# Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee

Jeff Da, Rowan Zellers, Robert Mankoff, Yejin Choi

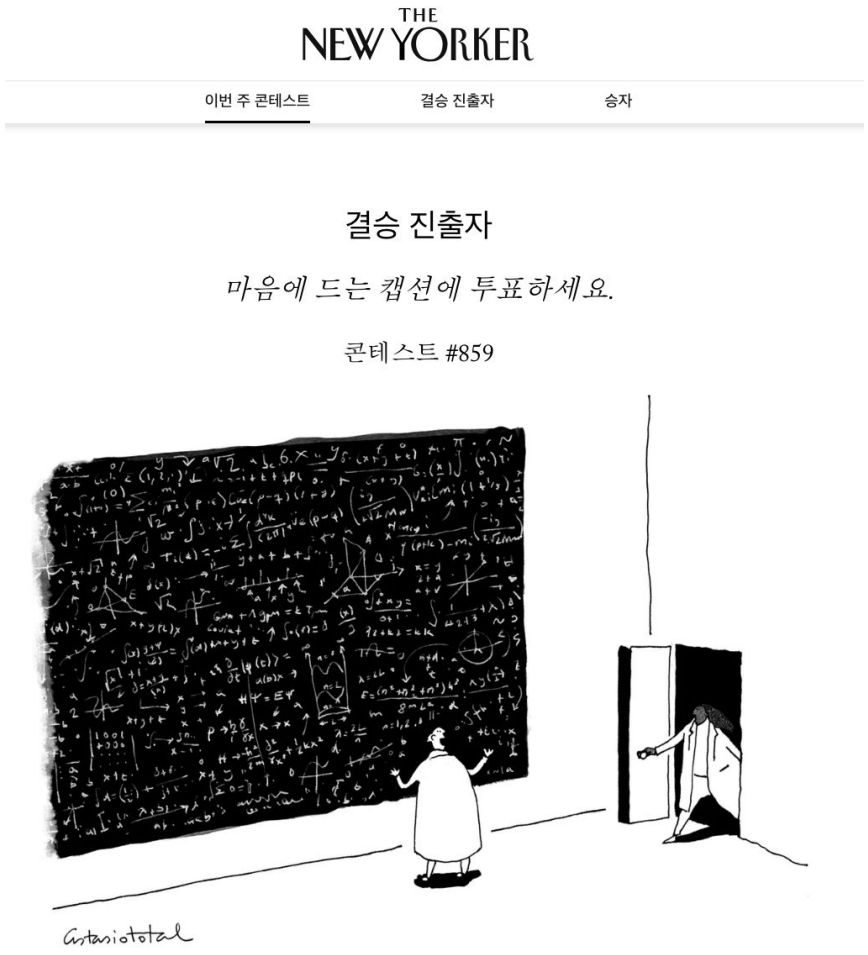
ACL 2023 Best Paper

발제자: 정현우

23-08-10

# Introduction : The New Yorker Contest

# Introduction : The New Yorker Contest




- *The New Yorker* 의 편집진 구성원 (이하 "심사위원")이 각 콘테스트에서 3명의 결선 진출자를 선정합니다.

- 각 콘테스트의 우승자는 캡션이 대중으로부터 가장 많은 유효 투표("투표")를 받고 모든 규칙을 충족하는 사람("자격을 갖춘 우승자")이 됩니다.

