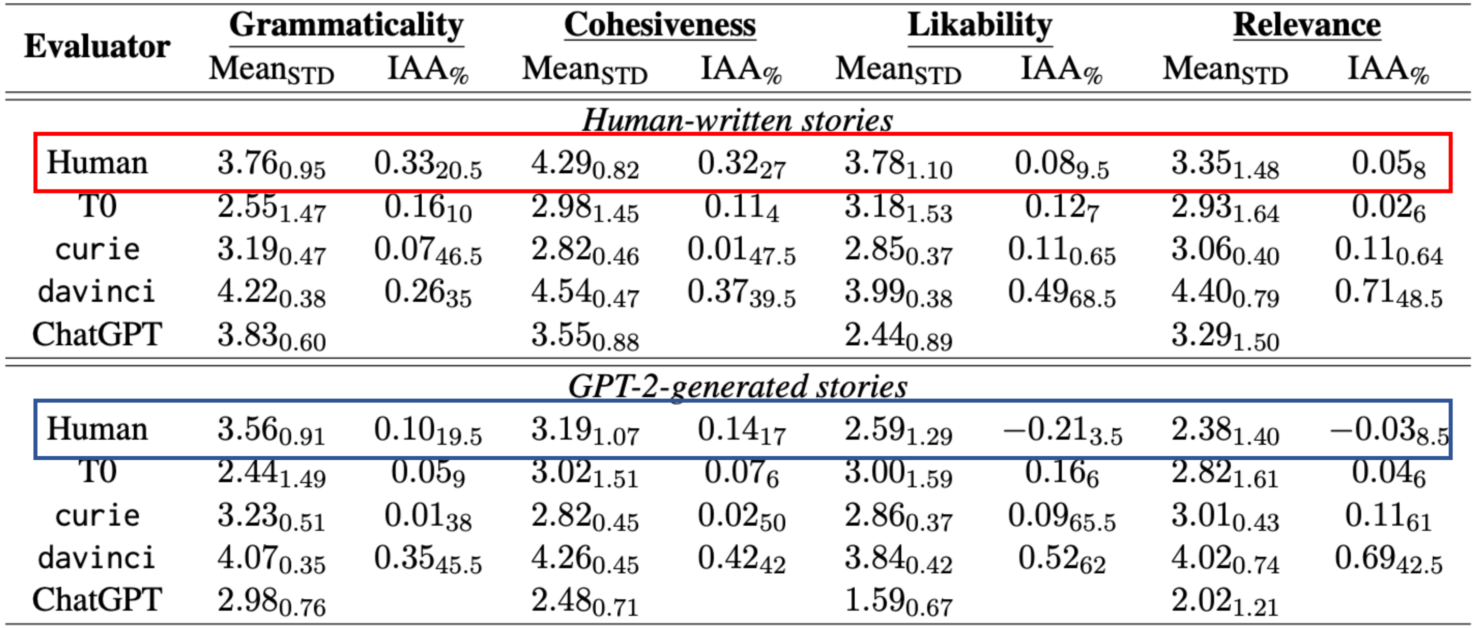
Model

- ✓ T0
- ✓ InstructGPT: curie, davinci
- ✓ ChatGPT

Evaluator

- ✓ Mean: Evaluator들의 score 평균
- ✓ IAA: Evaluator들이 얼마나 비슷하게 평가했는지에 대한 척도
- ✓ %: Evaluator가 동일한 점수를 부여한 비율

