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**[†‡]
[†] The Allen Institute for AI [◇] University of Utah [○] Cornell University [•] OpenAI
[‡] University of Washington [▲] Air Mail and Cartoon Collections
jackh@allenai.org ana.marasovic@utah.edu jenah@allenai.org llee@cs.cornell.edu
{jzda,rowanz}@cs.washington.edu bob@bobmankoff.com yejin@cs.washington.edu

Abstract

Large neural networks can now generate jokes, but do they really "understand" humor? We challenge AI models with three tasks derived from the New Yorker Cartoon Caption Contest: matching a joke to a cartoon, identifying a winning caption, and explaining why a winning caption is funny. These tasks encapsulate progressively more sophisticated aspects of "understanding" a cartoon; key elements are the complex, often surprising relationships between images and captions and the frequent inclusion of indirect and playful allusions to human experience and culture. We investigate both multimodal and language-only models: the former are challenged with the cartoon images directly, while the latter are given multifaceted descriptions of the visual scene to simulate human-level visual understanding. We find that both types of models struggle at all three tasks. For example, our best multimodal models fall 30 accuracy points behind human performance on the matching task, and, even when provided ground-truth visual scene descriptors, human-authored explanations are preferred head-to-head over the best machine-authored ones (few-shot GPT-4) in more than 2/3 of cases. We release models, code, leaderboard, and corpus, which includes newly-gathered annotations describing the image's locations/entities, what's unusual in the scene, and an explanation of the joke.

1 Introduction

Humor can be dissected, as a frog can, but the thing dies in the process and the innards are discouraging to any but the pure scientific mind.
– White, E. B. (1941)

Each week, *The New Yorker* publishes a uncaptioned cartoon image, inviting readers to submit their funniest English-language caption for it. Editors choose three finalists from sometimes thousands of submissions; then, readers vote to pick the final winner. We develop a suite of three progressively harder tasks built around this contest to test how well AI models "understand" humor across vision and language: 1) matching jokes to cartoons, 2) identifying a winning caption, and 3) generating an explanation of why an image/caption combination is funny.

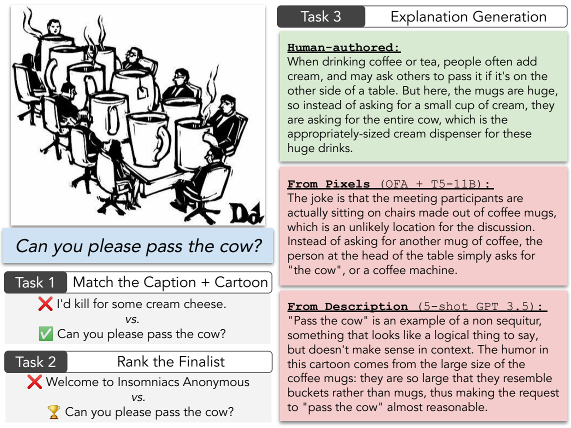


Figure 1: We formulate three tasks using over a decade of New Yorker caption contests: models must 1) recognize a caption written about a cartoon (vs. options that were not); 2) evaluate that caption's "quality" by scoring it more highly than a non-finalist/non-winner from the same contest; and 3) explain why the joke is funny. (Cartoon by Drew Dernavich, winning caption by Bennett Ellenbogen).

These tasks are difficult because the connection between a winning caption and image can be quite subtle, and the caption can make playful allusions to human experience, culture, and imagination. Consider the image and winning caption "Can you please pass the cow?" in Figure 1. Unlike literal image captions such as in MSCOCO (Lin et al., 2014), here, the caption's relation to the image is indirect:[1] the size of the mugs must first be recognized as unusual, and then, the caption invokes

[1] The (relatable) experience of "not getting" a New Yorker cartoon often results from inability to identify the image/text relationship.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 688–714
July 9-14, 2023 ©2023 Association for Computational Linguistics

Figure 2: Instances of our three tasks. *Matching* requires models to select the finalist caption for the given cartoon from among distractors that were finalists, but for *other* contests. *Quality ranking* requires models to differentiate a finalist from a non-finalist, both written for the given cartoon. *Explanation* requires models to generate free-text explanations of how a high-quality caption relates to the cartoon. Cartoons by Robert Mankoff and Mick Stevens.

an association between a large mug and a large amount of cream/milk — perhaps a whole cow's worth. Further, matching a caption to an image is not sufficient: non-finalist entries (e.g., "...Insomniacs Anonymous" in Figure 1) also match the image, but something else makes one seem funnier than the other. Finally, even if a model can accurately identify winning submissions, we would like it to also be able to explain *why* a particular highly rated/relevant caption is funny.

We cover our three tasks in two settings: in the *from pixels* setting, models are given access only to the cartoon images at test time, and must perform computer vision; in the *from description* setting, we allow models access to a newly-collected, human-authored corpus of cartoon descriptions, thus simulating access to a human-level computer-vision system — or, alternately, facilitating benchmarking of models that don't have a built-in image-processing component. The annotations we collect and release are rich and multifaceted: they describe the image overall and its locations and entities, what's unusual about the image, and an explanation of the joke. We view this effort as a significant contribution of our work.

Our results reveal a gap between AI and human-level humor "understanding." In the *from pixels* setting, our best multimodal model (fine-tuned CLIP ViT-L/14 (Radford et al., 2021)) achieves 62% accuracy on a 5-way multiple choice task, but humans achieve 94% in the same setting. Even with significant manual annotation of the cartoons in the *from description* setting (and despite significant improvements in language modeling performance since this work's submission[2]) large language models still fall short: human explanations are still preferred in more than two-thirds of cases compared to our best explanation model, 5-shot GPT-4.

We release our challenging NLP/vision benchmarks,[3] annotations, models, leaderboard, and code at https://capcon.dev/. Beyond AI research, we also hope that our work will spur progress in human-AI collaboration tools for cartoonists, contest entrants, and beyond (see Appendix G for AI-generated captions).

2 Datasets and Task Setups

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; and (4) for some contests, quality estimates for each submission collected via crowdsourcing.[4]

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), starting from #508.[5] Crowd ratings in this corpus are gath-

[2]GPT-3 (Brown et al., 2020) was the most performant in Jan. 2023 when this work was submitted, but we have since updated our results.

[3]Our data may contain offensive jokes. We manually removed a handful of cases we observed to target specific protected classes. We do not endorse the jokes in the corpus, but rather, view them as interesting objects of study.

[4]We regret that The New Yorker does not currently have an alliterative-paragraph contest.

[5]We manually corrected some errors in the corpus.

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

Table 1: Basic size statistics for our three tasks. We extend Shahaf et al. (2015); Radev et al. (2016); Jain et al. (2020) by (a) proposing matching, quality ranking, and explanation tasks; (b) providing new, dense annotations for each cartoon (see Figure 3); (c) authoring a set of 651 joke explanations.

ered via the NEXT platform (Jamieson et al., 2015; Tanczos et al., 2017), where 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). We also sample three additional top captions that aren't editorial picks to serve as additional "finalists."

The second corpus, due to 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), but no crowd ratings. We remove by hand 55 contests whose images' resolutions are too low, and identify 80 low resolution (but usable) cases, taking special care when annotating this set (§2.2).

2.1 Task Setups

We pose three tasks. Matching and explanation are novel, whereas quality ranking extends the formulations introduced in Shahaf et al. (2015); Radev et al. (2016).

Matching. *Can a model recognize when a caption is appropriate for a given cartoon?* Five choices are given, only one of which truly corresponds. For the example in Figure 1, we supply the following possibilities:

(a) *O.K. I'm at the window. To the right? Your right or my right?*
(b) *I'd kill for some cream cheese.*
(c) *Bob just came directly from work.*
(d) **Can you please pass the cow?**
(e) *They only allow one carry-on.*

The correct caption is a finalist for the cartoon. Negative choices are randomly selected finalists from other contests, and as a result, are great captions for some *other* contest's image.[6] In some cases, matching depicted objects to their textual references may suffice, but in other cases, the relationship is more indirect. For example, Figure 2 (top) contains a subtle reference to Jane Goodall, thus requiring external knowledge; Figure 2 (bottom) relies on a stereotype of pharmaceutical companies being untrustworthy, hence requiring reasoning beyond the literal text.

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. As preprocessing, we run one round of text-only filtering to discard submissions that are easily identifiable as low quality, and also perform semantic deduplication; more details in Appendix C. Here is the end result for Figure 1:

(a) **Can you please pass the cow?**
(b) Welcome to Insomniacs Anonymous.

Which caption a particular individual prefers can be a matter of personal taste; but there is a general preference among our human annotators for the true finalist (see §3).

Explanation. *Can a model generate as good an explanation as a human for why a caption-and-image combination is funny?* Free-form explanations of why captions are funny/appropriate for their corresponding image were written by an author of this paper.[7] The rough annotation guidance was: "In a few sentences, explain the joke as if to a friend who doesn't 'get it' yet." Starting from a random finalist for each contest, 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). We consider a model to succeed at this task if human judges, presented with (unlabeled) pairs of author/machine-generated explanations, do not show a preference for the author-generated ones.