# Introduction : The New Yorker Contest

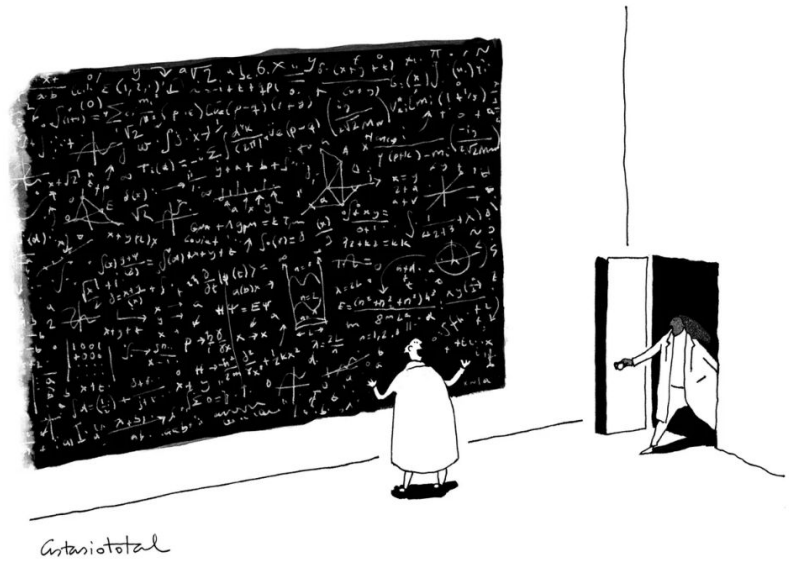
THE NEW YORKER

이번 주 콘테스트 | 결승 진출자 | 승자

## 결승 진출자

*마음에 드는 캡션에 투표하세요.*

콘테스트 #859



*"나는 증명할 것이 남아 있지 않습니다."*

Colin Mills, 보스턴, 매사추세츠

*"저는 학생들을 겁주기 위해 이렇게 합니다. 사실 그게 무슨 뜻인지 전혀 모르겠습니다."*

Kyle Sasloe, 노워크, 코네티컷

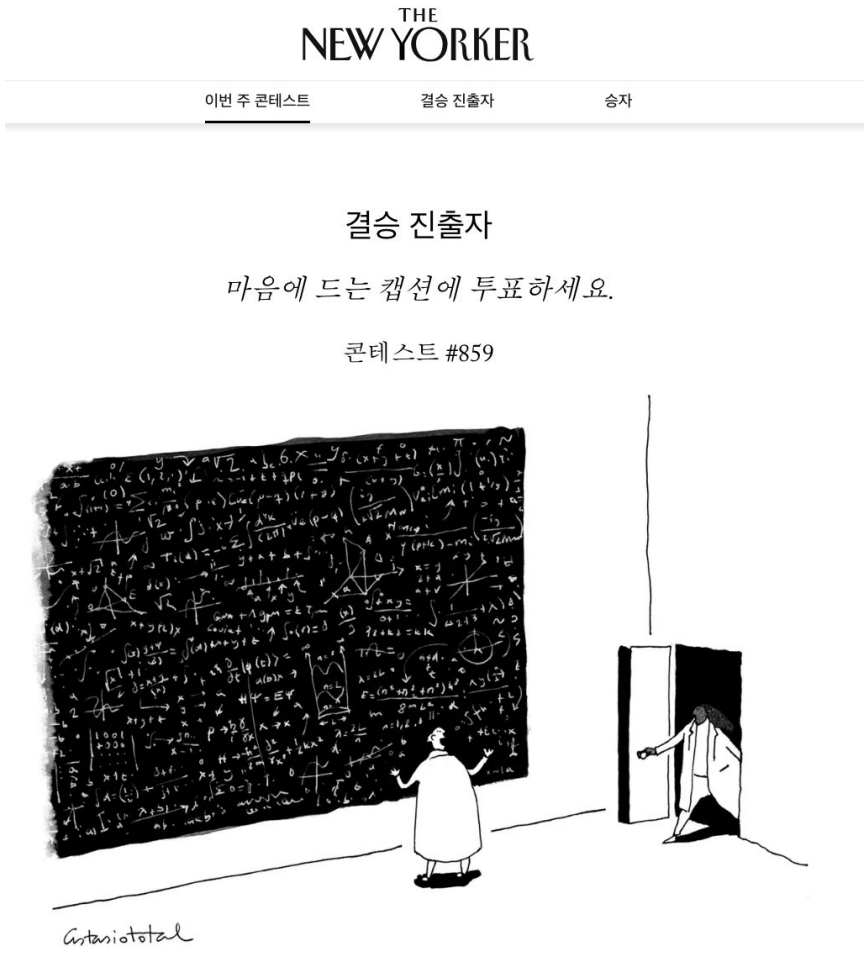
*"나는 항상 약간의 의심의 여지를 남겨둔다."*

Carol Lasky, 보스턴, 매사추세츠

투표하려면 로그인하세요

응모는 2023년 8월 13일 오후 11시 59분까지 접수되어야 합니다 . 공식 규칙을 보려면 여기를 클릭하십시오.

# Introduction : Three Tasks




- Task 1: Match the Caption + Cartoon
  - 이미지와 맞는 Caption 찾기
  - 후보작도 괜찮음.
- Task 2 : Rank the Finalist
  - 우승자에게 더 높은 점수를 주어야 함.
- Task 3: Explanation Generation
  - 사람과 비교해서 어떤 것이 더 나은지 판별해야 함.

# Introduction : Matching 예시


A) I always figured hell would be less ironic.

*B) You both know Jane*

C) I'd better give it a little longer. It's a really tough case.

D) And then I thought 'Wow, my cat really is kind of sexy.'

E) We'll eventually miss him.

A) 월요일 좋아

B) 혼자 먹으면 그게 1인분이다

C) 주먹으로 보자기 이기는 법

D) 교수님 제발 B만 주세요

나머지 선택지들은 다른 콘테스트에 관한 것임.

# Introduction : Quality Ranking 예시


🏆

*You both know Jane*

-vs-

Accounting meet archives.

마동석을 화나게 하면?

-vs-

주먹으로 보자기 이기는 법

다른 결선 진출자들과 비교해서 최종 당선작을 고르는 Task 이다.

# Introduction：Explanation Generation 예시

## 주먹으로 보자기를 이기는 법

일반적으로 주먹은 보자기를 이길 수 없다. 일반적인 가위바위보의 상성에 의해서 주먹은 보자기에게 항상 지기 때문이다. 하지만 보자기를 뚫을 정도의 주먹이라면 보자기도 이길 수 있을 것이다. 가위 바위 보에서 아무리 보자기를 내보아도 힘 센 주먹 앞에서는 분명히 질 것이다.

# Datasets

- Our corpus compiles 14 years of weekly New Yorker caption contests.

Each contest consists of:
(1) a captionless cartoon;
(2) that week's entries;
(3) the three finalists, selected by New Yorker editors;
(4) for some contests, quality estimates for each submission collected via crowdsourcing.

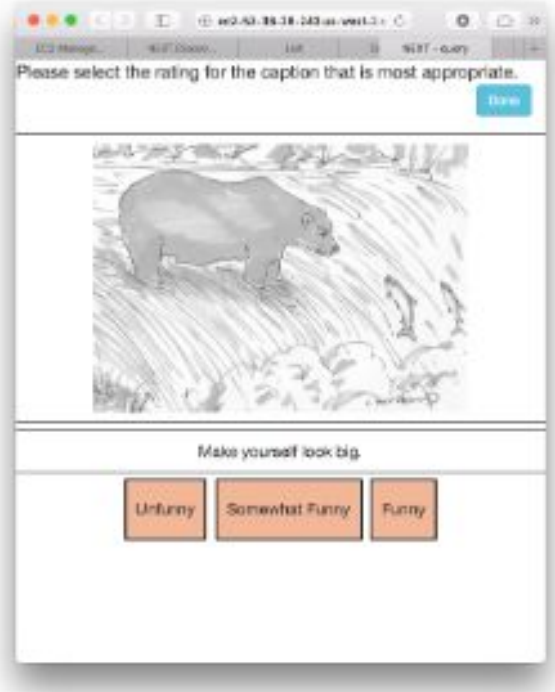
The corpus was constructed from two sources.
The first is Jain et al. (2020), from which we obtain roughly 250 contests (mean/median 6.1K/5.7K unique captions per contest; 1.5M total)
Crowd ratings in this corpus are gathered via the NEXT platform (Jamieson et al., 2015;
Tanczos et al., 2017)

The second corpus, due to Shahaf et al. (2015); Radev et al. (2016) and derived from contests #1-#507, includes 2M unique captions, but no crowd ratings. We

# Datasets

- Crowd ratings in this corpus are gathered via the NEXT platform

"이 캡션이 얼마나 재밌나요?"




"어떤 자막이 더 웃긴가요?"

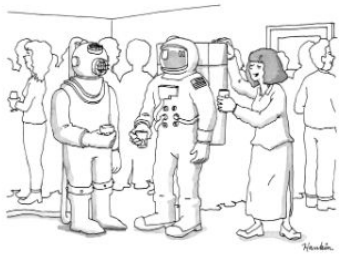
# Datasets

- Crowd ratings in this corpus are gathered via the NEXT platform

| | rank | funny | somewhat_funny | unfunny | count | score | precision | contest | caption |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2575 | 2327 | 1783 | 6685 | 2.118474 | 0.009769 | 586 | Of course it looks alien to you. You've never ... |
| **1** | 2 | 2164 | 2448 | 1500 | 6112 | 2.108639 | 0.009806 | 586 | I have a feeling they are going to treat us li... |
| **2** | 3 | 1771 | 1543 | 1474 | 4788 | 2.062030 | 0.011865 | 586 | Left on Pennsylvania Avenue, big white house o... |
| **3** | 4 | 1901 | 2088 | 1595 | 5584 | 2.054799 | 0.010564 | 586 | And they say life can't exist in a vacuum. |
| **4** | 5 | 1114 | 1221 | 957 | 3292 | 2.047691 | 0.013801 | 586 | Wow. Aliens really do perform the jobs most Am... |

# Datasets

- The first is Jain et al. from which we obtain roughly 250 contests

| | | Top rated | The New Yorker Winner | | |
|---|---|---|---|---|---|
| 676 대시보드 |  | "계란을 평생 먹을 수 있었던 거 알지?" | "모으는 방법은 말하지 않겠습니다." | 2019년 9월 23일 | 913,555 |
| 675 대시보드 |  | "여보, 친절하게 대하세요... 그는 훨씬 더 많은 압박을 받고 있어요." | "둘이 이상하게 잘 어울리는 것 같군." | 2019년 9월 16일 | 655,786 |
| 674 대시보드 |  | "그가 당신의 공을 지키게 합시다." | "그가 당신의 공을 지키게 합시다." | 2019년 9월 9일 | 1,150,449 |