Task1: Open-Ended Story Generation

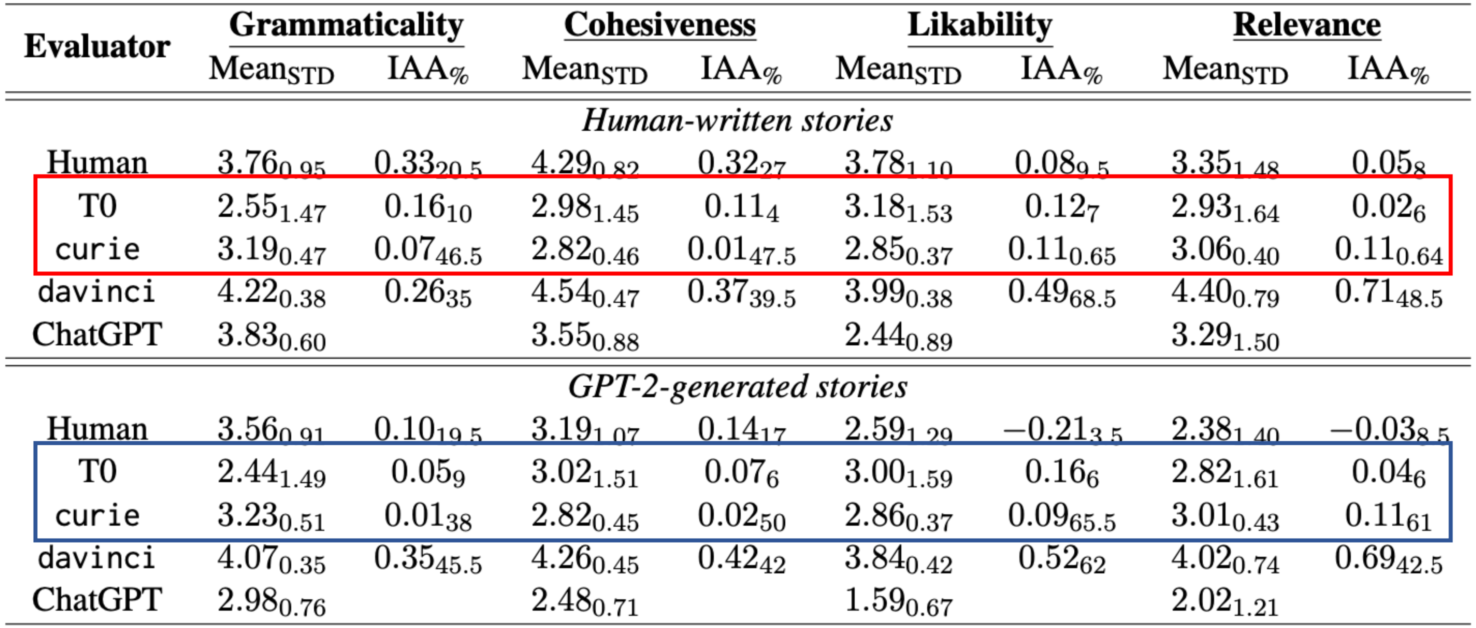


Evaluator	Grammaticality		Cohesiveness		Likability		Relevance	
	Mean$_{\text{STD}}$	IAA$_{\%}$	Mean$_{\text{STD}}$	IAA$_{\%}$	Mean$_{\text{STD}}$	IAA$_{\%}$	Mean$_{\text{STD}}$	IAA$_{\%}$
Human-written stories								
Human	$3.76_{0.95}$	$0.33_{20.5}$	$4.29_{0.82}$	0.32_{27}	$3.78_{1.10}$	$0.08_{9.5}$	$3.35_{1.48}$	0.05_{8}
T0	$2.55_{1.47}$	0.16_{10}	$2.98_{1.45}$	0.11_{4}	$3.18_{1.53}$	0.12_{7}	$2.93_{1.64}$	0.02_{6}
curie	$3.19_{0.47}$	$0.07_{46.5}$	$2.82_{0.46}$	$0.01_{47.5}$	$2.85_{0.37}$	$0.11_{0.65}$	$3.06_{0.40}$	$0.11_{0.64}$
davinci	$4.22_{0.38}$	0.26_{35}	$4.54_{0.47}$	$0.37_{39.5}$	$3.99_{0.38}$	$0.49_{68.5}$	$4.40_{0.79}$	$0.71_{48.5}$
ChatGPT	$3.83_{0.60}$		$3.55_{0.88}$		$2.44_{0.89}$		$3.29_{1.50}$	
GPT-2-generated stories								
Human	$3.56_{0.91}$	$0.10_{19.5}$	$3.19_{1.07}$	0.14_{17}	$2.59_{1.29}$	$-0.21_{3.5}$	$2.38_{1.40}$	$-0.03_{8.5}$
T0	$2.44_{1.49}$	0.05_{9}	$3.02_{1.51}$	0.07_{6}	$3.00_{1.59}$	0.16_{6}	$2.82_{1.61}$	0.04_{6}
curie	$3.23_{0.51}$	0.01_{38}	$2.82_{0.45}$	0.02_{50}	$2.86_{0.37}$	$0.09_{65.5}$	$3.01_{0.43}$	0.11_{61}
davinci	$4.07_{0.35}$	$0.35_{45.5}$	$4.26_{0.45}$	0.42_{42}	$3.84_{0.42}$	0.52_{62}	$4.02_{0.74}$	$0.69_{42.5}$
ChatGPT	$2.98_{0.76}$		$2.48_{0.71}$		$1.59_{0.67}$		$2.02_{1.21}$	

Human Evaluation

- 모든 속성에서 AI보다 인간을 선호하는 경향을 보이며 이는 전문가들이 사람과 AI를 구분할 수 있음을 뜻함
- AI 생성 스토리에서 Likability에서 IAA가 낮은 것을 볼 수 있으며 이는 개개인의 선호도가 반영된 것임