Evaluation metrics. For matching and quality ranking, we evaluate using accuracy. For quality ranking, we report *NYAcc* — the average accuracy over instances where the finalist was an official New Yorker finalist — and *CrowdAcc*, where the

[6] Distractors are balanced across instances so that a model that only examines the answer choices cannot achieve better than chance accuracy.

[7] Several attempts to solicit explanations from crowdworkers were not satisfactory; similarly unsuccessful were prompting experiments with GPT-3 inspired by Wiegreffe et al. (2022); Marasović et al. (2022) — too few of the sampled explanations were correct to bootstrap a corpus.

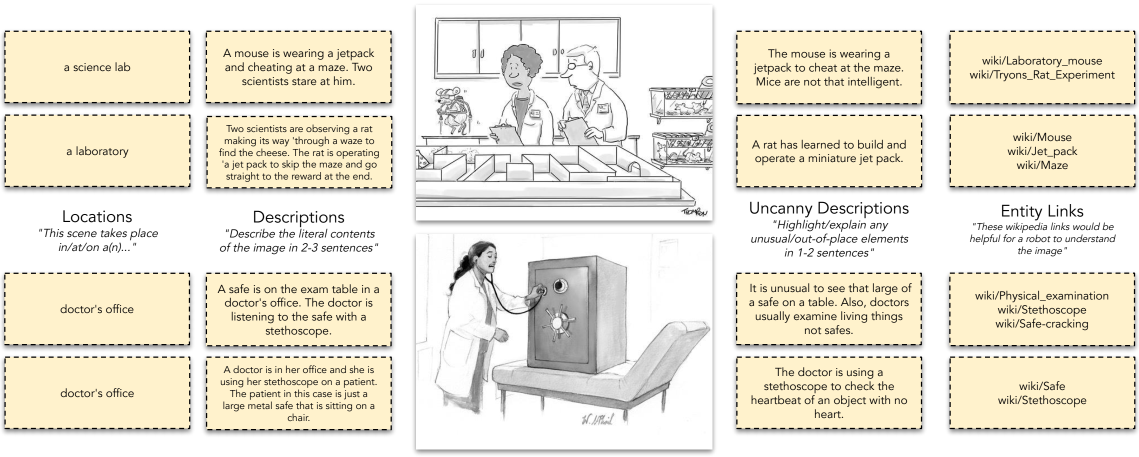


Figure 3: For each of 704 cartoons, we gather several types of annotations from human participants, including locations, descriptions, descriptions of uncanniness, and relevant entities in the form of English Wikipedia links. Annotations shown are true random samples from the corpus. Cartoons by Mark Thompson and Will McPhail.

"finalist" caption was selected by the crowd as high quality. These two measures allow us to account for different audience tastes. For explanation, we conduct pairwise human evaluations to test several hypotheses detailed in §3.2. To complement these human evaluations, we also report in Appendix E automatic metrics that take into account the human-written reference: (a) BLEU-4 (Papineni et al., 2002) using Post (2018)+ROUGE-L (Lin, 2004); and (b) word-level perplexity.

From Pixels + From Description. We consider two experimental settings. In **From Pixels (FP)**, a vision+language model undertakes image processing, i.e., at test time, the only contest information available is the image itself. In the second setting, which we call **From Description (FD)**, we factor out visual processing by providing the model with human written annotations, described in §2.2. FD models thus simulate access to a human-level computer-vision system.

2.2 Annotation of cartoons.

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. For each cartoon, we gather:

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

A random sample of annotations is shown in Figure 3. We used Amazon Mechanical Turk, and paid crowdworkers a minimum of $15/hr. Low-resolution images involved special treatment: 1) we offered additional pay to crowdworkers; and 2) at least one of the annotations is conducted by an author of this work using the same HIT interface. Details including qualification rounds, screenshots of the HITs, etc. are given in Appendix A.

3 Experiments

We split the 704 cartoons into 5 cross-validation splits such that entire contests are held out at test time. Task construction details are in Appendix C; modeling details (e.g., hyperparameter sweeps, task formatting) are in Appendix B.

From Pixels (FP) Models

We explore two vision+language models.

CLIP. We fine-tune CLIP `ViT-L/14@366px` (Radford et al., 2021) (428M parameters), which consists of a text Transformer (Vaswani et al., 2017) and a vision Transformer (Dosovitskiy et al., 2021) pretrained to align images/captions in the WebImageText corpus (400M pairs). 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`

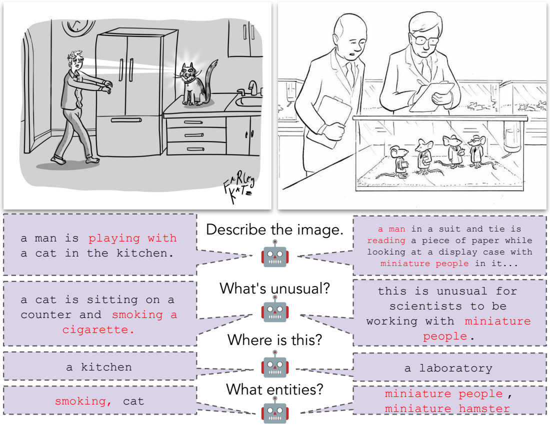


Figure 4: Example predictions by fine-tuned OFA-Huge from images. The model recognizes many objects/actions/locations, but makes some mistakes (indicated in red): for the left image, for example, it falsely indicates that the cat is smoking, and, on the right, that the mice are small people or hamsters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

`winning caption`. CLIP isn't generative, so we can't use it for explanation.

OFA → LM. We use OFA Huge (930M parameters) (Wang et al., 2022), a seq2seq model that supports image/text inputs/outputs; it is pretrained on a variety of vision+language tasks. We fine-tune on the New Yorker corpus by training it to map from (cartoon, prompt) → descriptions for the four types of annotations described in §2.2; see Figure 4 for example predictions. We organize the OFA-predicted outputs in the same format as the human-authored descriptions in our From Description (FD) models detailed below (except the inputs are the outputs of OFA), and pass the result to a language model:[8] this composition can be considered a Socratic Model (Zeng et al., 2022).

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 (Raffel et al., 2020); these encoder-decoder transformer models have 770M and 11.3B parameters respectively. For explanation, we sample with temperature 1.0 and nucleus sampling with p=.95 (Holtzman et al., 2020).

GPT-3, GPT-3.5, GPT-4. We use these three OpenAI models as both zero-shot and few-shot models. We provide the models with a description of the task, and, for the few-shot case, 5 random labelled in-context examples. Specifically, for GPT-3 we use `text-davinci-002` (175B) (Brown et al., 2020), and for GPT-3.5/GPT-4, we use the May 12, 2023 versions (OpenAI, 2023). For GPT-3, we also consider a fine-tuned version (which is unavailable for GPT3.5/GPT-4).[9] For zero-shot GPT-3.5/GPT-4, early experiments revealed that prompting models to "think" step-by-step with chain-of-thought (CoT) was helpful (Wei et al., 2022; Kojima et al., 2022). See §B.6 for GPT-3 details, and §B.7 for GPT-3.5/GPT-4 details.

Baselines

Caption Only. In addition to a **Random**-guess baseline, we fine-tune T5-11B given just the caption, i.e., without knowledge of the cartoon (Trichelair et al., 2019; Poliak et al., 2018).

Human performance estimates. Three people (two authors and one person familiar with the project) each attempted 100 randomly sampled instances from both the matching and quality ranking tasks.[10] It is important to note that *human performance is not an upper bound for model performance on matching and quality ranking* because labels are not generated by a single human and tastes can vary; it can (and does, see §3.1) happen that a machine might be able to reconstruct New Yorker editor preferences more reliably than an untrained human. Annotators were given access to the images, but not the descriptions (akin to the FP setting).

Hardware+software details.

T5, CLIP, and OFA were trained using 8 A100 GPUs in pytorch (Paszke et al., 2019). We use the Transformers (Wolf et al., 2020) implementation of T5: T5-11B was trained with deepspeed (Rasley

[8] We found that fine-tuning OFA directly was less effective.

[9] https://beta.openai.com/docs/guides/fine-tuning; for explanation, we use the default settings; for multiple choice, we set `prompt_loss_weight` to zero. The validation set is not used by the API for early stopping, so we concatenate it with the training set and perform no validation.

[10] Matching instances were sampled such that there were no repeated options, i.e., annotators couldn't use process of elimination across instances. 595 total responses were collected.