# Datasets

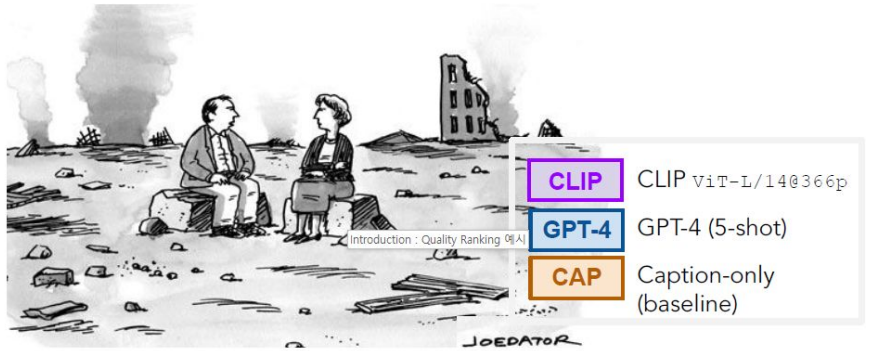
- The second corpus, due to Shahaf et al. (2015); Radev et al. (2016)

What's it going to take to get you in this car today?
Relax! It just smells the other car on you.
It runs entirely on legs.
Just don't tailgate during mating season.
It's only been driven once.
He even cleans up his road kill.
The spare leg is in the trunk.
Comfortably eats six.
She runs like a dream I once had.

**Figure 1: Example cartoon from the New Yorker contest, with the shortlist of submitted captions.**

# Task Setups

- Matching : Five choices are given, only one of which truly corresponds.



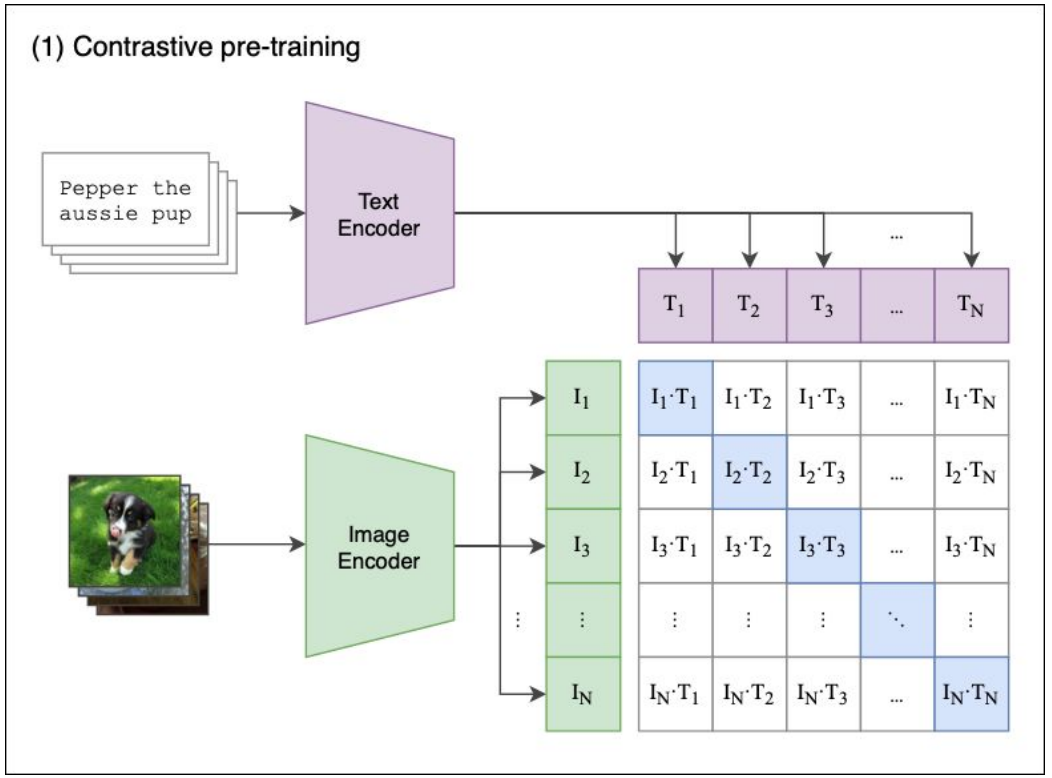

1. From Pixels Setting : 이미지와 Caption만 보고 Task를 수행하는 세팅

2. From Description Setting : 이미지에 대한 설명과 Caption을 보고 Task를 수행하는 세팅

# Task Setups

1. From Pixels Setting : 이미지와 Caption만 보고 Task를 수행하는 세팅

1.1 fine-tune CLIP (ViT-L/14@366px) : explanation task에서는 사용 X



# Task Setups

1. From Pixels Setting : 이미지와 Caption만 보고 Task를 수행하는 세팅
1.2 OFA -> T5 : same format as the human-authored descriptions 으로 이미지에 대한 설명을 생성. 이 설명을 언어모델의 입력으로 넣어줌.

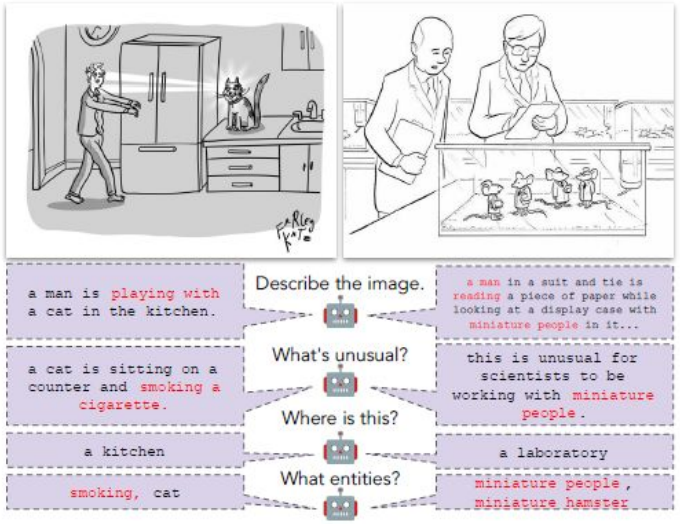



Figure 4: Example predictions by fine-tuned OFA-Huge from images. The model recognizes many objects/actions/locations, but makes some mistakes (indicated in red): for the left image, for example, it falsely indicates that the cat is smoking, and, on the right, that the mice are small people or hamsters (hamsters have stubby tails). Cartoons by Farley Katz and Paul Noth.

in Figure 3. To facilitate this, in addition to providing OFA with the image, we also provide it with a per-annotation-type prompt:

1. for locations: "Where does this take place?"
2. for descriptions: "Describe this image."
3. for uncanny: "What's unusual about this image?"
4. for entities: "What entities are there?"

