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

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da,
Rowan Zellers, Robert Mankoff, Yejin Choi

Reporter: Xiachong Feng

ACL 2023 Best Paper Awards

Best Paper Awards

- 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 and Yejin Choi
- What the DAAM: Interpreting Stable Diffusion Using Cross Attention
  Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin and Ferhan Ture
- From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models
  Shangbin Feng, Chan Young Park, Yuhan Liu and Yulia Tsvetkov

- **Task**
- **Resource**

Authors

Jack Hessel
Research Scientist @ AI2
PhD in Cornell University

Ana Marasovi´c
Assistant Professor in University of Utah
Ph.D. in Heidelberg University
Postdoctoral, University of Washington

Jena D. Hwang
Research Scientist @ AI2
Ph.D. in University of
Colorado Boulder

Lillian Lee
Professor @ Cornell University

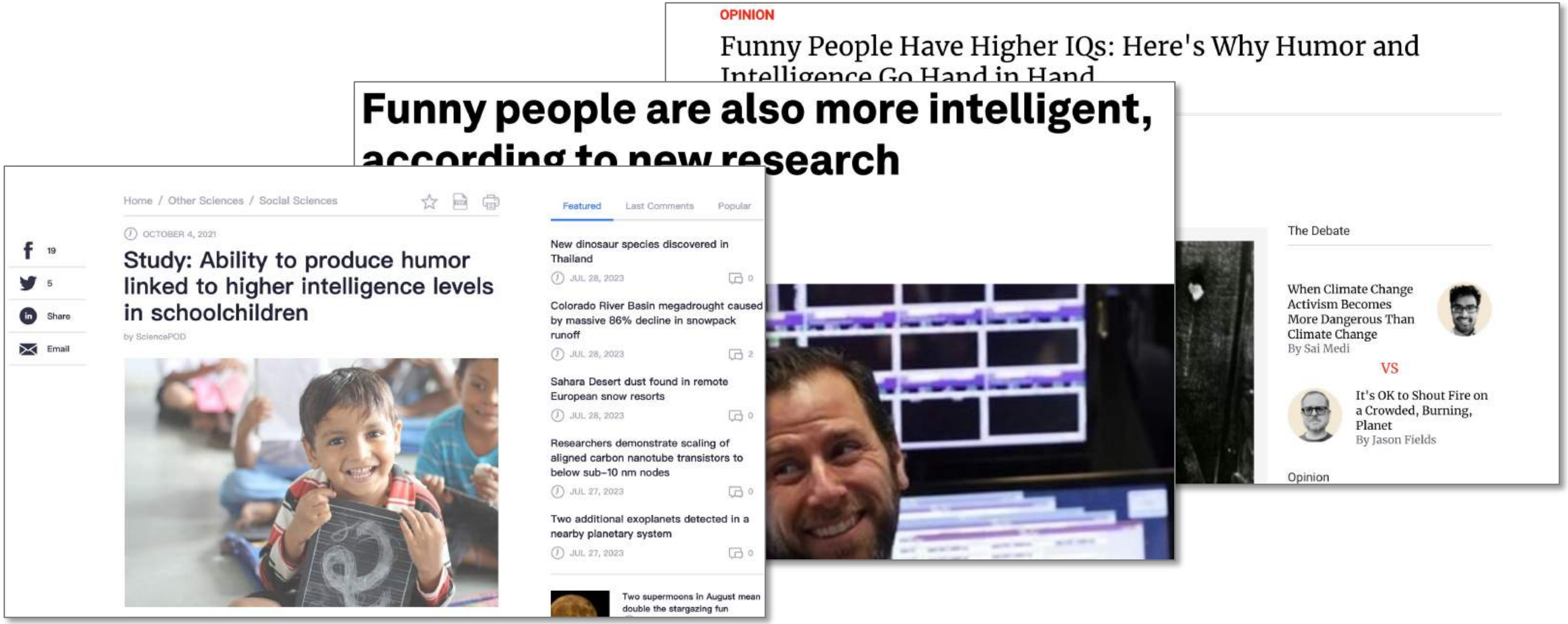