		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

Table 2: Prediction results for the matching and quality ranking tasks: averages over 5 cross-validation splits. <u>Underlined</u> results are the best model in the *From Pixels* (FP) setting, where at test time, models only have access to the cartoon images. **Bold** results are best in the *From Description* (FD) setting, where at test time, models have access to human-authored descriptions of the cartoons. Appendix D presents these results visually. Right: sample predictions by CLIP (finetuned), GPT-4 (5-shot), and the caption-only baseline over a matching/ranking instance. Cartoon by Joe Dator.

et al., 2020); T5-Large and CLIP were trained with Accelerate.[11]

3.1 Matching and quality ranking results

Table 2 contains the results. Among the *from description* models, GPT-4 (5-shot) generally performs best, e.g., achieving 84.5% accuracy on matching. It (and fine-tuned GPT-3) also perform *better* at predicting New Yorker editor selections than our three humans (column NYAcc: GPT-3 69.8 vs. Human estimate, 64.6), but underperform at predicting crowd selections (CrowdAcc column: GPT-4 73.3 vs. 83.7).[12] We also see that our *from pixels* models leave significant headroom compared to the human performance estimates.

Other observations include: 1) both *from pixels* and *from description* models mostly outperform the Caption Only baseline (even for smaller model sizes), suggesting that the models are truly using feature interactions between cartoons/captions to improve their predictive accuracy; 2) fine-tuning CLIP tends to do best for matching in the *from pixels* setting, but OFA+T5-11B is competitive for quality ranking (and supports generation, see §3.2); and 3) the performance difference between T5 vs. OFA→T5 exemplifies the effect of suboptimal visual recognition when shifting from the *from pixels* setting to the *from description* setting. Finally, while performance drops are incurred universally for zero-shot models, pointing towards the utility of the new annotated corpus we are releasing (§2.2), GPT-4's zero-shot chain-of-thought incurs a smaller performance drop compared to other zero-shot models; see §B.7 for a sample chain-of-thought.

3.2 Human evaluation of explanation.

We gather judgments from 3 crowd-workers per test instance by asking them which of a pair of explanations they prefer, and take a majority vote to determine a winner. Results and annotator agreement are in Table 3, and samples of GPT-3, GPT-4, and human joke explanations are in Figure 5. Our evaluations address seven questions:

Q1: Do models utilize the image context of the caption to generate better explanations? *Test: T5-11B vs. Caption-only T5-11B.* 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.

Q2: Is computer vision a bottleneck for top-quality explanation generation? *Test: T5-11B (in the FD setting) vs. OFA → T5-11B.* Answer: **Yes.** Compared to the same model trained with access to human written descriptions available at test

[11] https://huggingface.co/docs/accelerate

[12] Also, crowd selectors greatly outnumber New Yorker editors, so crowd rankings may be a more dependable target, statistically speaking.

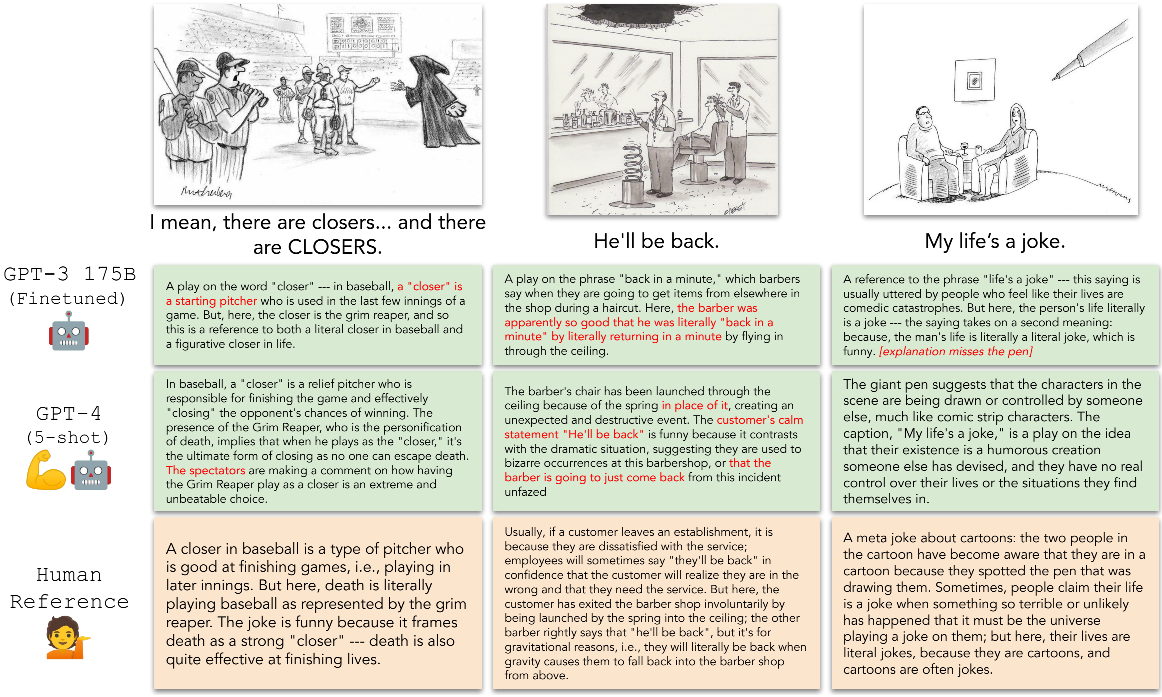


Figure 5: A random sample of caption explanations generated by a fine-tuned version of GPT-3, GPT-4 with 5 shots, and human-written references. Errors are highlighted in red. Machine-authored generations are often on the right track, but frequently contain mistakes, e.g., by referring to a closing pitcher as a starter (GPT-3, left image) or suggesting that a barber, rather than a customer, was launched (GPT-4, middle image). Cartoons by Mort Gerberg, Tom Cheney, and Mick Stevens.

	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.

time (i.e., the *from description* setting), the model trained with access only to OFA-predictions loses in 74.6% of cases.

Q3: Do bigger T5 models generate better explanations? *Test: T5-11B vs. T5-Large.* Answer: **Yes.** T5-11B with access to the same information at test time as T5-Large (770M) is preferred in 68.5% of cases.

Q4: Does fine-tuning an LLM model help vs. in-context learning for explanation generation? *Test: FT-GPT3 vs. In context (=5-shot) GPT3.* Answer: **Not really.** In contrast to the multiple choice tasks, 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).[13] 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.

Q5: Do supervised explanations help, even with GPT-4? *Test: 5-shot GPT-4 vs. Zero-shot 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.

Q6: Does GPT-4 outperform GPT-3? *Test: 5-shot GPT-4 vs. 5-shot 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

[13] A disparity not mirrored in the word-overlap metrics BLEU-4 and Rouge-L, also reported in Appendix E.

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.

Q7: Does our best model, GPT-4, explain jokes as well as humans? *Test: Human vs. Few-shot GPT-4.* Answer: **No.** Human-written explanations are preferred by annotators in 68% of pairwise cases.[14] 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 cavepeople 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).

3.3 Error Analysis for Matching

We conduct an error analysis of a performant *from pixels* model (`CLIP ViT-L/14@336px` finetuned), and a performant *from description* model (GPT3-175B finetuned). We concatenate the test set predictions over the 5 cross validation splits, and ask:

Q8: Are some contests more difficult than others? Answer: **Yes.** *Details:* We conduct a χ^2 test by 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 ($p < .05$ for both CLIP and GPT-3).[15] There's a moderate Spearman correlation between the per-contest accuracy between the models ($\rho = .28, p \ll .001$), but (as a null hypothesis) only a slight correlation between contest date and difficulty for either (later contests easier, GPT3/CLIP $\rho = .07/.08, p = .08/.05$). When the models' predictions agree, they are correct 87% of the time. When GPT-3 is wrong, CLIP is right only 38% of the time; under the null hypothesis that their errors are uncorrelated, CLIP's accuracy would be 62% ($p \ll .001$ errors are uncorrelated, permutation test). 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.

4 Related Work

Humor. Raskin (1979) and Attardo (2008) highlight three "great families" of theories of the roots of humor: 1) *hostility,* claims of superiority over someone or something (Gruner, 1978; Billig, 2005); 2) *release* of a constraint (Freud, 1905; Fry, 1963; Mindess, 1971) and 3) *incongruity,* (sometimes "incongruity-resolution"; Mulder and Nijholt, 2002) the introduction (and subsequent resolution) of generally incompatible contexts (Schopenhauer, 1818; Shultz, 1976). Shahaf et al. (2015) note that most New Yorker caption contest cartoons involve incongruous situations.

NLP + The Caption Contest. King et al. (2013), Shahaf et al. (2015), and Radev et al. (2016) analyze 5, 16, and 50 New Yorker Caption Contests, respectively. Best-performing features for identifying the funniest among a set of caption choices include: perplexity, match to image setting and uncanniness description, readability, proper nouns (Shahaf et al., 2015), overlap with WordNet's (Fellbaum, 1998) "person" and "relative" synsets, lexical centrality among submissions (Radev et al., 2016, inspired by Mihalcea and Pulman (2009)), and sentiment (both papers). Our "location" and "uncanny description" annotations are direct analogs of the "context" and "anomaly" tags of Shahaf et al. (2015), and our data incorporates that generously released by the previous researchers. Our extensions are (a) the addition of two novel tasks; (b) using new data/resources/models to curate ranking pairs (see

[14]For a similar, earlier set of experiments with FT-GPT-3 vs. human, human was preferred in 87.8% of pairwise cases.