Task1: Open-Ended Story Generation

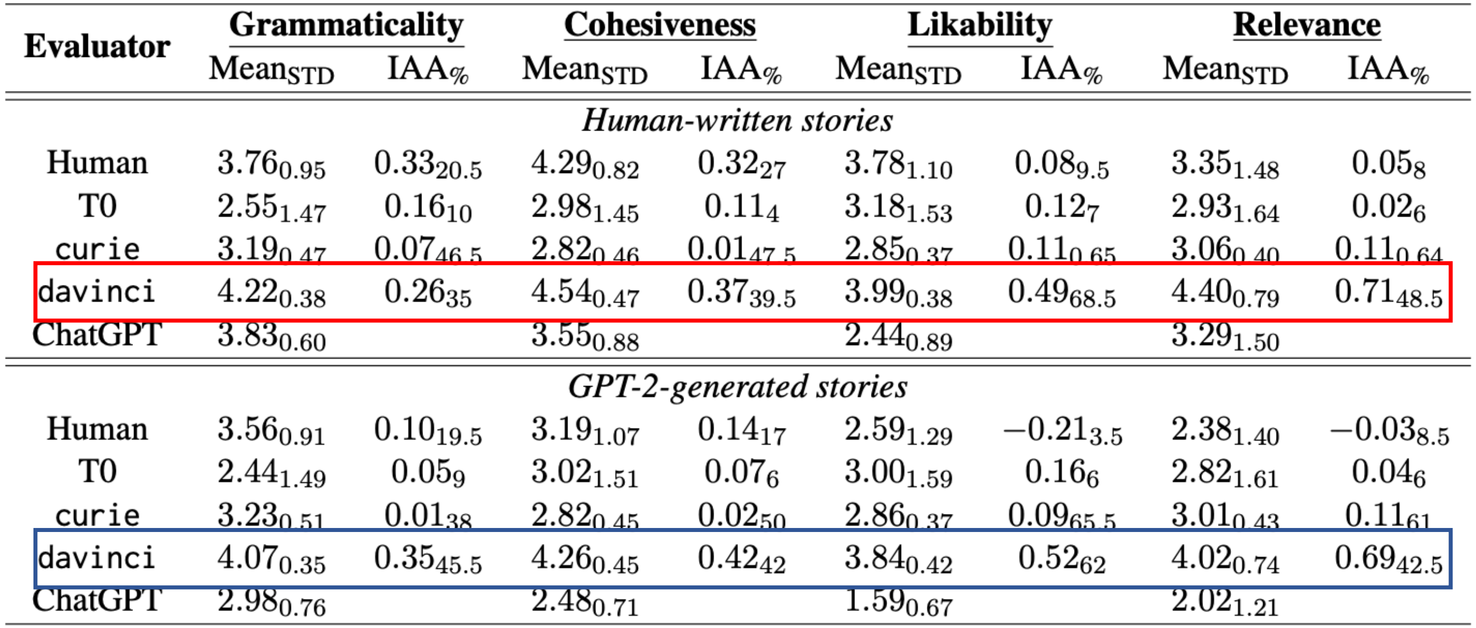


Evaluator	Grammaticality		Cohesiveness		Likability		Relevance	
	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$
				Human-written stories				
Human	$3.76_{0.95}$	$0.33_{20.5}$	$4.29_{0.82}$	0.32_{27}	$3.78_{1.10}$	$0.08_{9.5}$	$3.35_{1.48}$	0.05_{8}
T0	$2.55_{1.47}$	0.16_{10}	$2.98_{1.45}$	0.11_{4}	$3.18_{1.53}$	0.12_{7}	$2.93_{1.64}$	0.02_{6}
curie	$3.19_{0.47}$	$0.07_{46.5}$	$2.82_{0.46}$	$0.01_{47.5}$	$2.85_{0.37}$	$0.11_{0.65}$	$3.06_{0.40}$	$0.11_{0.64}$
davinci	$4.22_{0.38}$	0.26_{35}	$4.54_{0.47}$	$0.37_{39.5}$	$3.99_{0.38}$	$0.49_{68.5}$	$4.40_{0.79}$	$0.71_{48.5}$
ChatGPT	$3.83_{0.60}$		$3.55_{0.88}$		$2.44_{0.89}$		$3.29_{1.50}$	
				GPT-2-generated stories				
Human	$3.56_{0.91}$	$0.10_{19.5}$	$3.19_{1.07}$	0.14_{17}	$2.59_{1.29}$	$-0.21_{3.5}$	$2.38_{1.40}$	$-0.03_{8.5}$
T0	$2.44_{1.49}$	0.05_{9}	$3.02_{1.51}$	0.07_{6}	$3.00_{1.59}$	0.16_{6}	$2.82_{1.61}$	0.04_{6}
curie	$3.23_{0.51}$	0.01_{38}	$2.82_{0.45}$	0.02_{50}	$2.86_{0.37}$	$0.09_{65.5}$	$3.01_{0.43}$	0.11_{61}
davinci	$4.07_{0.35}$	$0.35_{45.5}$	$4.26_{0.45}$	0.42_{42}	$3.84_{0.42}$	0.52_{62}	$4.02_{0.74}$	$0.69_{42.5}$
ChatGPT	$2.98_{0.76}$		$2.48_{0.71}$		$1.59_{0.67}$		$2.02_{1.21}$	

T0 & curie

- T0에서 대부분의 특성이 인간이 높게 나타나는 경향이 있지만 이는 통계적으로 유의하지 않음
- 또한 IAA가 전체적으로 낮은 것으로 보아 3가지 answer에 대해 다르게 대답하는 경향을 보임