# Task Setups

1. From Pixels Setting : 이미지와 Caption만 보고 Task를 수행하는 세팅
1.2 OFA -> T5 : same format as the human-authored descriptions 으로 이미지에 대한 설명을 생성.
이 설명을 언어모델의 입력으로 넣어줌.

In this task, you will see a description of an uncanny situation. Then, you will see five jokes — only one of which was written about the described situation. Pick which of the five choices truly corresponds to the described scene.
###
This scene takes place in the following location: boardroom. Four birds are in an office. They're perched around a table. Birds don't have offices. The scene includes: Parrot, Speech repetition.
one of the following funny captions is most relevant to the scene:

A) Just be glad he's not wearing his kilt today.
B) The founding fathers were clear. You must win by two.
C) She'll appreciate you're wearing protection.
D) We have to stop eating the seed money.
E) Can I interest you in opening an offshore account?

the funny caption that matches the scene is:

# Task Setups

2. From Description Setting : 이미지에 대한 설명과 Caption을 보고 Task를 수행하는 세팅

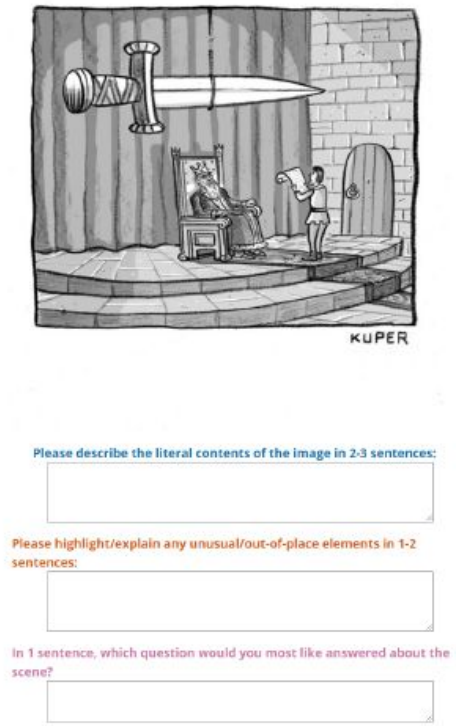
We collect several types of annotations about the 704 cartoons; these either serve as input to models in the from description setting or as additional information available only at training time in the from pixels setting.

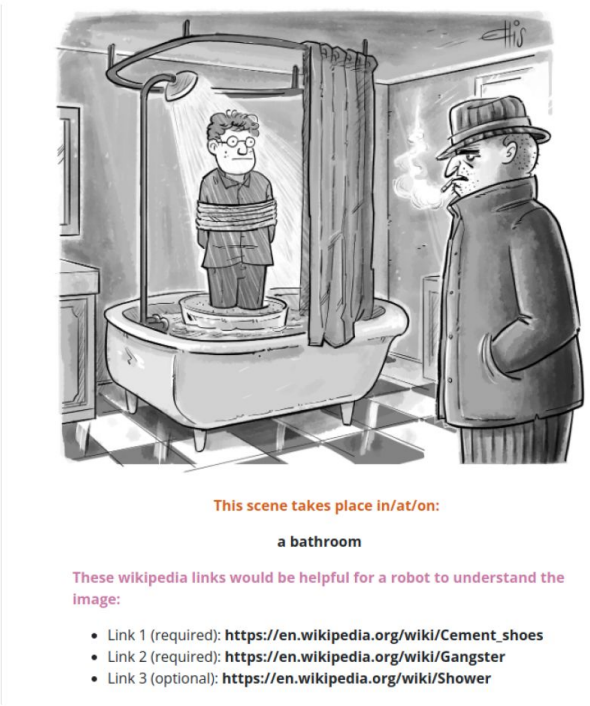
(i) A phrase describing the setting of the scene, e.g., “an office” or “the park” (2 per cartoon)

(ii) A literal 1–3 sentence description of the scene (3 per cartoon)

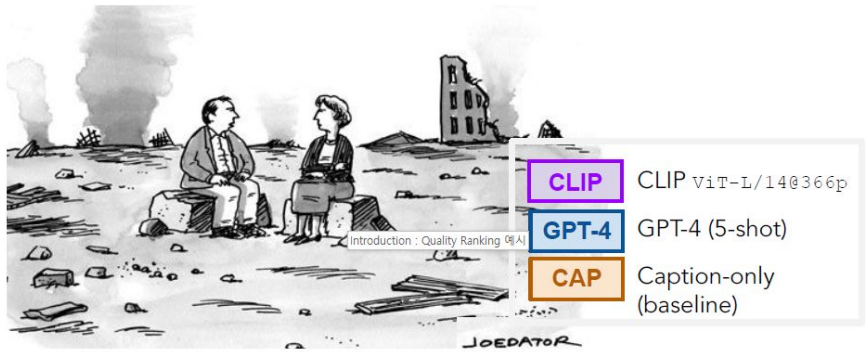
(iii) A 1–3 sentence description or explanation of what makes the scene unusual (3 per cartoon)

(iv) 2–3 EnglishWikipedia links that an annotator identified as relevant, to serve as a proxy for world knowledge (2 per cartoon)



# Experiments




**Matching**

A) You should be happy. How many husbands even notice window treatments? CAP ✗

B) I've led a good life, but now it's time to meet my raker.

*C) I'd like to see other people.* GPT-4 ✓

D) I think it's called an air B&B.