[15]Similar χ^2 tests find no evidence of correlation between correctness and (a) cross-validation split (5-by-2 table; p=.84/.14 for GPT3/CLIP); or (b) which captions are randomly assigned as negative choices (2646-by-2 table, p=.92/.79 for GPT3/CLIP).

§2); and (c) evaluating two distinct audience preferences: New Yorker editors vs. "the crowd". Appendix H highlights efforts beyond the scope of peer reviewed AI venues, e.g., blog posts.

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.

Multimodal and computational humor. Chandrasekaran et al. (2016) explore humor recognition in images, and Castro et al. (2019); Hasan et al. (2019); Patro et al. (2021); Hasan et al. (2021) explore laughter prediction in TED-talks/sitcoms. Tsakona (2009); Fallianda et al. (2018) study political cartoons. Chakrabarty et al. (2022) recently proposed a version of NLI for figurative language, which can be humorous. Some work has tried to detect whether a sentence is humorous or not (Blinov et al., 2019; Annamoradnejad and Zoghi, 2020). More difficult to evaluate (Valitutti, 2011) are setups where the goal is to automatically generate humorous content in various contexts (Binsted and Ritchie, 1994; Stock and Strapparava, 2003; Mihalcea and Strapparava, 2005, 2006; Wang and Wen, 2015; Chandrasekaran et al., 2018; Yoshida et al., 2018; Sundaram, 2018; Shimomoto et al., 2019); a survey is provided by Amin and Burghardt (2020).

Explaining humor. In the taxonomy of Tan (2022), joke explanations are most related to proximal mechanisms: "This type of explanation attempts to provide the mechanism behind the predicted label, i.e., how to infer the label from the text", or efficient cause a la Aristotle (Lombrozo, 2006). Chowdhery et al. (2022) undertake a qualitative exploration of (non-visual) joke explanations.

5 Conclusion

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. However, the partial capacity of today's AI is still substantial, and may be sufficient for models to serve as creative collaborators, e.g., as brainstorming assistants for humorists/cartoonists. Specifically: 1) our matching/quality ranking models could help entrants receive quantitative feedback on the relevance/predicted quality of their submissions, and 2) the annotated corpus+explanations we introduce could be repurposed for generation (we explore generation of novel cartoons/captions in Appendix G). 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.

A last remark. We cannot claim to know whether the human-machine 'humor understanding gap' will be closed sooner or later.[16] But we encourage other researchers to have as much fun with the topic as we did!

6 Limitations

The New Yorker Cartoon Caption Contest represents a narrow slice of humor, deriving from a particular language, region, history, culture, style, and set of conventions. Hence, the results of this study do not represent or cover all types of humor.

Our framing of the quality ranking task could be interpreted as seemingly prescriptive (i.e., that joke A is "objectively" better than joke B), but New Yorker editorial selections should not be taken as ground truth for funniness; disagreement about what is funny is expected and valid. 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.

Finally, the explanations in our annotated corpus were largely written by a single author of this paper. While a larger pool of the crowdworkers judged these explanations to be of higher quality in comparison to machine generations, future work would be well-suited to compare the person-to-person variance in explaining why particular jokes are funny.

[16]Or never. Is never good for you?

7 Acknowledgements

We thank the cartoonists and contest entrants for their wonderful efforts! We additionally thank our crowd annotators for their diligent work, Lisa Watkins for contributing to the human performance estimates, and the anonymous reviewers for their constructive comments. This work was funded in part by DARPA MCS through NIWC Pacific (N66001-19-2-4031), the Allen Institute for AI, and a Google Focused Research Award. Jack Hessel conducted initial work while at Cornell University. Ana Marasović conducted this work while at The Allen Institute for AI. Rowan Zellers conducted this work while at University of Washington.

References

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*.

Issa Annamoradnejad and Gohar Zoghi. 2020. ColBERT: Using BERT sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*.

Salvatore Attardo. 2008. A primer for the linguistics of humor. *The primer of humor research*, 8:101–55.

Michael Billig. 2005. *Laughter and ridicule: Towards a social critique of humour*. Sage.

Kim Binsted and Graeme Ritchie. 1994. An implemented model of punning riddles. In *AAAI*.

Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. In *ACL*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *NeurIPS*.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *ACL*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: figurative language understanding and textual explanations. In *EMNLP*.

Arjun Chandrasekaran, Devi Parikh, and Mohit Bansal. 2018. Punny captions: Witty wordplay in image descriptions. In *NAACL*.

Arjun Chandrasekaran, Ashwin K. Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2016. We are humor beings: Understanding and predicting visual humor. In *CVPR*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Fallianda, Rani Yuni Astiti, and Zulvy Alivia Hanim. 2018. Analyzing humor in newspaper comic strips using verbal-visual analysis. *Lingua Cultura*, 12(4):383–388.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Sigmund Freud. 1905. *Jokes and their Relation to the Unconscious*, volume 8 of *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth, London.

William F. Fry. 1963. *Sweet madness: A study of humor*. Pacific Books, Palo Alto.

Charles R. Gruner. 1978. *Understanding laughter: The workings of wit & humor*. Nelson-Hall, Chicago.

Kilem Gwet. 2014. *Handbook of Inter-Rater reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*, 4th edition edition. Advanced Analytics, LLC.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *AAAI*.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: a multimodal language dataset for understanding humor. In *EMNLP*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Lalit Jain, Kevin Jamieson, Robert Mankoff, Robert Nowak, and Scott Sievert. 2020. The New Yorker cartoon caption contest dataset.

Kevin G. Jamieson, Lalit Jain, Chris Fernandez, Nicholas J. Glattard, and Rob Nowak. 2015. NEXT: A system for real-world development, evaluation, and application of active learning. In *NeurIPS*.

Ben King, Rahul Jha, Dragomir Radev, and Robert Mankoff. 2013. Random walk factoid annotation for collective discourse. In *ACL*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E. Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of NAACL*.

Rada Mihalcea and Stephen Pulman. 2009. Characterizing humour: An exploration of features in humorous texts. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, page 337–347, Berlin, Heidelberg. Springer-Verlag.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *EMNLP*.

Rada Mihalcea and Carlo Strapparava. 2006. Technologies that make you smile: Adding humor to text-based applications. *IEEE Intelligent Systems*, 21(5):33–39.

Harvey Mindess. 1971. *Laughter and Liberation*. Nash.

Pamela Mishkin, Matt Daniels, Russell Goldenberg, Ilia Blinderman, and James Yu. 2022. The pudding caption contest experiments. https://pudding.cool/projects/caption-contest/. Accessed: 2022-04-01.

Matthijs P. Mulder and Antinus Nijholt. 2002. *Humour research: State of the art*. Centre for Telematics and Information Technology, University of Twente.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh, Hunar Singh, and Vinay P. Namboodiri. 2021. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *WACV*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In **SEM*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *WMT*.

Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. 2016. Humor in collective discourse: Unsupervised funniness detection in the New Yorker cartoon caption contest. In *LREC*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

Victor Raskin. 1979. Semantic mechanisms of humor. In *Annual Meeting of the Berkeley Linguistics Society*, volume 5, pages 325–335.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*.

Arthur Schopenhauer. 1818. *The world as will and idea*, volume 1.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *KDD*.

Erica K Shimomoto, Lincon S Souza, Bernardo B Gatto, and Kazuhiro Fukui. 2019. News2meme: An automatic content generator from news based on word subspaces from text and image. In *Conference on Machine Vision Applications*.

Thomas R Shultz. 1976. *A cognitive-developmental analysis of humour.* Transaction Publishers.

Oliviero Stock and Carlo Strapparava. 2003. Getting serious about the development of computational humor. In *IJCAI*.

Rajesh Shanmuga Sundaram. 2018. *Generation of Humorous Caption for Cartoon Images Using Deep Learning*. Ph.D. thesis, Texas A&M University-Commerce.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *NAACL*.

Ervin Tanczos, Robert Nowak, and Bob Mankoff. 2017. A KL-LUCB algorithm for large-scale crowdsourcing. In *NeurIPS*.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the Winograd schema challenge and SWAG. In *EMNLP*.

Villy Tsakona. 2009. Language and image interaction in cartoons: Towards a multimodal theory of humor. *Journal of Pragmatics*, 41(6):1171–1188.

Alessandro Valitutti. 2011. How many jokes are really funny? In *Human-Machine Interaction in Translation: Proceedings of the 8th International NLPCS Workshop*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*.

David Wallace. 2022. Lecture notes for MIT 2.00b toy product design: Innovation and associations.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.