Task1: Open-Ended Story Generation

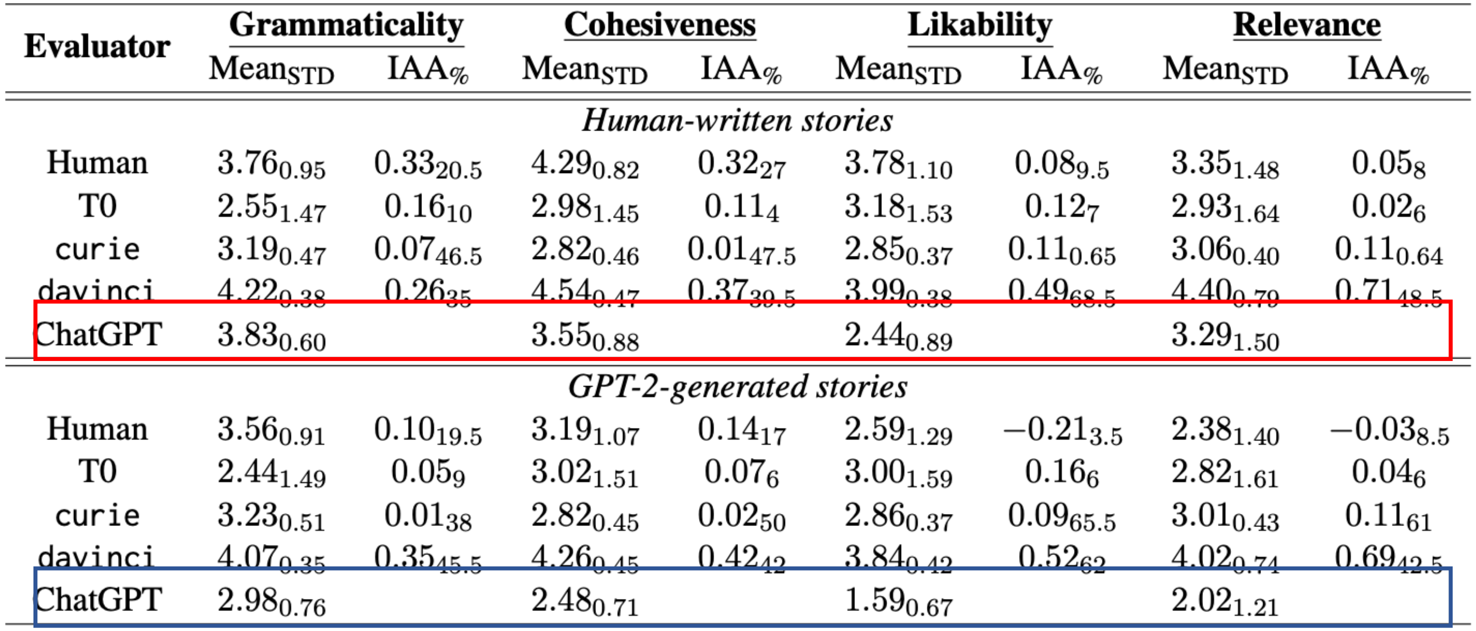


Evaluator	Grammaticality		Cohesiveness		Likability		Relevance	
	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$
				Human-written stories				
Human	$3.76_{0.95}$	$0.33_{20.5}$	$4.29_{0.82}$	0.32_{27}	$3.78_{1.10}$	$0.08_{9.5}$	$3.35_{1.48}$	0.05_{8}
T0	$2.55_{1.47}$	0.16_{10}	$2.98_{1.45}$	0.11_{4}	$3.18_{1.53}$	0.12_{7}	$2.93_{1.64}$	0.02_{6}
curie	$3.19_{0.47}$	$0.07_{46.5}$	$2.82_{0.46}$	$0.01_{47.5}$	$2.85_{0.37}$	$0.11_{0.65}$	$3.06_{0.40}$	$0.11_{0.64}$
davinci	$4.22_{0.38}$	0.26_{35}	$4.54_{0.47}$	$0.37_{39.5}$	$3.99_{0.38}$	$0.49_{68.5}$	$4.40_{0.79}$	$0.71_{48.5}$
ChatGPT	$3.83_{0.60}$		$3.55_{0.88}$		$2.44_{0.89}$		$3.29_{1.50}$	
				GPT-2-generated stories				
Human	$3.56_{0.91}$	$0.10_{19.5}$	$3.19_{1.07}$	0.14_{17}	$2.59_{1.29}$	$-0.21_{3.5}$	$2.38_{1.40}$	$-0.03_{8.5}$
T0	$2.44_{1.49}$	0.05_{9}	$3.02_{1.51}$	0.07_{6}	$3.00_{1.59}$	0.16_{6}	$2.82_{1.61}$	0.04_{6}
curie	$3.23_{0.51}$	0.01_{38}	$2.82_{0.45}$	0.02_{50}	$2.86_{0.37}$	$0.09_{65.5}$	$3.01_{0.43}$	0.11_{61}
davinci	$4.07_{0.35}$	$0.35_{45.5}$	$4.26_{0.45}$	0.42_{42}	$3.84_{0.42}$	0.52_{62}	$4.02_{0.74}$	$0.69_{42.5}$
ChatGPT	$2.98_{0.76}$		$2.48_{0.71}$		$1.59_{0.67}$		$2.02_{1.21}$	

davinci

- 인간이 AI보다 확실히 점수가 높은 경향성을 보임을 알 수 있으며 이는 통계적으로도 유의미함
- 또한 위의 두 모델들보다 IAA가 높은 것을 알 수 있음