Jeff Da

Rowan Zellers
Researcher at OpenAI

Robert Mankoff

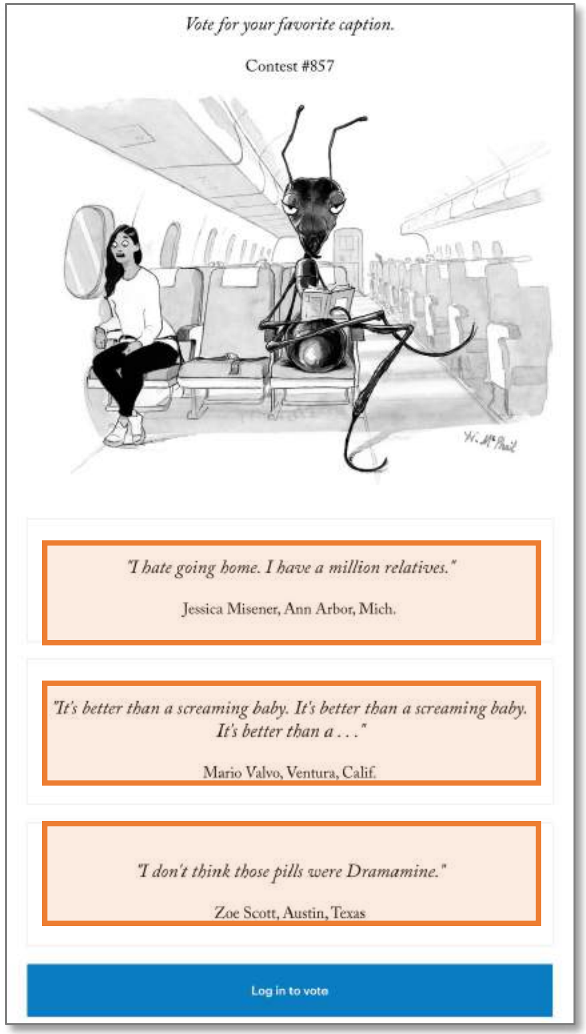
Yejin Choi
MacArthur Fellow

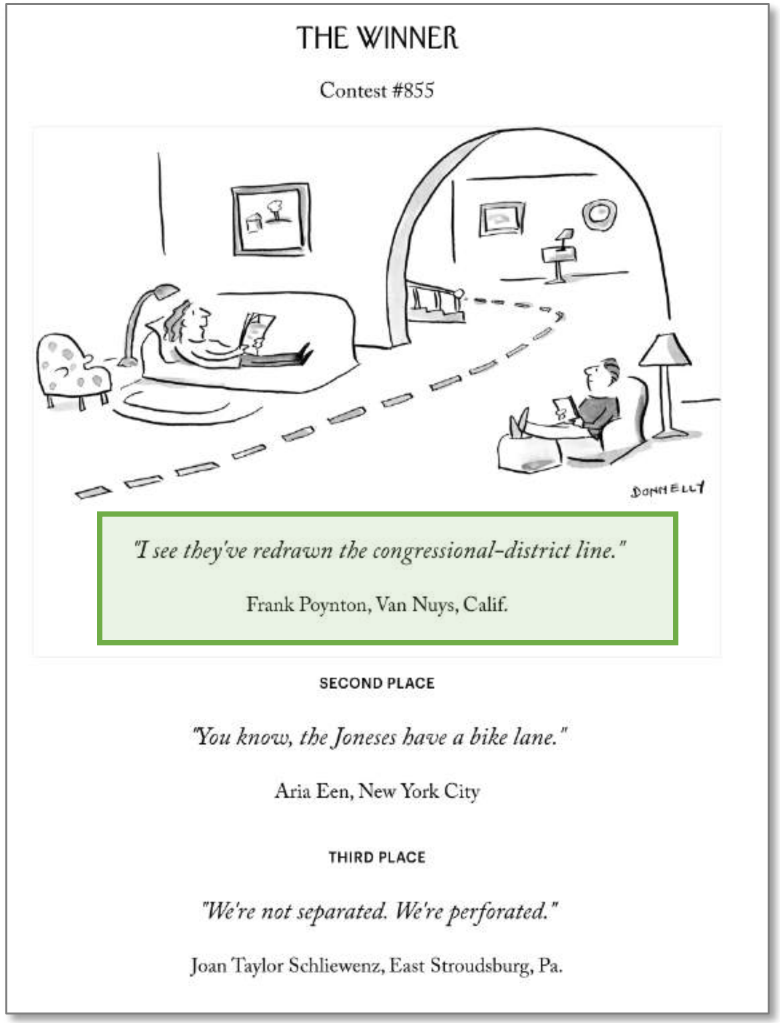
Background: Humor



Humor is a sign of intelligence.

Background: New Yorker Cartoon Caption Contest





https://www.newyorker.com/cartoons/contest

Challenge: Implicit Intelligence

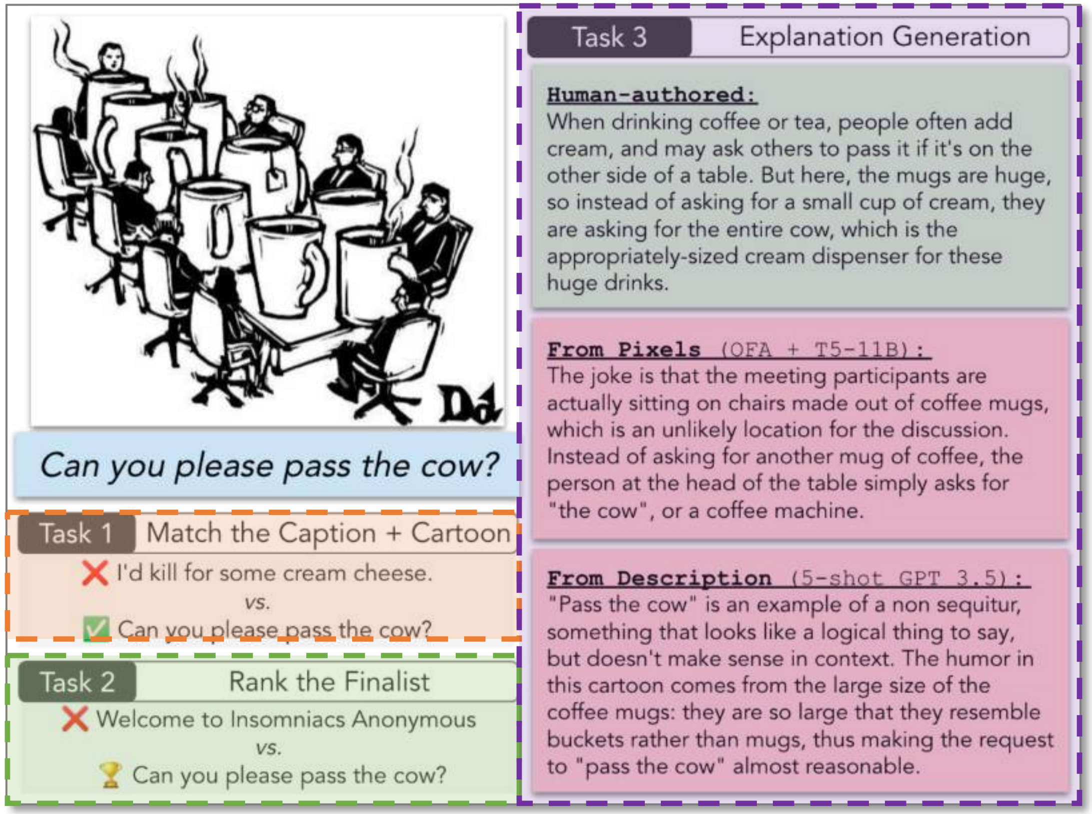
Indirect and playful allusions to human experience and culture.