William Yang Wang and Miaomiao Wen. 2015. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *NAACL*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

White, E. B. 1941. Preface. In E. B. White and Katherine S. White, editors, *A Subtreasury Of American Humor*, page xvii. The original version of this quote appeared as a preview in *The Saturday Review* (1941), credited to both Whites. But, the quote appears in the preface to *A Subtreasury* (1941) with authorship solely credited to E.B.. We thus credited the quote itself to E.B., and credited both E.B. and K.S. as editors of the anthology in which it appears in non-preview form.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *NAACL*.

Hannah Wilson. 2019. Project four - nobody knows you're a bot.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*.

Kota Yoshida, Munetaka Minoguchi, Kenichiro Wani, Akio Nakamura, and Hirokatsu Kataoka. 2018. Neural joking machine: Humorous image captioning. In *CVPR Language & Vision Workshop*.

Michael Zelenko and Frank Bi. 2015. On the internet, nobody knows you're a machine.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

A Crowdworking Details

We use three Mechanical Turk interfaces to gather data. These are:

1. *Cartoon description* (Figure 6). We ran this HIT 3 times per cartoon.
2. *Cartoon wikipedia links* (Figure 7). We ran this HIT 2 times per cartoon.
3. *Pairwise explanations* (Figure 8). We ran this HIT 2.7K times to facilitate the comparisons in §3.2

Qualification+training rounds. To ensure our set of crowdworkers were properly trained for the annotations, we ran two types of qualification rounds: one for the description/link HITs, and one for the pairwise explanation HITs.

For the description/link HITs, our qualification round was based off an earlier and more involved HIT that involved a joint setup where, for 3 cartoons, users described cartoons, highlighted image regions, explained jokes, etc. We allowed users from $\{$AU, CA, NZ, GB, US$\}$ with 10K prior approved HITs and a minimum acceptance rate of 97% on their previous HITs to participate. Some of the cartoons and captions contain mature themes; we provided the recommended disclaimer for this and other HITs: “WARNING: This HIT may contain adult content. Worker discretion is advised.” We manually graded the responses of 30 annotators in a qualification round, and qualified 21. Through a mix of the older, more involved HITs and the streamlined HIT in Figure 6, which is a pared-down version of the original HIT without captions, we gathered descriptions of the cartoons. We also gathered the locations/Wikipedia entity links from the qualified annotators. These annotations were gathered in mid-to-late 2021.

About 9 months later, we conducted a second set of Mechanical Turk studies for pairwise judgment evaluations for explanation. A second qualification round was run, in which we asked annotators to rate the quality of several joke explanations which we manually selected to be good/bad across various desirable axes. We qualified 29 out of 51 annotators who attempted the HIT via manual inspection of their responses. This set of annotators were given access to the final pairwise-judgment HITs.

Crowdworking studies of standard computer vision corpora (involving no personal disclosures) are not required by our IRB to be reviewed by them. While the authors of this work are not lawyers and this is not legal advice, this opinion is based on United States federal regulation 45 CFR 46, under which this study qualifies and as exempt. We hashed crowdworker IDs in the public release so annotations cannot be back-traced to individual workers.

B Additional Experimental Details

B.1 From Description details

For each cartoon, we have multiple annotations of each type, as detailed in §2.2. During training, we utilize all location/description/uncanny description/sets of links, but at test time, we randomly sample a single set of these four annotation types such that inference requires only a single forward pass. For fair comparison, the randomly sampled description available at test time is held constant between all methods.

More detail about how we managed multiple annotations: because we have 2 locations $\times$ 3 descriptions $\times$ 3 uncanny descriptions $\times$ 2 entity links, there are potentially 36 possible combinations we could use to form a *from description* instance for each cartoon. However: tuples are constructed at the annotator level to account for potential dependencies between annotation types: because descriptions/uncanny descriptions were were collected in the same HIT, the uncanny description may reference entities from the description because they were authored at the same time by the same annotator in sequence. Similarly, the (locations, links) were collected in the same HIT. So, we instead consider all six possible tuples holding author constant between HITs, i.e., 3 (description, uncanny description) $\times$ 2 (location, link). For test time, select a single random valid tuple of annotations for evaluation, which is fixed for all comparisons.

B.2 CLIP

For fine-tuning results, we do linear warmup for 200 steps and conduct a small learning rate search on the validation set for each cross-validation split independently between $\{5e\text{-}5, 1e\text{-}5, 5e\text{-}6\}$, keeping batch size fixed at 32. To keep the entire cartoon in the 336px square input, we resize and pad. At training time, we perform data augmentations on the image, including: random horizontal flipping, random color jittering, and random grayscaling.

NOTE: The instructions have probably changed since the last time you did this HIT!
While this HIT is similar, please take a moment to familiarize yourself with the modifications

Your task is to analyze a given image.

There are three parts

First, **describe the literal contents of the image image** by writing a *2-3 sentence summary*. Consider focusing on:

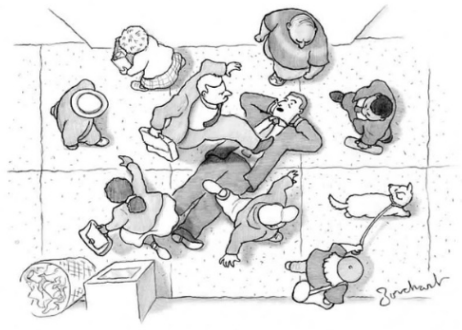
- Where is the scene taking place?
- Who/what is in the scene? What are they doing?
- What objects and actions are being depicted?
- Is anyone particularly happy/unhappy/mad/etc.?

There's no need to be too formal, but please do your best to write full, grammatical sentences (see the examples below).

Second, these images may depict interesting/unusual situations. **Highlight these uncanny/unusual elements**, by giving a 1-2 sentence explanation of why they are uncanny, *e.g., "[object/character/...] is unusual/uncanny/out-of-place because ..."*. Consider focusing on:

- Which objects, actions, entities, etc. are out-of-place and why?
- Are the actions any characters are undertaking strange?
- Do the characters have any unusual identifying characteristics?

Third, in a single sentence, please write **the question that you most want answered about the scene**, based on the image, your description, and your highlight of which parts are unusual/uncanny (see examples below). Optionally, you can include a second question that you would like answered if there are multiple uncanny elements to the scene.



Please describe the literal contents of the image in 2-3 sentences:

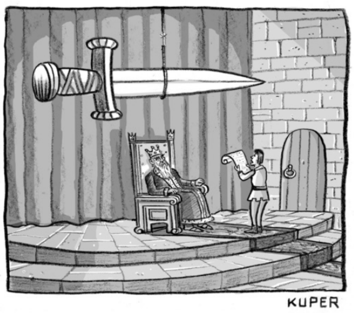
A man in a suit is lying down on a sidewalk in a busy city as pedestrians walk over him. The pedestrians seem to be frustrated and confused that the lying down man is blocking their way, while the man himself seems to be carefree.

Please highlight/explain any unusual/out-of-place elements in 1-2 sentences:

It's unusual that the man is lying in the middle of a sidewalk not only because this action is disruptive to other pedestrians, but also because he's in a nice suit that is likely to become dirty. Furthermore, his carefree expression indicates that, despite these downsides, he doesn't care and is in no rush to move.

In 1 sentence, which question would you most like answered about the scene?

Why is the man lying on the sidewalk?



Please describe the literal contents of the image in 2-3 sentences:

Please highlight/explain any unusual/out-of-place elements in 1-2 sentences:

In 1 sentence, which question would you most like answered about the scene?

(optional) In 1 sentence, which question would you also like answered about the scene?

Figure 6: Instructions, example, and interface for the Cartoon Description HIT. We gather, but do not use, the final "Which question?" annotation in our experiments.

NOTE: Please read the instructions, even if you've done a similar HIT before. The instructions have probably changed! Please take a moment to familiarize yourself with the modifications.

Your task is to provide context for a given image.

There are two parts to this task.