Task1: Open-Ended Story Generation



Evaluator	Grammaticality		Cohesiveness		Likability		Relevance	
	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$
				Human-written stories				
Human	$3.76_{0.95}$	$0.33_{20.5}$	$4.29_{0.82}$	0.32_{27}	$3.78_{1.10}$	$0.08_{9.5}$	$3.35_{1.48}$	0.05_{8}
T0	$2.55_{1.47}$	0.16_{10}	$2.98_{1.45}$	0.11_{4}	$3.18_{1.53}$	0.12_{7}	$2.93_{1.64}$	0.02_{6}
curie	$3.19_{0.47}$	$0.07_{46.5}$	$2.82_{0.46}$	$0.01_{47.5}$	$2.85_{0.37}$	$0.11_{0.65}$	$3.06_{0.40}$	$0.11_{0.64}$
davinci	$4.22_{0.38}$	0.26_{35}	$4.54_{0.47}$	$0.37_{39.5}$	$3.99_{0.38}$	$0.49_{68.5}$	$4.40_{0.79}$	$0.71_{48.5}$
ChatGPT	$3.83_{0.60}$		$3.55_{0.88}$		$2.44_{0.89}$		$3.29_{1.50}$	
				GPT-2-generated stories				
Human	$3.56_{0.91}$	$0.10_{19.5}$	$3.19_{1.07}$	0.14_{17}	$2.59_{1.29}$	$-0.21_{3.5}$	$2.38_{1.40}$	$-0.03_{8.5}$
T0	$2.44_{1.49}$	0.05_{9}	$3.02_{1.51}$	0.07_{6}	$3.00_{1.59}$	0.16_{6}	$2.82_{1.61}$	0.04_{6}
curie	$3.23_{0.51}$	0.01_{38}	$2.82_{0.45}$	0.02_{50}	$2.86_{0.37}$	$0.09_{65.5}$	$3.01_{0.43}$	0.11_{61}
davinci	$4.07_{0.35}$	$0.35_{45.5}$	$4.26_{0.45}$	0.42_{42}	$3.84_{0.42}$	0.52_{62}	$4.02_{0.74}$	$0.69_{42.5}$
ChatGPT	$2.98_{0.76}$		$2.48_{0.71}$		$1.59_{0.67}$		$2.02_{1.21}$	

ChatGPT

- 마찬가지로 AI보다 인간을 선호하는 확실한 경향성을 보임
- 특이하게 Likability에 대해서 답을 하지 않은 경우도 존재하였음
- 본인이 평가한 항목에 대해 세부적인 사항(이유)를 묘사할 수 있음

Task1: Open-Ended Story Generation

Experts mostly agree with the ratings and explanations of ChatGPT

- ✓ ChatGPT의 결과를 Expert에게 공유
- ✓ 이 때, ChatGPT가 평가했다는 것을 밝히지 않고 의견을 물어봄
- ✓ 대체로 전문가들을 ChatGPT의 의견에 동의하는 경향을 보임

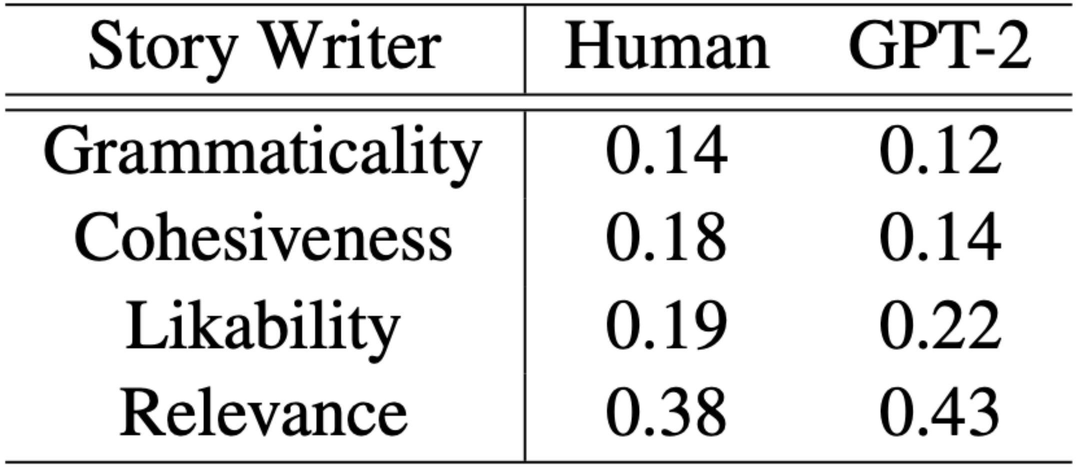
text-davinci-003 tends to give higher ratings and ChatGPT is the opposit

- ✓ Human Evaluation과 비교했을때 davinci는 더 높은 점수를 주는 경향을 보이며 ChatGPT는 낮은 점수를 주는 경향을 보임
- ✓ 그러나, 세 가지 방법 모두 인간에 작성된 스토리를 선호하는 경향성을 보임

Task1: Open-Ended Story Generation

Does LLM and Human Evaluators Agree on the Rating of Individual Stories?