The size of the mugs must first be recognized as unusual, and then, the caption invokes an association between a large mug and a large amount of cream/milk — perhaps a whole cow's worth.

Can you please pass the cow?

Overview (three tasks)

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



Task 1: Matching

- *Can a model recognize when a caption is appropriate for a given cartoon?*
- **Five** choices are given, only **one** of which truly corresponds.

1. O.K. I'm at the window. To the right? Your right or my right?
2. I'd kill for some cream cheese.
3. Bob just came directly from work.
4. Can you please pass the cow?
5. They only allow one carry-on.

Negative choices are **randomly** selected finalists from other contests

Task 2: Quality Ranking

- *Can a model identify highly rated captions?*
- For each finalist, we sample for comparison a caption that was not selected as a finalist, and **ask models to identify which one** (the real one or the distractor) **was rated as higher quality.**

Can you please pass the cow?

Welcome to Insomniacs Anonymous.

Preprocessing

One round of text-only filtering to discard submissions that are easily identifiable as low quality perform semantic deduplication

Task 3: Explanation

- *Can a model generate as good an explanation as a human for why a caption-andimage combination is funny?*
- Free-form explanations of why captions are funny/appropriate for their corresponding image were written by an author of this paper.
- The rough annotation guidance was: **"In a few sentences, explain the joke as if to a friend who doesn't 'get it' yet."**
- After filtering out cases where the author did not understand the joke, **a corpus of 651 human-created joke explanations to serve as comparison points was formed** (mean/median 60/59 words, 39.3K total)

Human-authored:
When drinking coffee or tea, people often add cream, and may ask others to pass it if it's on the other side of a table. But here, the mugs are huge, so instead of asking for a small cup of cream, they are asking for the entire cow, which is the appropriately-sized cream dispenser for these huge drinks.

Three Tasks

	Matching	Quality Ranking	Explanation Generation
	A) I always figured hell would be less ironic. <u>*B) You both know Jane*</u> 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.	🏆 <u>*You both know Jane*</u> -vs- Accounting meet archives.	<u>*You both know Jane*</u> A reference to Jane Goodall. Goodall is a scientist who is famous for studying chimpanzees, as represented by the ape at the party. This party is likely a scientific conference on biology, but the unusual part is that the subject of the study, the chimp, is invited. Both the peer scientist and the chimpanzee know Goodall, but for different reasons.
	A) Can I interest you in an offshore account? B) So how much of the story is autobiographical? C) Don't give me that holier-than-thou attitude! D) They give me free drinks if I keep my tray table down. <u>*E) Publicly, we are still saying there are no side effects*</u>	I'll admit he may look ugly, but his resume is beautiful. -vs- 🏆 <u>*Publicly, we are still saying there are no side effects*</u>	<u>*Publicly, we are still saying there are no side effects*</u> This is a board meeting of a shady pharmaceutical company. The drug the company makes has the side effect of turning people into cartoon monsters, and most everyone at the company has taken it. Nonetheless, they are choosing not to warn the public. This plays upon a common belief that pharmaceutical companies care more about profits than they do the well-being of their patients.

Datasets

- 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.

# Train/val/test Matching	1.6K / 538 / 538
# Train/val/test Quality ranking	1.6K / 523 / 523
# Train/val/test Explanation	391 / 130 / 130

Jain et al. (2020) starting from #508

- Roughly 250 contests (mean/median 6.1K/5.7K unique captions per contest; 1.5M total),
- Readers rate captions as "funny", "somewhat funny", or "unfunny"; we use the per-caption mean. There are over 114M ratings total (mean/median of 445K/471K per contest).
- Sample three additional top captions that aren't editorial picks to serve as additional "finalists."

Shahaf et al. (2015); Radev et al. (2016) and derived from contests #1-#507

- Includes 2M unique captions (mean/median 5.2K/5.0K per contest)
- *No crowd ratings.*
- Remove by hand 55 contests whose images' resolutions are too low.
- Identify 80 low resolution (but usable) cases, taking special care when annotating this set.