First, in a few words, you'll complete the sentence **This scene takes place in/at/on...** .
Examples of reasonable completions include:

- a bar
- a medieval castle
- a city street
- a dessing room

Next, you'll choose **at least 2 English Wikipedia links** (and up to 3) that could help a **robot** understand what is expected, and what is weird, about the provided scene. You'll need to use the search function provided here https://en.wikipedia.org/wiki/Main_Page to find the article URLs, and then copy-and-paste from your Internet browser's URL bar.
Examples of reasonable links include:

- if Paris were referenced: ***https://en.wikipedia.org/wiki/Paris***
- if President Obama was referenced, ***https://en.wikipedia.org/wiki/Barack_Obama***
- if the characters were eating Fondue, ***https://en.wikipedia.org/wiki/Fondue***
- if the scene is in space, ***https://en.wikipedia.org/wiki/Outer_space***
- if the characters were at a Lūʻau, ***https://en.wikipedia.org/wiki/Lūʻau***
- if the characters were fencing, ***https://en.wikipedia.org/wiki/Fencing***
- if the characters are drunk, ***https://en.wikipedia.org/wiki/Alcohol_intoxication***
- if the characters are from Sleeping Beauty, ***https://en.wikipedia.org/wiki/Sleeping_Beauty***
- if a character is psychic, ***https://en.wikipedia.org/wiki/Psychic***

Rules for links:

- **Provide valid links:** A valid English wikipedia link should begin with "https://en.wikipedia.org/wiki/" and have no section links. Don't provide *https://en.wikipedia.org/wiki/Sleeping_Beauty#Plot*, use *https://en.wikipedia.org/wiki/Sleeping_Beauty*.
- **Try not to provide the wikipedia article for the answer you gave in part 1:** If you wrote that the scene takes place in "a bar," try your best to provide links beyond the wikipedia article for "bar," unless it's particularly relevant, or there are no other options.
- **Provide specific/relevant links:** If the scene happens to contain people, don't just provide a link to the wikipedia page for "Person." Also, focus on relevant information, e.g., don't include links for "Shoe" if the person happens to be wearing a shoes, unless the shoe is relevant to the scene.
- **For proper nouns, specific is better:** linking "Sleeping Beauty" is better than "Fairy Tale"; linking "New York City" is better than "City"; linking "Barack Obama" is better than than "President"

This scene takes place in/at/on:

a bathroom

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

- Link 1 (required): **https://en.wikipedia.org/wiki/Cement_shoes**
- Link 2 (required): **https://en.wikipedia.org/wiki/Gangster**
- Link 3 (optional): **https://en.wikipedia.org/wiki/Shower**

Figure 7: Instructions and example for the Wikipedia links HIT.

In this HIT, you will be presented with a cartoon from the New Yorker, and a caption someone wrote about that cartoon. The captions will relate to the image in a clever/funny way.

You will also be presented with **two explanations** of the joke: these explanations may be written by humans or by machines. Your job is to **select the explanation that you think is the best one.**

Aside from fluency, grammaticality, etc., qualities of good explanations include:

- they offer a complete explanation of why the caption is funny;
- they reference appropriate external factors like real-world knowledge, etc.;
- they are not overly long or overly short;
- they are not overly repetitive/redundant;
- and, they don't contain false information, e.g., references to objects, people, etc. that are not in the image.

If both explanations are reasonable, choose the one you prefer. If both explanations are bad, do your best, and choose the one that is less wrong/makes more correct references. **Please take a moment to understand the captions/cartoons --- some can be hard to understand!** Choices for explanations will appear after a ten second delay.

Thanks again for your participation and your work!

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	● Explanation 2
"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.	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.

Figure 8: Instructions and interface for the pairwise explanation judgment HIT.

B.3 OFA

We use validation-set early stopping on cross-entropy loss, and fine-tune OFA separately for each cross-validation split. After fine-tuning, we select the top-1 prediction according to beam search (n=5). We finetune OFA Huge with a learning rate of $5e$-5, which was determined via a small grid search over the first cross-validation split. We use label-adjusted smoothed cross entropy loss as implemented by the OFA authors[17] with smoothing of 0.1. We train for a maximum of 7 epochs with a warmup ratio of 6%. For each image, we query for the four different types of annotations shown 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?"

In early experiments, instead of composing with a language model, we did attempt to fine-tune OFA directly for the explanation task. However, we found that the resulting perplexity (roughly 300) was significantly higher than for other fine-tuned models, with the errors difficult to diagnose.

B.4 T5-Large/T5-11B.

For T5-Large, we conduct a small, per-cross-validation split learning rate search between $\{1e\text{-}4, 1e\text{-}5, 5e\text{-}5\}$ and keep batch size fixed at 64. For T5-11B we use a fixed learning rate of $1e$-5 and a batch size of 64.

B.5 GPT-3 Zero Shot/In Context

We use GPT-3's `davinci-text-002` model for our main zero shot and in-context learning experiments. Examples of zero-shot prompts for all tasks are given in Figure 9. The in-context prompts are similar, except they contain 5 random samples from the training set. A full, randomly selected in-context prompt for the explanation generation task is given in Figure 10.

B.6 GPT-3 Fine-tuning

We use the OpenAI fine-tuning API to fine-tune `davinci`, a 175B parameter language model.[18] While the precise details of how the API works are not currently available (e.g., which parameters are updated, or which version of `davinci` is used), we use the same cross-validation setup as for the other models so that the results are comparable. The total fine-tuning cost is approximately (3 tasks) $\times$ (5 cross-val splits) $\times$ (40 dollars per fine-tune) = 600 dollars.

[17] https://github.com/OFA-Sys/OFA

[18] https://beta.openai.com/docs/guides/fine-tuning

B.7 GPT 3.5/GPT-4 Details

Between submitting this work and its acceptance, OpenAI released two new models, GPT-3.5 (sometimes called ChatGPT when accessed through the chat interface) and GPT-4; we updated our results to include these models. Figure 11 provides an example of a prompt/response in the new "Chat" API, which requires a more structured conversational prompt compared to the GPT-3 "Completion" API; this prompt includes a "system" prompt, which describes the desired behavior of the model, e.g., "You are CaptionContestGPT..." We sample with default hyperparameters in all cases. The cost of GPT 3.5 is an order of magnitude less than GPT-4. In total our GPT-4 queries cost on the order of \$4K.

C Task Construction Details

Identification of High Quality Captions. For each contest, our first step is to identify a set of high quality captions; these are involved in construction of instances for all three tasks. For cases where we have access to the three official New Yorker finalists, all are automatically added to the high quality set. Next, for cases where we have crowd ratings, we consider the top 5 crowd ranked captions according to the `mean` score provided by Jain et al. (2020). From these top 5, we select 3 diverse candidates among these using a semantic deduplication method: specifically, we compute the SBERT (Reimers and Gurevych, 2019) vector for each candidate using `paraphrase-MiniLM-L6-v2`, compute a hierarchical clustering of the candidates, and sample a single candidate from each cluster — the result is a set of candidates that is representative of all clusters. In total, there are 2.7K high quality captions across 704 contests. Each contest either has 3 high quality captions (coming from the official New Yorker finalists or, if those aren't available, highly crowd-rated options), or 6 (if both official finalists and crowd rated are available).

```
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:
```

```
In this task, you will see a description of an uncanny situation. Then, you will see two jokes that were written about the situation. One of the jokes is better than the other one. Pick which of the two jokes is the one rated as funnier by people.
###

This scene takes place in the following location: a cave. A caveman is drawing a picture of an elephant on his cave wall. The elephant is standing by as a model. The elephant is friends with a man. The scene includes: Caveman, Mammoth, Cave_painting. choices:
A) Trust me. One day your portrait will be used as the symbol of a political party even more primitive than we are.
B) So I've added the pointy trunk. Were there any other unique characteristics the mugger had that you remember?

the funnier is:
```

Figure 9: Example GPT-3 zero-shot prompts for Matching (top) and Quality ranking (bottom) tasks. In-context prompts are similar, except 5 random labelled training examples are also provided in the prompt.

	Matching	Quality Ranking		Explanation		
	Accuracy (↑)	CrowdAcc (↑)	NYAcc (↑)	B-4 (↑)	Rouge-L (↑)	PPL (↓)
Random	20.0	50.0	50.0	-	-	-
Caption Only (T5-11B finetuned)	19.4	59.4	64.5	3.61	17.8	34.0
`text-ada-001` (in context, n=5)	20.1	50.8	49.9	2.04	15.9	2367
`text-babbage-001` (in context, n=5)	19.0	51.3	51.1	2.18	17.2	137
`text-curie-001` (in context, n=5)	20.4	51.0	50.0	2.99	18.1	108
`text-davinci-001` (in context, n=5)	35.6	54.4	53.8	3.79	19.5	151
`text-davinci-002` (in context, n=5)	**57.2**	**55.1**	**54.8**	**5.07**	**20.5**	**107**

Table 4: GPT-3 scaling experiment results, averaged over 5 cross-validation splits. In all cases, models are given access to the same sample of 5 in-context examples. Overall, `text-davinci-002` performs best — this appears to be both because of scale (e.g., `text-davinci-001` generally outperforms `text-curie-001`) and also because of training improvements in the updated `002` version of the model.

Forming Matching Instances. For each high quality caption, we create a matching instance that serves as the correct answer. Next, we randomly assign captions to mismatched contests to form negative, mismatched sets to serve as false options. While the assignment is random, we have two constraints: 1) we assign within cross-validation splits only, to ensure that training/validation/testing captions are disjoint; and 2) we construct the corpus with no answer-only biases by performing the negative assignment such that each answer appears exactly once as a correct answer and exactly 4 times as an incorrect answer in other instances.