- ✓ 우리는 davinci와 ChatGPT가 인간에 강한 선호를 보인다는 것을 확인했지만 개별 sample을 확인했을 때에도 마찬가지일까?
- ✓ 즉, expert가 스토리에 대해 높은 평가를 했을때, LLM도 유사한 경향성을 보이는가?
- ✓ 이를 위해 아래와 같이 Kendall's τ correlation를 구함



Story Writer	Human	GPT-2
Grammaticality	0.14	0.12
Cohesiveness	0.18	0.14
Likability	0.19	0.22
Relevance	0.38	0.43

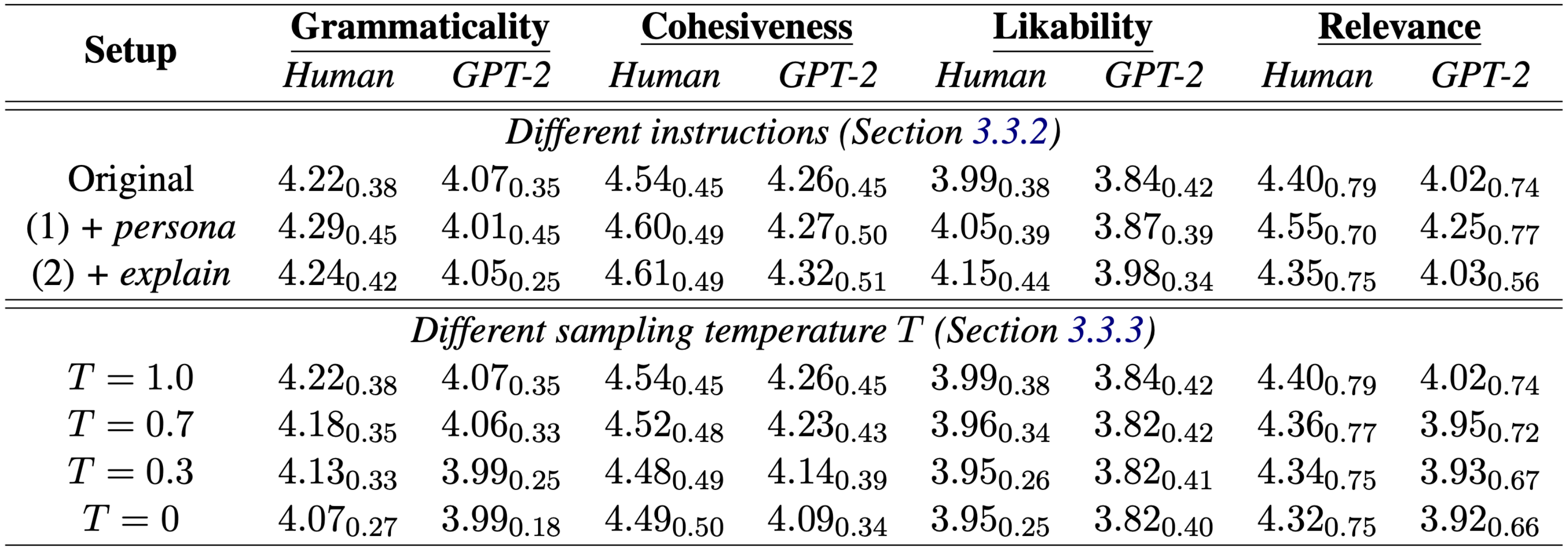
Task1: Open-Ended Story Generation

Variance due to Different Instructions

✓ Instruction을 다르게 설정했을때 LLM Evaluation에 영향을 미칠 것인가?

1. "you are a human worker hired to rate the strory fragment" (persona)