E) We have to turn back. I forgot my scarf. CLIP ✗

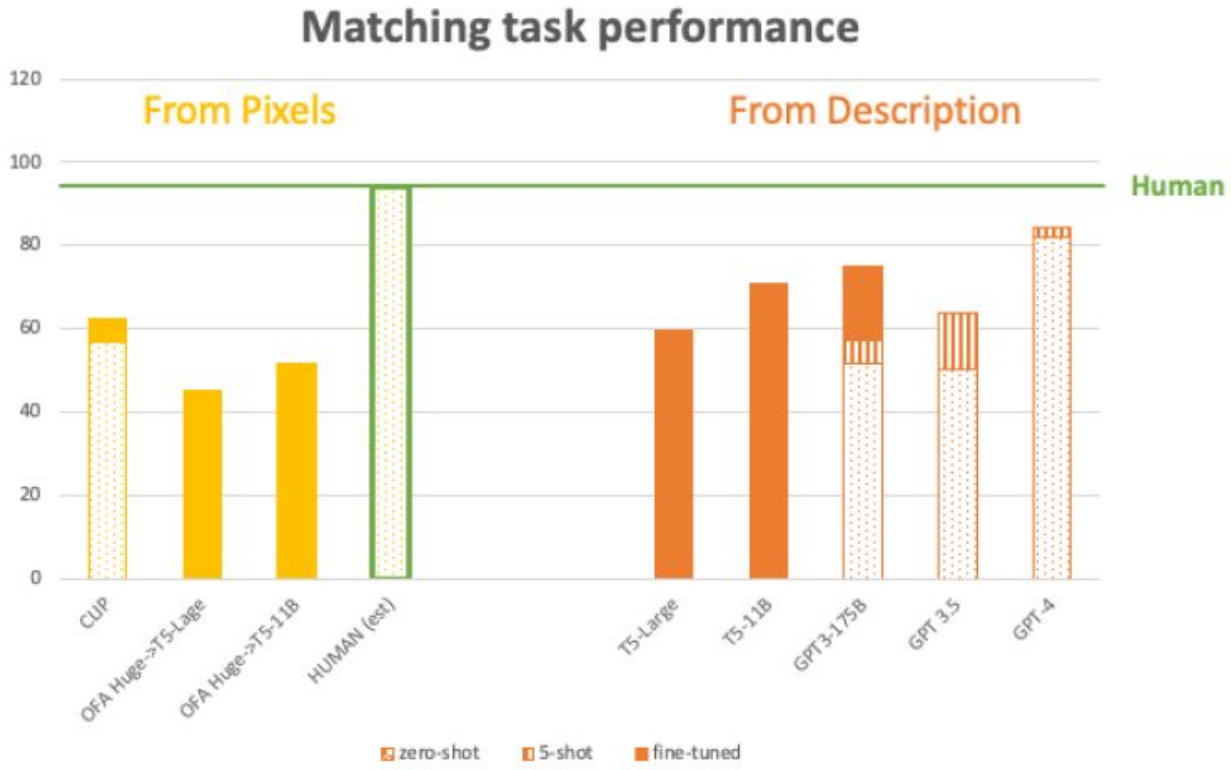
찍었을 때 확률

| | | Matching | Quality Ranking | |
|---|---|---|---|---|
| | | Accuracy (↑) | CrowdAcc (↑) | NYAcc (↑) |
| | Random | 20.0 | 50.0 | 50.0 |
| | Caption Only (T5-11B) | 19.4 | 59.4 | 64.5 |
| FP | CLIP ViT-L/14@336px (finetuned) | 62.3 | 57.0 | 66.9 |
| FP | ↳ Zero-shot | ↳ 56.6 | ↳ 55.8 | ↳ 56.8 |
| FP | OFA-Huge → T5-Large | 45.2 | 59.1 | 64.3 |
| FP | OFA-Huge → T5-11B | 51.8 | 60.3 | 65.0 |
| FD | T5-Large | 59.6 | 61.8 | 64.8 |
| FD | T5-11B | 70.8 | 62.3 | 65.6 |
| FD | GPT3-175B (finetuned) | 75.1 | 64.8 | **69.8** |
| FD | ↳ 5-shot | ↳ 57.2 | ↳ 55.1 | ↳ 54.8 |
| FD | ↳ Zero-shot | ↳ 51.6 | ↳ 56.2 | ↳ 55.6 |
| FD | GPT 3.5 (5-shot) | 63.8 | 55.6 | 55.2 |
| FD | ↳ Zero-shot+CoT | ↳ 50.4 | ↳ 52.8 | ↳ 55.4 |
| FD | GPT-4 (5-shot) | **84.5** | **73.3** | 68.2 |
| FD | ↳ Zero-shot+CoT | ↳ 81.9 | ↳ 66.2 | ↳ 64.3 |
| | Human Estimate From Pixels (FP) | 94.0 | 83.7 | 64.6 |

편집자 취향

# Experiments

| | | Matching | Quality Ranking | |
|---|---|---|---|---|
| | | Accuracy (↑) | CrowdAcc (↑) | NYAcc (↑) |
| | Random | 20.0 | 50.0 | 50.0 |
| | Caption Only (T5-11B) | 19.4 | 59.4 | 64.5 |
| FP | CLIP ViT-L/14@336px (finetuned) | 62.3 | 57.0 | 66.9 |
| FP | ↳ Zero-shot | ↳ 56.6 | ↳ 55.8 | ↳ 56.8 |
| FP | OFA-Huge → T5-Large | 45.2 | 59.1 | 64.3 |
| FP | OFA-Huge → T5-11B | 51.8 | 60.3 | 65.0 |
| FD | T5-Large | 59.6 | 61.8 | 64.8 |
| FD | T5-11B | 70.8 | 62.3 | 65.6 |
| FD | GPT3-175B (finetuned) | 75.1 | 64.8 | **69.8** |
| FD | ↳ 5-shot | ↳ 57.2 | ↳ 55.1 | ↳ 54.8 |
| FD | ↳ Zero-shot | ↳ 51.6 | ↳ 56.2 | ↳ 55.6 |
| FD | GPT 3.5 (5-shot) | 63.8 | 55.6 | 55.2 |
| FD | ↳ Zero-shot+CoT | ↳ 50.4 | ↳ 52.8 | ↳ 55.4 |
| FD | GPT-4 (5-shot) | **84.5** | **73.3** | 68.2 |
| FD | ↳ Zero-shot+CoT | ↳ 81.9 | ↳ 66.2 | ↳ 64.3 |
| | Human Estimate From Pixels (FP) | 94.0 | 83.7 | 64.6 |




Finetuned CLIP tends to do best for matching in the from pixels setting but OFA+T5-11B is competitive for quality ranking.

Among the from description models, GPT-4 (5-shot) generally performs Best. It (and fine-tuned GPT-3) also perform better at predicting New Yorker editor selections than our three humans.

# Experiments

In this task, you will see a description of an uncanny situation. Then, you will see a joke that was written about the situation. Explain how the joke relates to the situation and why it is funny.
###

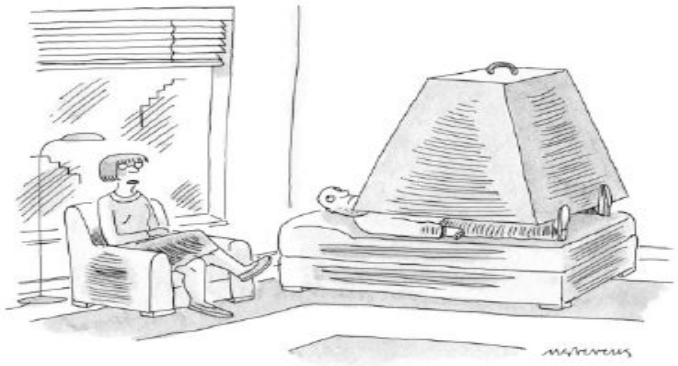
… (4-shot examples)

###
This scene takes place in the following location: a desert. There is a man rowing a boat across a sandy desert who is looking down at another man next to his boat. The other man is crawling on all fours and is
looking up and talking to the man in the boat. It is unusual that a man is rowing a boat through sand. It is also unusual to find a man crawling through a desert on all fours. The scene includes: Rowing, Thirst.
caption: Long time no sea.
explanation of the caption:
A play on the term “sea” — “long time no see” is something people say to each-other after they meet up for the first time in an extended period. But here, ”sea” transforms the meaning to mean: “it’s been a long
time since you’ve been in the ocean,” which is an accurate comment for someone who has been paddling a boat through the desert for many miles.
###
This scene takes place in the following location: an office. There are three man in the room of an office building. One man is carefully trying to place files inside a filing cabinet. There is a tiger on a leash sitting
on top of the filing cabinet. There is a wild tiger inside of an office building. The scene includes: Filing cabinet, Tiger.
caption: You know, I’m gonna miss this when we go paperless.
explanation of the caption:

# Experiments

| | A | B | % A wins | # ratings | G-$\gamma$ |
|---|---|---|---|---|---|
| Q1 | T5-11B | Caption only | 84.7% | 393 | 64.4 |
| Q2 | T5-11B | OFA $\rightarrow$ T5-11B | 74.6% | 393 | 41.6 |
| Q3 | T5-11B | T5-Large | 68.5% | 390 | 45.9 |
| Q4 | FT-GPT-3 | In context GPT-3 | 50.0% | 396 | 23.2 |
| Q5 | 5-shot GPT-4 | Zero-shot GPT-4 | 64.3% | 396 | 19.7 |
| Q6 | 5-shot GPT-4 | 5-shot GPT-3 | 93.0% | 384 | 86.4 |
| Q7 | Human | 5-shot GPT-4 | 67.7% | 390 | 20.9 |

Table 3: Pairwise human evaluations for explanation, with per-instance agreement according to Gwet's (2014) $\gamma$. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.



I wish you'd come in when it was just a chip on your shoulder.

Please take a moment to think about the above caption/cartoon. Explanation choices will appear after 10 seconds.

Select the more complete, relevant, and convincing explanation of the joke.
*If they are both bad, do your best to pick the more relevant/correct one.*

**Explanation 1**

"Chip" is an expression describing mental health issues, such as depression, anxiety, addiction, etc. Here, the patient is an old man with a chip on his shoulder. It's funny in part because the old man is saying "chip" differently from the usual way we say "chip on your shoulder", i.e., when someone has a chip on their shoulder, they might just carry it around all day, and not realize it's bothering him.

**Explanation 2**

Typical patient narratives involve someone "coming in with a chip on your shoulder" and 'getting treated'. But, this patient is so morbid, he literally carries his entire mental illness around on his body. So, here, it's not like he was just a chip on his shoulder; instead, he's carrying a weight bearing down on his head.

# Experiments

| | A | B | % A wins | # ratings | G-$\gamma$ |
|---|---|---|---|---|---|
| Q1 | T5-11B | Caption only | 84.7% | 393 | 64.4 |
| Q2 | T5-11B | OFA $\rightarrow$ T5-11B | 74.6% | 393 | 41.6 |
| Q3 | T5-11B | T5-Large | 68.5% | 390 | 45.9 |
| Q4 | FT-GPT-3 | In context GPT-3 | 50.0% | 396 | 23.2 |
| Q5 | 5-shot GPT-4 | Zero-shot GPT-4 | 64.3% | 396 | 19.7 |
| Q6 | 5-shot GPT-4 | 5-shot GPT-3 | 93.0% | 384 | 86.4 |
| Q7 | Human | 5-shot GPT-4 | 67.7% | 390 | 20.9 |

Table 3: Pairwise human evaluations for explanation, with per-instance agreement according to Gwet's (2014) $\gamma$. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Is computer vision a bottleneck for top quality explanation generation?

Does GPT-4 outperform GPT-3?

Does our best model, GPT-4, explain jokes as well as humans?

# Experiments

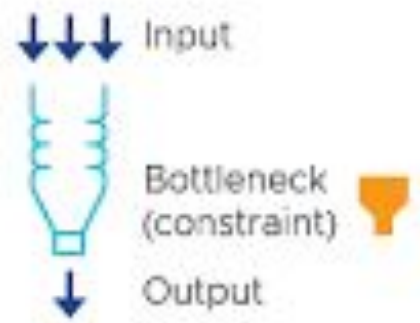
| | A | B | % A wins | # ratings | G-$\gamma$ |
|---|---|---|---|---|---|
| Q1 | T5-11B | Caption only | 84.7% | 393 | 64.4 |
| Q2 | T5-11B | OFA $\rightarrow$ T5-11B | 74.6% | 393 | 41.6 |
| Q3 | T5-11B | T5-Large | 68.5% | 390 | 45.9 |
| Q4 | FT-GPT-3 | In context GPT-3 | 50.0% | 396 | 23.2 |
| Q5 | 5-shot GPT-4 | Zero-shot GPT-4 | 64.3% | 396 | 19.7 |
| Q6 | 5-shot GPT-4 | 5-shot GPT-3 | 93.0% | 384 | 86.4 |
| Q7 | Human | 5-shot GPT-4 | 67.7% | 390 | 20.9 |

Table 3: Pairwise human evaluations for explanation, with per-instance agreement according to Gwet's (2014) $\gamma$. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Is computer vision a bottleneck for top quality explanation generation?

Yes. Compared to the same model trained with access to human written descriptions available at test time
(i.e., the from description setting),
the model trained with access only to OFA-predictions loses in 74.6% of cases.

왼쪽은 Description Setting (인간의 시각)



# Experiments

| | A | B | % A wins | # ratings | G-$\gamma$ |
|---|---|---|---|---|---|
| Q1 | T5-11B | Caption only | 84.7% | 393 | 64.4 |
| Q2 | T5-11B | OFA $\rightarrow$ T5-11B | 74.6% | 393 | 41.6 |
| Q3 | T5-11B | T5-Large | 68.5% | 390 | 45.9 |
| Q4 | FT-GPT-3 | In context GPT-3 | 50.0% | 396 | 23.2 |
| Q5 | 5-shot GPT-4 | Zero-shot GPT-4 | 64.3% | 396 | 19.7 |
| Q6 | 5-shot GPT-4 | 5-shot GPT-3 | 93.0% | 384 | 86.4 |
| Q7 | Human | 5-shot GPT-4 | 67.7% | 390 | 20.9 |

Table 3: Pairwise human evaluations for explanation, with per-instance agreement according to Gwet's (2014) $\gamma$. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Does GPT-4 outperform GPT-3?

Answer: Yes, definitely.
In our most definitive result, with equal amounts of supervision, GPT-4's explanations are preferred nearly universally — specifically, in 93% of cases. Interestingly, GPT-3 performs slightly better on automatic evaluation metrics for explanation like BLEU-4 and Rouge-L (see Appendix E),
which suggest that the earlier family of may fit the surface features of the generation task more effectively, e.g., 5-shot GPT-3 achieves 5.07 BLEU-4 compared to 4.99 for 5-shot GPT-4. This suggests that mirroring the surface form of our explanation corpus is not sufficient to generate the highest quality explanations.

# Experiments

| | A | B | % A wins | # ratings | G-$\gamma$ |
|---|---|---|---|---|---|
| Q1 | T5-11B | Caption only | 84.7% | 393 | 64.4 |
| Q2 | T5-11B | OFA $\rightarrow$ T5-11B | 74.6% | 393 | 41.6 |
| Q3 | T5-11B | T5-Large | 68.5% | 390 | 45.9 |
| Q4 | FT-GPT-3 | In context GPT-3 | 50.0% | 396 | 23.2 |
| Q5 | 5-shot GPT-4 | Zero-shot GPT-4 | 64.3% | 396 | 19.7 |
| Q6 | 5-shot GPT-4 | 5-shot GPT-3 | 93.0% | 384 | 86.4 |
| Q7 | Human | 5-shot GPT-4 | 67.7% | 390 | 20.9 |

Table 3: Pairwise human evaluations for explanation, with per-instance agreement according to Gwet's (2014) $\gamma$. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Does our best model, GPT-4, explain jokesas well as humans?