Evaluation

Matching	Quality ranking	Explanation
Accuracy	NYAcc The average accuracy over instances where the finalist was an official New Yorker finalist CrowdAcc “finalist” caption was selected by the crowd as high quality.	Pairwise human evaluations Automatic metrics

Settings

① From Pixels (FP)

Only contest information available is the image itself

② From Description (FD)

factor out visual processing by providing the model with human written annotations

An office
Many people are having a meeting

Description

Locations

"This scene takes place in/at/on a(n)..."

Cartoon 1	a science lab
	a laboratory
Cartoon 2	doctor's office
	doctor's office

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

Descriptions

"Describe the literal contents of the image in 2-3 sentences"

Cartoon 1	A mouse is wearing a jetpack and cheating at a maze. Two scientists stare at him.
	Two scientists are observing a rat making its way 'through a waze to find the cheese. The rat is operating 'a jet pack to skip the maze and go straight to the reward at the end.
Cartoon 2	A safe is on the exam table in a doctor's office. The doctor is listening to the safe with a stethoscope.
	A doctor is in her office and she is using her stethoscope on a patient. The patient in this case is just a large metal safe that is sitting on a chair.

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

Uncanny Descriptions

"Highlight/explain any unusual/out-of-place elements in 1-2 sentences"

Cartoon 1	The mouse is wearing a jetpack to cheat at the maze. Mice are not that intelligent.
	A rat has learned to build and operate a miniature jet pack.
Cartoon 2	It is unusual to see that large of a safe on a table. Also, doctors usually examine living things not safes.
	The doctor is using a stethoscope to check the heartbeat of an object with no heart.

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

Entity Links

"These wikipedia links would be helpful for a robot to understand the image"

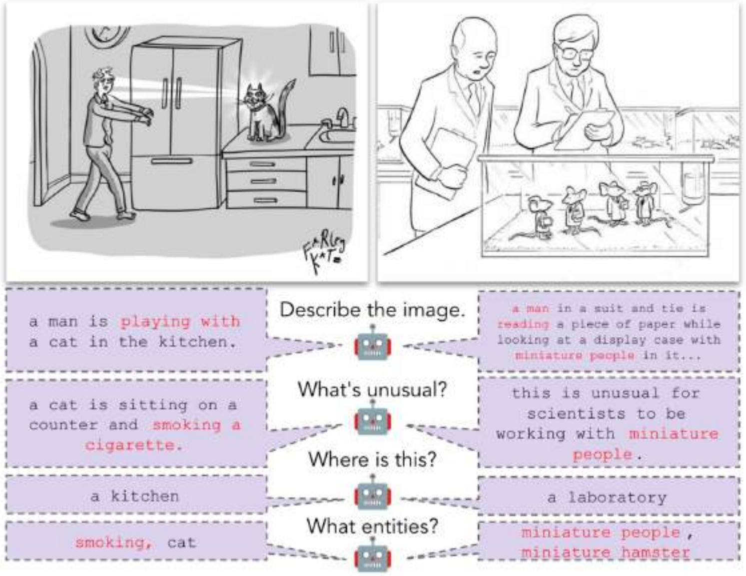
Cartoon 1	wiki/Laboratory_mouse wiki/Tryons_Rat_Experiment
	wiki/Mouse wiki/Jet_pack wiki/Maze
Cartoon 2	wiki/Physical_examination wiki/Stethoscope wiki/Safe-cracking
	wiki/Safe wiki/Stethoscope

2-3 English Wikipedia links that an annotator identified as relevant, to serve as a proxy for world knowledge (2 per cartoon)

From Pixels (FP) Models

- **CLIP**
  - Fine-tune CLIP ViT-L/14@366px
  - Pretrained to align images/captions in the WebImageText corpus
  - For multiple choice, we use InfoNCE (Oord et al., 2018) to encourage the cosine similarity of the cartoon/correct answer to be higher than the incorrect ones.
  - For zero-shot classification, we use the prompt a new yorker cartoon with winning caption
  - CLIP isn't generative, so we can't use it for explanation.