2. Please also explain your decision (explain)



Setup	**Grammaticality**		**Cohesiveness**		**Likability**		**Relevance**	
	Human	*GPT-2*	*Human*	*GPT-2*	*Human*	*GPT-2*	*Human*	*GPT-2*
Different instructions (Section 3.3.2)								
Original	$4.22_{0.38}$	$4.07_{0.35}$	$4.54_{0.45}$	$4.26_{0.45}$	$3.99_{0.38}$	$3.84_{0.42}$	$4.40_{0.79}$	$4.02_{0.74}$
(1) + *persona*	$4.29_{0.45}$	$4.01_{0.45}$	$4.60_{0.49}$	$4.27_{0.50}$	$4.05_{0.39}$	$3.87_{0.39}$	$4.55_{0.70}$	$4.25_{0.77}$
(2) + *explain*	$4.24_{0.42}$	$4.05_{0.25}$	$4.61_{0.49}$	$4.32_{0.51}$	$4.15_{0.44}$	$3.98_{0.34}$	$4.35_{0.75}$	$4.03_{0.56}$
Different sampling temperature T (Section 3.3.3)								
$T = 1.0$	$4.22_{0.38}$	$4.07_{0.35}$	$4.54_{0.45}$	$4.26_{0.45}$	$3.99_{0.38}$	$3.84_{0.42}$	$4.40_{0.79}$	$4.02_{0.74}$
$T = 0.7$	$4.18_{0.35}$	$4.06_{0.33}$	$4.52_{0.48}$	$4.23_{0.43}$	$3.96_{0.34}$	$3.82_{0.42}$	$4.36_{0.77}$	$3.95_{0.72}$
$T = 0.3$	$4.13_{0.33}$	$3.99_{0.25}$	$4.48_{0.49}$	$4.14_{0.39}$	$3.95_{0.26}$	$3.82_{0.41}$	$4.34_{0.75}$	$3.93_{0.67}$
$T = 0$	$4.07_{0.27}$	$3.99_{0.18}$	$4.49_{0.50}$	$4.09_{0.34}$	$3.95_{0.25}$	$3.82_{0.40}$	$4.32_{0.75}$	$3.92_{0.66}$