Forming Quality ranking Instances. For each high quality caption, we aim to sample from the larger set of all submissions for the contest captions that are just "okay." First, we note that 25 contests from early on in the contest's history were missing entries, so we are limited to sampling negatives for 679 contests. Next, because many entries are exact duplicates, we deduplicate on string matching, such that "okay" captions are not exact copies of 1) the identified high quality captions; and 2) any other sampled "okay" captions.

Next, for later contests from Jain et al. (2020), we have estimated quality ratings based on crowd feedback for each entry already: in that case, we discard the top third and bottom third of captions according to mean crowd rating — the middle tertile form the "okay" set we sample from.

But, for earlier contests, we do not have direct ratings: we only have access to New Yorker finalists and a large pool of entries. For those cases, we aim to eliminate captions that are clearly likely to be low quality. To accomplish this, we train a quality ranking model (conditioned just on the caption text, rather than any information about the contest) using crowdlabelled data from 253 contests provided by Jain et al. (2020). We sample a good/bad set by selecting from each contest the top and bottom 1000 entries according to their mean crowdsource score: the resulting dataset forms a corpus of 506K captions. We form two sets of labelled data based on the parity of the contest

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.
###

This scene takes place in the following location: a laboratory. A man in lab attire is sitting in a room with several monkies. Most are in cages, but one of them is in front of a typewriter It's unusual to see a monkey operating a typewriter The scene includes: Infinite_monkey_theorem, Mad_scientist.
caption: Have you considered writing this story in the third monkey rather than the first monkey?
explanation of the caption:
Stories can be told in first person (e.g., "I did X") or third person ("The character did X"), and editors will sometimes critique this choice or offer suggestions about the writing style. But here, the monkey is writing, so instead of first/third person, the suggestion about perspective is first/third "monkey".
###
This scene takes place in the following location: city. Devils are herding people in the middle of a street. A manhole is open and there is fire below. Devils in the middle of a city are pretty out of place, and hell being in the sewers is too. The scene includes: Devil, Sewerage.
caption: Watch your step, I think this street is paved with Good Intentions.
explanation of the caption:
A play on the figurative saying "The Road to Hell is Paved with Good Intentions" — this saying means that good intentions, left unchecked, can often lead to the worst outcomes. But here, the street is literally a path to hell, and so the man is saying it must be the street from the saying.
###
This scene takes place in the following location: a streetcorner. Two angels driving a police car look on as two other angels loiter and drink on a street corner. The scene is cloudy, which implies that this is happening in heaven Heaven is an idealistic place where one wouldn't expect police to be needed. Furthermore, the folks sitting on the sidewalk and drinking do not accord with a the platonic, perfect view of heaven — if things are so perfect, they probably wouldn't need to be drinking. The scene includes: Heaven, Public_intoxication, Angel.
caption: Except here it's always good cop, good cop.
explanation of the caption:
A play on the phrase "good cop, bad cop" — two homeless angels are loitering in heaven as two angel cops come by. Instead of the usual "good cop, bad cop" routine where one officer pretends to be nice and the other is mean as a strategy to get a confession, in heaven apparently all cops are good, so it's "good cop, good cop".
###
This scene takes place in the following location: an office. Two people are sitting across a desk from one another. They are both tied up in straight jackets. People don't normally wear straight jackets in offices. The scene includes: Straitjacket, Interview.
caption: I have some good news—we're letting you go.
explanation of the caption:
A play on the words "letting you go" – normally, if your boss was to "let you go" it would be bad news, because they would be firing you. In this context, "letting you go" means removing you from your straight jacket, which would be good news. Because the boss also has a straitjackets, the employee might be getting fired from this company where everyone wears straitjackets, but that would still be good news because this is a terrible workplace.
###
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:

`text-davinci-002`

The man is trying to put files into a filing cabinet, but there is a tiger sitting on top of it. The tiger is probably taking up a lot of space and making it difficult to put the files away. The man says he's going to miss this when they go paperless, meaning he's going to miss the struggle of trying to put files away with a tiger in the way, because once they go paperless there will be no more files to put away.

This is a joke about how going "paperless" (i.e., using digital files instead of physical paper) is often touted as being more environmentally friendly. But here, the man is saying he's going to miss having a tiger in the office when they switch to digital files — presumably because it won't be as exciting.

A play on the term "paperless" — "going paperless" is a goal for many companies because it would save money and be more efficient. But here, the company is so far from that goal that they have a tiger on top of their filing cabinet. So instead of "going paperless," this company is going "tiger-full."

`text-curie-001`

A play on the phrase "going paperless" — instead of carrying around a bunch of folders, documents, and other paperwork, people might move to a system where everything is electronic. But here, the tiger is a reminder that not everything will be digital, and there will still be times when people need to file things away in a physical filing cabinet.

A play on the phrase "going paperless" — in the past, when people had to file paper documents, it would take up space and be a hassle to manage. But now, with electronic files, the phrase "going paperless" means that all of these documents will be stored electronically, and so they will no longer take up space or be a hassle to manage. However, in this situation, the tiger will still be around, meaning that the transition to a paperless system has not been completely successful.

A play on the phrase "going paperless" — when people go paperless, they means they're going to stop using filing cabinets, which is a physical representation of the transition to a digital world. In this scenario, the joke is that the man is going to miss the tiger, which is a physical representation of the tiger being inside of the filing cabinet.

`text-babbage-001`

A play on the phrase "paperless" — in a paper-based office, the files would be placed on a paper filing cabinet, which would be taken away when the office moved to a paperless setting. The tiger might symbolize the fear of not being able to find or eliminate the files, as well as the idea of being lost in an office with no way to find your way out.

A play on the words "paper" and " loneliness." It is interesting to see the tiger in the office, because it is a reminder of the files that the man is trying to place inside the cabinet. In the same way, the tiger is a reminder of the man who is lost and lonely, and the office is a place where he can be alone and focus on his work.

A play on the phrase "paperless" — "paperless" usually means having to carry a lot of paperwork by hand. But in this context, it means that the office will be without the need for paper filings, which would be a big loss. Because the office is in an office building, it's also likely that the tiger will be released and roam the building freely.

`text-ada-001`

People often lose interest in paperless files when they do not receive any positive feedback from their users. This phrase is a good metaphor for how people feel about not being able to enjoy or enjoy their users. In fact, the man in the office who is trying to place files inside the filing cabinet might be missing the benefits of using the files himself.

You might be wondering why I'm saying these things, but one of the things I love about work is that it's a perfect-disrupted-bleeds-job kind of job. In a perfect-disrupted-bleeds-job, there would be no files, no only the Office, and no customer service centre. In fact, the customer service centre would be in a perfect-disrupted-bleeds-job. So it would be hard to find a perfect-disrupted-bleeds-job, because there would be no customer service, and no office.

A play on the term "documentary" — when you watch a documentary, it feels like one of those things you know is going to be funny. In this situation, the documentary-holder is trying to absent-moved the tiger from the office building, and the tiger is really enjoying it. The appearance of the tiger also means that the office building isn't being used as something as a place to store files, but rather it is being used as an opportunity for the tiger toatomicue.

Figure 10: An illustrative example of an in-context learning prompt for generating joke explanations (1095 tokens). 3 samples with temperature .8 from different GPT-3 engines are shown. According to our experiments, `text-davinci-002` performs the best; qualitatively, as model size decreases, explanations become more nonsensical.

`system prompt` (human authored)

You are CaptionContestGPT, an expert language model at understanding the famous New Yorker caption contest. You follow the contest each week, and understand what makes for a humorous caption for each cartoon. You are aware of the various theories of humor, and read/anaylze the caption contest entries and winners each week.

Some things to remember:
- You're well versed in the history of the New Yorker Caption contest, and the types of captions that are selected as finalists/winners vs. those that are not.
- You think step-by-step, but aren't overly verbose.
- You can express uncertainty in your thinking, but in the end, pick the single best answer in the requested format.

`user` (human authored)

I will describe a New Yorker cartoon to you. Then, I will give you 5 choices (labelled A-E) for captions. One of the captions was the winning caption for that cartoon, the other captions do not correspond to this cartoon. Your job is to first reason step-by-step about which answer might be correct, and, in the end, respond with "Answer: X" where X is either A, B, C, D, or E.

`assistant` (human authored)

Sure, please describe the New Yorker cartoon, and provide me with the 5 caption choices.

`user` (human authored)

OK. Here is a description of the cartoon followed by the five choices.

scene location: a book signing
description: There is an author signing at a bookstore. There is also a line of people with a big lizard in front.
uncanny description: The giant lizard in a bookstore does not belong.
entities: Godzilla, Autograph

Choices:
A: "MEET ??". I heard "EAT".
B: Trust me, I know the plague when I see it.
C: I'd like to see other people.
D: Shouldn't you be sitting in an ear?
E: So Major Tom... Do you know how fast you were going?

Which of the 5 options (A, B, C, D, or E) is the best fit? Think step-by-step and finish your response with "Answer: X" where X is either A, B, C, D, or E.