Answer: No. Human-written explanations are preferred by annotators in 68% of pairwise cases. We qualitatively examine the 39/130 cases where the human reference receives 3/3 annotator votes. In these cases, the machine-generated explanations usually incorrectly interpret the image, e.g., in one case, a caption jokes about two cave people in a hole looking at a caveman in a cave with the caption "Personally, I'm not a big fan of modern architecture."; GPT-4 incorrectly interprets the hole as "modern architecture" instead of the cave. We also examine the 8/130 cases where the GPT-4 produced caption was unanimously preferred: a close reading of these cases is provided in Appendix F. In 3 of these 8 cases, the human explanations, while on the right track, had slight inaccuracies, and in the remaining 5 cases, the human and machine explanations both express the same idea, but with different styles (GPT-4's sometimes arguably being more formal, detailed, or fluent).

# Conclusion

1. We demonstrate that today's vision and language models still cannot recognize caption relevance, evaluate (at least in the sense of reproducing crowdsourced rankings),or explain The New Yorker Caption Contest as effectively as humans can.

2. our matching/quality ranking models could help entrants receive quantitative feedback on the relevance/predicted quality of their submissions.

3. a promising avenue for future work focused on generating humorous captions (c.f. our focus of humor "understanding" benchmarks) would be to operationalize the feedback provided by our matching/ranking models.

4. We cannot claim to know whether the human-machine 'humor understanding gap' will be closed sooner or later. But we encourage other researchers to have as much fun with the topic as we did!

# Limitations

1. the results of this study do not represent or cover all types of humor.

2. Our tasks operationalize the prediction of only average preferences (rather than individual ones) and these preferences may include a partiality or bias towards items that conform to the characteristics of prior contest winners or published New Yorker cartoons.

3. Finally, the explanations in our annotated corpus were largely written by a single author of this paper. Future work would be well-suited to compare the person-to person variance in explaining why particular jokes are funny.

# Thank You

# Appendix：Measuring preferences over captions.

While humor is ultimately subjective, work on the contest has studied modeling average preferences of raters. Tanczos et al. (2017) design quality ranking algorithms for the caption contest, framed as identifying the best “arm” in a multi-armed bandit setting;

their crowdsourcing system NEXT (Jamieson et al., 2015) is used by The New Yorker. It does not directly use the content of the cartoons/contests. The result is Jain et al. (2020)’s continuously updated corpus, from which we draw some of our data.

# Appendix :

| Scene/caption | Our opinion re: annotator decision | Human-authored explanation | Machine-authored explanation (GPT-4 5-shot) |
|---|---|---|---|
| A person in a diver suit is out-of-place at a cocktail party. *Caption: This suit looked way better in the store.* | *Agree:* The human annotation assumes the person didn't notice; the machine explanation is more on point/coherent. | Sometimes, clothes you try on in a store look better when you are deciding to buy them or not. Here, the idea is taken to the extreme: the person in the diving suit bought the "suit" without even noticing it was a diving suit (and looks out of place for the party)/ (0/3) | We all experience situations where a garment looks great in the store, but when we wear it out in a public setting, we feel awkward or mismatched. Here, this scene humorously exaggerates that feeling by having the man wear an old diving suit to a seemingly professional office party. The situation creates comedic tension because it's an extreme example of being underdressed or out of place. (3/3) |
| A man is on a cell phone is in mid-air despite being thrown off a bucking horse. He is smiling. *Caption: Hi Hon. Can you meet me at the hospital in about ten minutes?* | *Somewhat agree:* The human explanation says "friend" instead of spouse/partner, but the machine authored explanation could also be improved ("mismatch" is less relevant). | The man on the phone is pre-planning his own hospital visit. This is unusual, because one would think that he would be focused on avoiding injury from the currently-occuring situation instead of, in a surprisingly carefree and cheerful manner, telling his friend to meet him at the hospital. (0/3) | The man in the suit is attempting to ride a bucking horse, which implies that he is not experienced in horse riding or else dressed inappropriately for it. The caption suggests he is bound to have an accident as a result of this mismatch, and so he is calling ahead to his spouse to be prepared to meet him at the hospital. It's a humorous way of showcasing that the man knows he is in a dangerous and ill-suited situation. (3/3) |
| A man in an office building is planting seedlings in his cubicle. *Caption: The third quarter is going to be delicious.* | *Mostly agree:* Although the machine explanation refers to documents that are not explicitly evident, the human explanation assumes a less likely interpretation of "delicious". | A play on the term "delicious" — in a usual office setting, if third quarter was going to be delicious, it would be figuratively so, i.e., with juicy gossip, people getting fired, lots of drama, etc. But here, the third quarter will be literally delicious because the person growing vegetables for the company will share them in the future, which will make for delicious food. It's funny in part because the usual interpretation is about dark enjoyment of office politics, but the literal interpretation is about a wholesome vegetable feast. (0/3) | In the business world, people often discuss quarters in terms of profits and financial performance. The joke here is that the man is literally "planting" invoices or financial documents, hoping to grow something metaphorically delicious or profitable. The humor comes from the absurdity of treating financial documents as plants and expecting positive results from this unorthodox method. (3/3) |