- **OFA → LM**
  - OFA Huge (930M parameters) (Wang et al., 2022), a seq2seq model that supports image/text inputs/outputs
  - Finetune on the New Yorker corpus by training it to map from (cartoon, prompt) → descriptions for the four types of annotations
  - We pass the OFA-predicted outputs to a language model



From Description (FD) Models

- We formulate multiple-choice tasks as text-to-text by concatenating the human-authored cartoon descriptions with the choices as input: the target is simply the letter corresponding to the answer, e.g., **E**.
- For explanation, we autoregressively generate the explanations conditioned on the descriptions/captions.

- **T5**
  - We fine-tune T5-Large and T5-11B
- **GPT-3, GPT-3.5, GPT-4**
  - As both zero-shot and few-shot models
  - Provide the models with a description of the task
  - For the few-shot case, 5 random labelled in-context examples.

Results: Matching and quality ranking results

		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)	<u>62.3</u>	57.0	<u>66.9</u>
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	<u>60.3</u>	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

Results: Human Evaluation of Explanation

	A	B	% A wins	# ratings	G-γ
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) γ. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Q1: Do models utilize the image context of the caption to generate better explanations?

Answer: Yes.
Compared to the same model trained with no access to image information, the model with image information wins in 84.7% of cases.

Results: Human Evaluation of Explanation

	A	B	% A wins	# ratings	G-γ
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) γ. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Q2: Is computer vision a bottleneck for top quality explanation generation?

Answer: 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.

Results: Human Evaluation of Explanation

	A	B	% A wins	# ratings	G-γ
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) γ. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Q3: Do bigger T5 models generate better explanations?

Answer: Yes.
T5-11B with access to the same information at test time as T5-Large (770M) is preferred in 68.5% of cases.

Results: Human Evaluation of Explanation

	A	B	% A wins	# ratings	G-γ
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) γ. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Q4: Does fine-tuning an LLM model help vs. in-context learning for explanation generation?

Answer: Not really.

- We find that in-context explanation generations are comparable to fine-tuned ones according to pairwise human evaluations, *even though the perplexity of the in-context model*, reported in Appendix E, is much higher (107 vs. 21.8).
- We expect that the fine-tuned model more closely mirrors the style of the corpus, but that the in-context explanations also contain similar content, e.g., relevant entities.

Results: Human Evaluation of Explanation

	A	B	% A wins	# ratings	G-γ
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) γ. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Q5: Do supervised explanations help, even with GPT-4?

Answer: Yes.
The zero-shot version of GPT-4 is missing access not only to the supervision of paired (caption, explanation) data, but also, explanations in the detailed style of our released corpus. Perhaps as a result, 5-shot GPT-4 (which also achieves significantly higher BLEU-4/Rouge-L) is preferred in 64% of cases.

Results: Human Evaluation of Explanation

	A	B	% A wins	# ratings	G-γ
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) γ. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Q6: Does GPT-4 outperform GPT-3?

Answer: Yes.

Results: Human Evaluation of Explanation

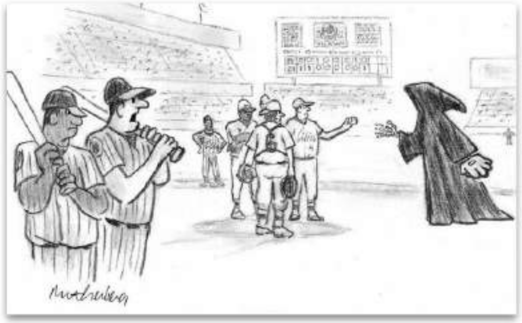
	A	B	% A wins	# ratings	G-γ
Q1	T5-11B	Caption only	84.7%	393	64.4
Q2	T5-11B	OFA $\rightarrow$ T5-11B	74.6%	393	41.6
Q3	T5-11B	T5-Large	68.5%	390	45.9
Q4	FT-GPT-3	In context GPT-3	50.0%	396	23.2
Q5	5-shot GPT-4	Zero-shot GPT-4	64.3%	396	19.7
Q6	5-shot GPT-4	5-shot GPT-3	93.0%	384	86.4
Q7	Human	5-shot GPT-4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

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

Q7: Does our best model, GPT-4, explain jokes as well as humans?