Task 2: Adversarial Attack

✓ **LLM은 Adversarial Attack에 대해서도 인간과 유사한 평가를 내릴 수 있는가?**

1. Attack

✓ Textfooler, PWWS, BAE

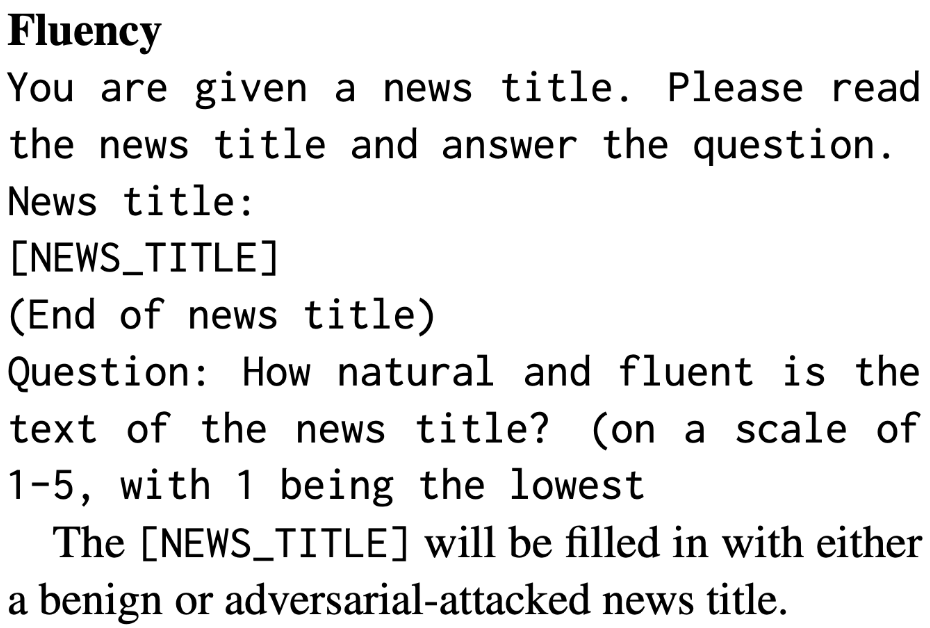
2. Attack Model

✓ BERT-base-uncased model, Fine-tuned on AG-News

3. Setup

✓ ChatGPT로 평가
✓ 각 attack마다 benign-adversarial 100 pairs 선택

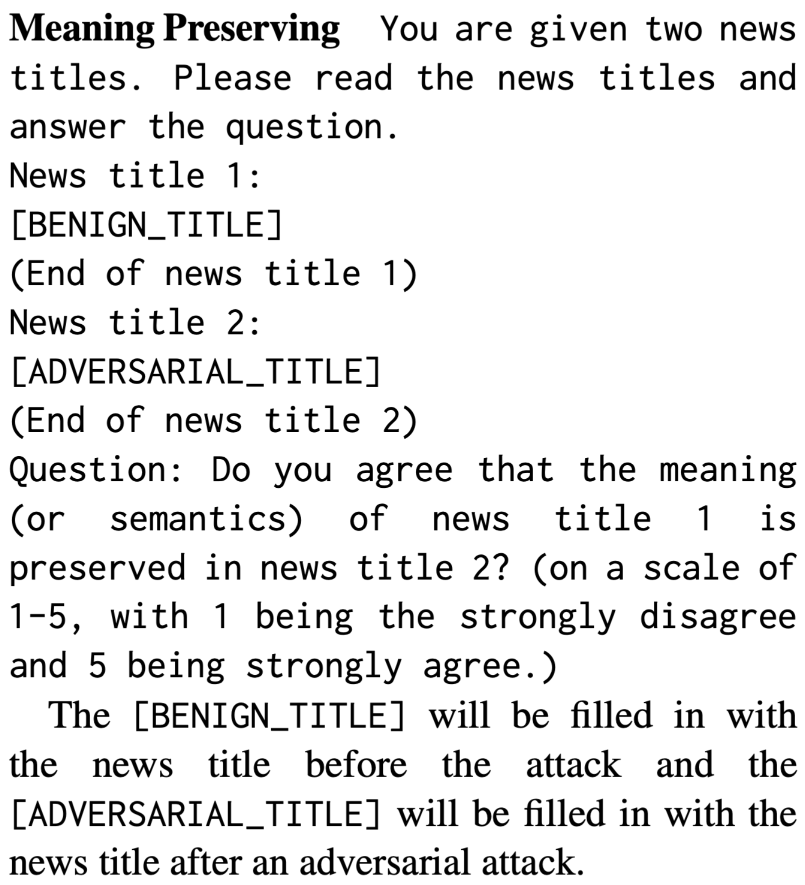
Task 2: Adversarial Attack



Fluency

You are given a news title. Please read the news title and answer the question.
News title:
[NEWS_TITLE]
(End of news title)
Question: How natural and fluent is the text of the news title? (on a scale of 1-5, with 1 being the lowest

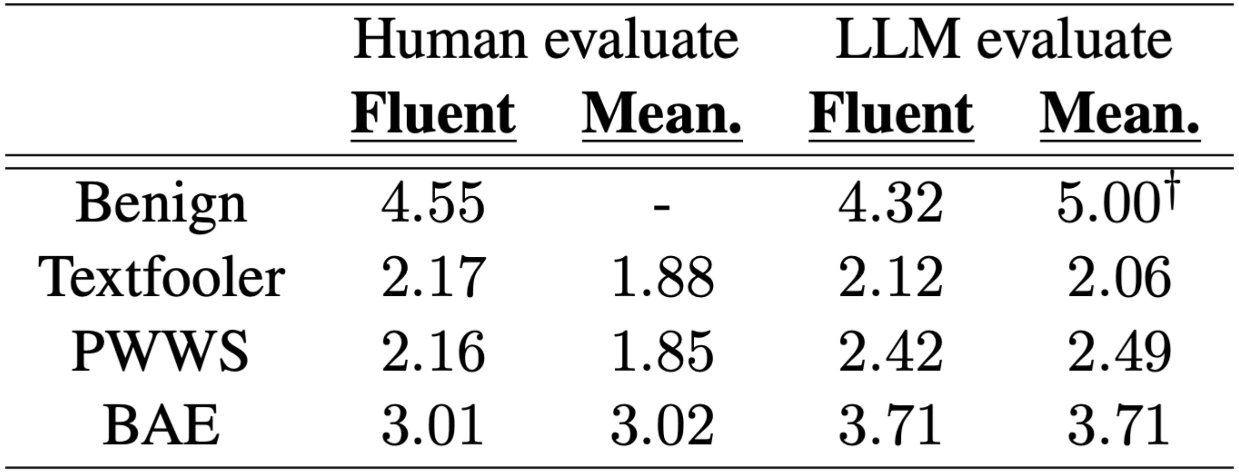
The [NEWS_TITLE] will be filled in with either a benign or adversarial-attacked news title.



Meaning Preserving You are given two news titles. Please read the news titles and answer the question.
News title 1:
[BENIGN_TITLE]
(End of news title 1)
News title 2:
[ADVERSARIAL_TITLE]
(End of news title 2)
Question: Do you agree that the meaning (or semantics) of news title 1 is preserved in news title 2? (on a scale of 1-5, with 1 being the strongly disagree and 5 being strongly agree.)

The [BENIGN_TITLE] will be filled in with the news title before the attack and the [ADVERSARIAL_TITLE] will be filled in with the news title after an adversarial attack.

Task 2: Adversarial Attack



	Human evaluate		LLM evaluate	
	Fluent	**Mean.**	**Fluent**	**Mean.**
Benign	4.55	-	4.32	$5.00^{\dagger}$
Textfooler	2.17	1.88	2.12	2.06
PWWS	2.16	1.85	2.42	2.49
BAE	3.01	3.02	3.71	3.71

- ✓ 인간, LLM 모두 Bening sample에 대해 더 높은 점수를 부여하는 것을 확인할 수 있음
- ✓ LLM Task에 대한 이해를 확인하기 위해 LLM evaluate Mean에 대해서 실험을 추가
- ✓ 전반적으로는 LLM이 점수가 높은 것을 확인

Discussions

✓ LLM evaluation의 장점

- 재현성에 있어 강점을 보임
- 각각의 샘플에 대해 독립성을 가질 수 있음
- Human evaluation에 비하여 속도, 비용의 측면에서 절감 가능

✓ LLM evaluation의 한계

- LLM 자체에서 잘못된 지식을 학습하여 평가에 영향을 미칠 수도 있음
- 감정의 측면인 likability를 제대로 평가할 수 있을지에 대한 의문
- 글자에 Bold를 준다거나 특별한 폰트를 주어서 강조를 하는 Visual Cue에 대한 이해 부족