# Appendix :

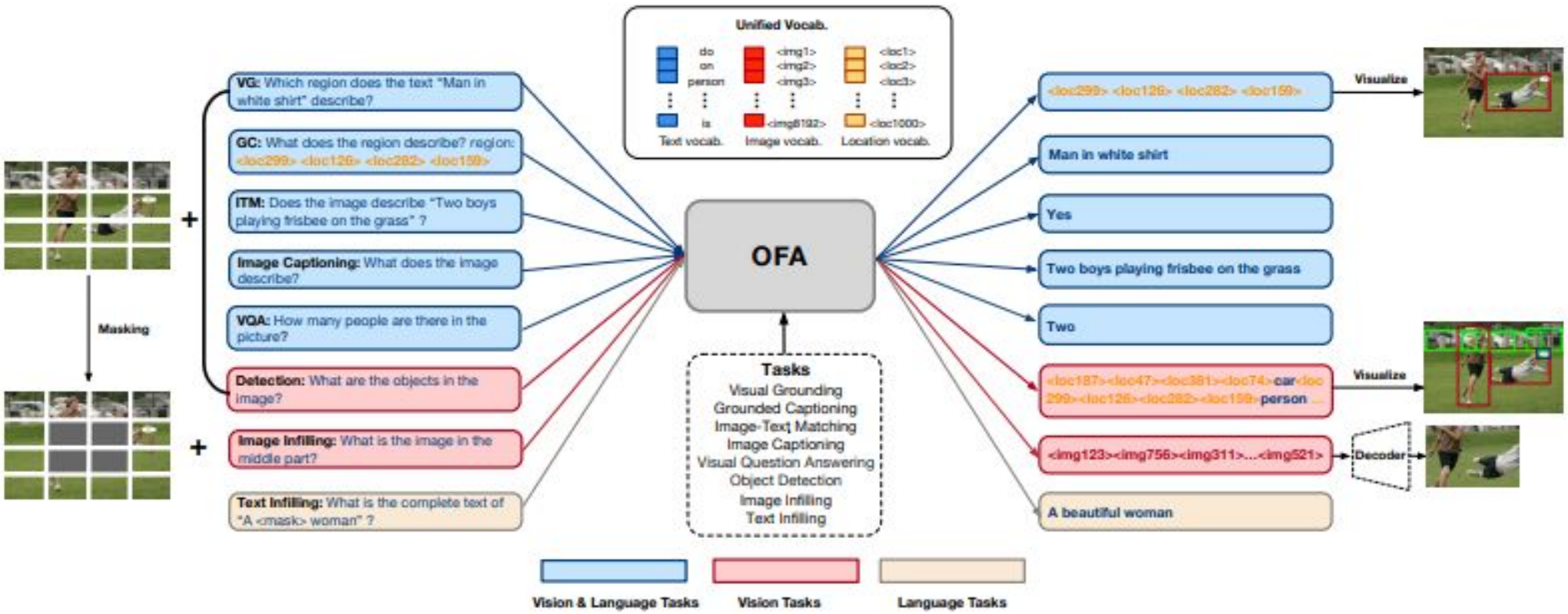
| | | | |
|---|---|---|---|
| In a 17th century scene, a woman is walking in town, but her ruff collar is comically large. *Caption: It deployed when her carriage rear-ended an ox cart.* | *Mostly agree:* Both are good, but the machine explanation has a more specific conclusion. | The woman's oversized ruff collar is juxtaposed against the modern invention of an air bag. Air bags deploy during car crashes — and here, the collar inflated during a collision between a carriage she was riding in and an ox cart as a safety mechanism. It's funny in part because, while it explains the oversized collar, collars are cosmetic rather than functional. (0/3) | The joke is a play on modern-day car safety features, like airbags, that deploy in an accident to protect passengers. In this case, instead of an airbag, the huge Victorian-era collar "deployed" when the woman's carriage rear-ended an ox cart. This absurdity of applying modern safety concepts to historical clothing makes the caption funny. (3/3) |
| A wolf trying on a sheep skin as if it were a costume, looks in a mirror as a butler looks on, holding various other outfits. *Caption: I'll take this and the granny dress.* | *Mostly agree:* The machine explanation is more specific. | A reference to the wolf from Little Red Riding Hood. A "wolf in sheeps clothing," as depicted, is an idiom meaning a nefarious actor who appears innocent: here the wolf is the same (also nefarious) wolf as in Little Red Riding Hood: who kills a grandmother and wears her clothes as a disguise. (0/3) | The wolf trying on a sheep's skin and asking for a granny dress is a humorous reference to two popular fairy tales involving wolves pretending to be something else—first being the idiom "a wolf in sheep's clothing," which means someone pretending to be harmless, and the second referencing "Little Red Riding Hood," where the wolf famously impersonates the girl's grandmother. The joke highlights how the wolf is taking these disguises quite literally in a modern-day retail setting. (3/3) |
| A flying saucer trails a man and a woman driving down the road. *Caption: When they said, "Take us to your leader." I don't think they meant your mother's house.* | *Slightly disagree:* Both take the joke in slightly different ways, but the human explanation seems more plausible. | This is a joke about in-laws. Pop culture depictions of aliens often have them requesting that the first humans they make contact with take them to the leaders of humanity. The woman is driving them to her mother's house, and the man, who feels stifled by his mother-in-law, is reminding his wife that her mom isn't their all-powerful leader. It's funny that this benign concern is on his mind in light of the fact that a giant spaceship is tailing them. (0/3) | When people think of aliens asking to be taken to our leaders, they usually mean they want to meet important authority figures or decision-makers like the president or other government personnel. The joke here is that the car driver, misunderstanding or maybe just messing with the aliens, is taking the aliens to their mother's house, which is vastly different from what they usually expect but incidentally drives the point people's mothers can sometimes be really dictatorial figures even if they aren't technically their leader. (3/3) |
| A man looks in a bathroom mirror as he gets ready, but he looks like Abraham Lincoln in the mirror. *Caption: On second thought let's skip the play.* | *Slightly disagree:* The human one expresses the same idea more succinctly. | A joke about Abraham Lincoln's assassination, which took place at a play. This man is getting ready to go to a play and has a vision of himself as Lincoln; because he wants to avoid being assassinated, due to this premonition, he suggests to his partner that they not go to the play after all. (0/3) | The joke plays on two ideas. First, it suggests that the man's reflection, which appears to be Abraham Lincoln, might be causing the man to think he has some connection or is somehow embodying Lincoln. Second, it references a historical event, as Lincoln was assassinated at a theater while watching a play. So, when the man says "let's skip the play," he is humorously trying to avoid the same fate as Lincoln. (3/3) |

# Appendix :

| | | | |
|---|---|---|---|
| A man driving talks to a a woman in the passenger seat. In the backseat is a cramped giant whale. *Caption: He wants to know if you can move your seat up.* | *Both are good:* both explanations express the same idea in a reasonable way. | Moving a front seat up is an everyday request from people sitting in the backseat of cars if they don't have enough room. The whale is making this request: the contrast between how unusual it is that the whale is in the backseat and the benign-ness of the request is what makes it funny. (0/3) | The humor lies in the absurdity of having a whale in the backseat of a car, a situation that should never occur. The whale, being so large, would naturally require more room, so it's funny that the only concern anyone in the cartoon appears to have is whether the front seat passenger can move their seat up to give the whale more space, instead of questioning why there's a whale in the car in the first place. (3/3) |

Figure 14: Examples of the small number of cases (8/130) where the machine-generated caption (5-shot GPT4) was preferred to the human-authored explanation by 3/3 annotators in our pairwise judgment setting.

# Appendix :


Unified Vocab.
Text vocab.
Image vocab.
Location vocab.
OFA
Masking
VG: Which region does the text "Man in white shirt" describe?
GC: What does the region describe? region:
ITM: Does the image describe "Two boys playing frisbee on the grass" ?
Image Captioning: What does the image describe?
VQA: How many people are there in the picture?
Detection: What are the objects in the image?
Image Infilling: What is the image in the middle part?
Text Infilling: What is the complete text of "A <mask> woman" ?
Tasks
Visual Grounding
Grounded Captioning
Image-Text Matching
Image Captioning
Visual Question Answering
Object Detection
Image Infilling
Text Infilling
Man in white shirt
Yes
Two boys playing frisbee on the grass
Two
Visualize
Visualize
Decoder
A beautiful woman
Vision & Language Tasks
Vision Tasks
Language Tasks