Answer: No.
The machine-generated explanations usually incorrectly interpret the image

Results: Human Evaluation of Explanation

	 I mean, there are closers... and there are CLOSERS.	 He'll be back.	 My life's a joke.
GPT-3 175B (Finetuned)	A play on the word "closer" --- in baseball, a "closer" is a starting pitcher who is used in the last few innings of a game. But, here, the closer is the grim reaper, and so this is a reference to both a literal closer in baseball and a figurative closer in life.	A play on the phrase "back in a minute," which barbers say when they are going to get items from elsewhere in the shop during a haircut. Here, the barber was apparently so good that he was literally "back in a minute" by literally returning in a minute by flying in through the ceiling.	A reference to the phrase "life's a joke" --- this saying is usually uttered by people who feel like their lives are comedic catastrophes. But here, the person's life literally is a joke --- the saying takes on a second meaning: because, the man's life is literally a literal joke, which is funny. *[explanation misses the pen]*
GPT-4 (5-shot)	In baseball, a "closer" is a relief pitcher who is responsible for finishing the game and effectively "closing" the opponent's chances of winning. The presence of the Grim Reaper, who is the personification of death, implies that when he plays as the "closer," it's the ultimate form of closing as no one can escape death. The spectators are making a comment on how having the Grim Reaper play as a closer is an extreme and unbeatable choice.	The barber's chair has been launched through the ceiling because of the spring in place of it, creating an unexpected and destructive event. The customer's calm statement "He'll be back" is funny because it contrasts with the dramatic situation, suggesting they are used to bizarre occurrences at this barbershop, or that the barber is going to just come back from this incident unfazed	The giant pen suggests that the characters in the scene are being drawn or controlled by someone else, much like comic strip characters. The caption, "My life's a joke," is a play on the idea that their existence is a humorous creation someone else has devised, and they have no real control over their lives or the situations they find themselves in.
Human Reference	A closer in baseball is a type of pitcher who is good at finishing games, i.e., playing in later innings. But here, death is literally playing baseball as represented by the grim reaper. The joke is funny because it frames death as a strong "closer" --- death is also quite effective at finishing lives.	Usually, if a customer leaves an establishment, it is because they are dissatisfied with the service; employees will sometimes say "they'll be back" in confidence that the customer will realize they are in the wrong and that they need the service. But here, the customer has exited the barber shop involuntarily by being launched by the spring into the ceiling; the other barber rightly says that "he'll be back", but it's for gravitational reasons, i.e., they will literally be back when gravity causes them to fall back into the barber shop from above.	A meta joke about cartoons: the two people in the cartoon have become aware that they are in a cartoon because they spotted the pen that was drawing them. Sometimes, people claim their life is a joke when something so terrible or unlikely has happened that it must be the universe playing a joke on them; but here, their lives are literal jokes, because they are cartoons, and cartoons are often jokes.

Results: Error Analysis for Matching

Q8: Are some contests more difficult than others?

- **Answer: Yes.**
- Forming a contest-by-correctness (704-by-2) contingency table, aggregating over the 3-6 matching instances for each contest, and find that errors are clustered according to contest.

	Contest 1	Contest 2	Contest 3	Contest 4	Contest 5
Correct					
Wrong					

($p < .05$ for both CLIP and GPT-3) ➔ there is a difference.

- However, when we attempt to identify consistent factors that predict contest difficulty using various visual/linguistic predictors, **we find hard vs. easy difficult to predict a priori**; our best classifiers perform only slightly above random. We will distribute the hard vs. easy contest lists as a resource for future work.

Conclusion

- Our matching/quality ranking models could help entrants receive quantitative feedback on the relevance/predicted quality of their submissions
- The annotated corpus+explanations we introduce could be repurposed for generation.
- Finally, 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 in an reinforcement learning from human feedback (RLHF) loop

Thanks!