`assistant` (machine generated; used to derive the prediction)

Let's analyze the options step-by-step:

A: This caption seems to play on Godzilla misinterpreting the book signing event ("MEET") for "EAT," creating a humorous situation. This caption relates to the cartoon and its elements.

B: The mention of the plague doesn't relate to the cartoon description. It seems unrelated to the scene and entities mentioned.

C: This caption about seeing other people doesn't make sense in the context of the cartoon description and lacks a connection to the scene and entities.

D: The reference to sitting in an ear also seems unrelated to the cartoon description. It doesn't connect to the entities or context.

E: This caption about Major Tom and speeding are again unrelated to the cartoon description. This caption does not fit the scene or entities.

Through this step-by-step analysis, we can narrow it down to A as the best option.

Answer: A

Figure 11: An example of a zero-shot prompt+completion for GPT-4 (OpenAI, 2023) when applied to the matching task. In contrast to the text completion API of GPT-3, the GPT-4 chat API requires a more structured input involving a "system" prompt specifying the behavior of the model, followed by an interleaved conversation between a system and a user. While the training process of GPT-4 is opaque, in general, its "chain of thought" generations loop over all options and attempt to reason about how/why a caption might relate to the given scene.

number (i.e., even vs. odd). We train/validate two T5-Large models based on this split for the binary classification task. While the average validation accuracy we achieve is 65%, we achieve higher precision in identifying the “bad” label: precision-at-10 is 83, precision-at-20 is 77, precision-at-30 is 72. It appears to be harder to identify very good captions than very low rated ones: precision-at-10 is 77, precision-at-20 is 73, precision-at-30 is 70. Upon training these models, we perform inference on all captions in contests without crowd ratings, and discard the 25% of entries with the lowest predicted score. Entries with very low scores have some common characteristics, e.g., they don’t have the *gestalt* of a New Yorker caption, they have many typos/formatting issues, they include the contact information of the submitter, etc. Examples of discarded captions (some are obfuscated for privacy reasons) are:

- THEY COULDN’T WAIT TO MARRY SO THEY CAME TO RECITE THEIR VOWS BETWEEN TAKES FROM “ PRIMITIVE LOVE LIFE”
- You’re hurting me, will we ever break up?” (@ technology)
- The stressed is so “Bad’ in the world. “you or me ” did not see(BIG)(“FOOT)
- Too mammalian, needs reptile.” [NAME], [STATE] [EMAIL]@gmail.com

After identifying a set of captions that are not obviously bad, nor apparently among the top quality submissoins, our second step is to deduplicate entries. Because submitted captions for each contest are frequently identical to other submissions or play off the same core joke concept, we perform the same SBERT+hierarchical clustering semantic deduplication step as we did for sampling the diverse high quality set (described above). Specifically, we extract SentenceBERT embeddings (Reimers and Gurevych, 2019) for each of the N entries, and then compute a hierarchical clustering of the embeddings into $.7 \cdot N$ clusters, sampling only a single representitive from each cluster to form a less-redundant set. This removes 30% of the data with close neighbors in the final set: for example, for a contest depicting two monsters eating buildings in New York City, this step downsamples 100 “tastes like chicken” jokes (which end up in a single cluster) to a single exemplar.

After filtering, for all contests, we are left with a (softly) deduplicated pool of candidate entries that are likely to be at least okay, but unlikely to be as good as the verifiably high quality entries. For each high quality entry, we sample an “okay” caption with: 1) similar estimated quality according to the text-only models; 2) similar length in words; 3) similar length in characters; 4) similar amount of punctuation; 5) a dissimilar SBERT embedding.

Explanation corpus. After several attempts to solicit high-quality explanations from crowdworkers fell short, one of the authors of this paper decided to simply annotate a corpus of explanations themselves. For each contest, a high quality caption was sampled for annotation — this high quality caption was sampled arbitrarily from the set of New Yorker finalists if they were available, and, in the few cases where New Yorker finalists weren’t available, from the set of high quality crowd captions. Of the 704 sampled explanations, the author reported understanding 651 of them, and wrote an explanation for each. This was a substantial effort: the resulting corpus has a mean of 60 words of explanation per cartoon, and the total length, 39.3K words, is comparable in length to a novella.

D Graphical version of matching and ranking results.

In Figure 12, we use vertically-stacked bars to illustrate the difference between zero-shot (small dots), five-shot (vertical stripes), and fine-tuned (solid) versions of various models. Human results are set off by dark green lines.

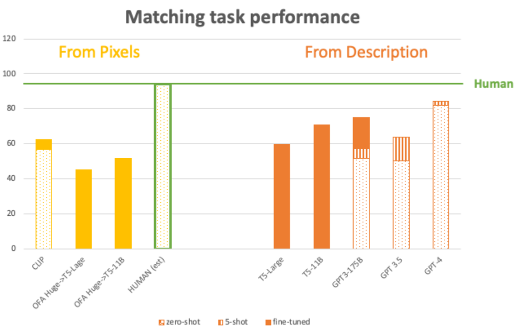


Figure 12: Graphical version of the matching results given in Table 2.

The scatter-plot in Figure 13 uses the same graphical conventions to display the quality-ranking results. Recall our caveat that crowd accuracy may be more statistically reliable, in the sense that crowd selectors, whose tastes underlie the y-axis results, vastly outnumber New Yorker

		Explanation		
		BLEU-4 (↑)	Rouge-L (↑)	PPL (↓)
	Caption Only (T5-11B)	3.61	17.8	34.0
FP	OFA-Huge → T5-Large	3.36	17.5	50.7
FP	OFA-Huge → T5-11B	<u>3.63</u>	<u>17.9</u>	<u>30.3</u>
FD	T5-Large	3.54	18.2	41.2
FD	T5-11B	4.33	19.0	23.7
FD	GPT3-175B (finetuned)	**5.42**	20.1	**21.8**
FD	↳ 5-shot	↳ 5.07	↳ **20.5**	↳ 107
FD	↳ Zero-shot	↳ 3.12	↳ 18.8	↳ 225
FD	GPT 3.5 (5-shot)	3.94	18.8	-
FD	↳ Zero-shot+CoT	↳ 2.40	↳ 17.3	-
FD	GPT-4 (5-shot)	4.99	20.0	-
FD	↳ Zero-shot+CoT	↳ 3.42	↳ 19.0	-

Table 5: Results for the explanation task using automatically computed metrics. Results are averages over 5 cross-validation splits. <u>Underlined</u> results are the best model in the *From Pixels* (FP) setting, where at test time, models only have access to the cartoon images. **Bold** results are best in the *From Description* (FD) setting, where at test time, models have access to human-authored descriptions of the cartoons. GPT-3.5 and GPT-4's API does not provide log probabilities, so we can't compute perplexity for those models.

editors, whose tastes underlie the x-axis results.

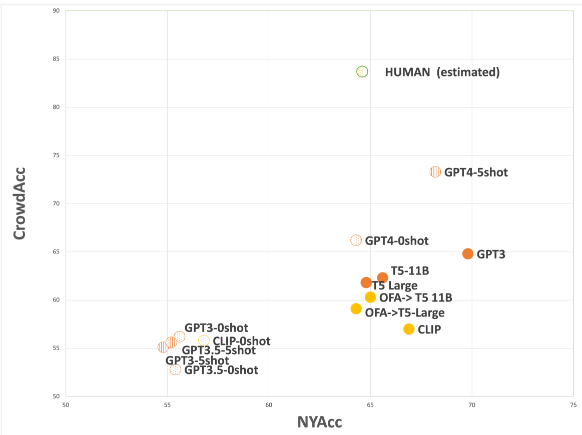


Figure 13: Graphical version of the ranking results given in Table 2.

E Automatic evaluation of explanations

For completeness, we provide the results for automatically-calculated explanation-evaluation metrics in Table 5. (Log probabilities are unavailable for GPT-3.5/GPT-4 so we cannot report perplexity for them.) However, we believe that the human evaluations reported in the main body of the text are better quality measures.

F Machine explanations that were preferred over human ones

GPT-4 In 8/130 cases, for our human vs. GPT-4 5-shot experiments, the machine generation was preferred to the human reference by 3/3 annotators. In Figure 14 we conduct a close reading of these 8 instances to understand where the human references fell short. In all cases, both were topical, but, for a handful of cases, the machine generation is arguably better because it's more succinct, or offers a more meaningful detail.

GPT-3 We also include a close reading of several instances where a majority of annotators preferred GPT-3 annotations vs. our human ones. This occured in 16/131 cases for our human vs. GPT-3 experiments: in 15 cases, 2/3 annotators preferred the machine generation, and in 1 case, 3/3 annotators preferred the machine generation. We present a few examples of these cases with comments in Figure 15. Similar to GPT-4, most commonly, both the explanations were reasonable, in one case, the human-written explanation missed a pop culture reference to "The Attack of the 50 Foot Woman" (1958) but GPT-3 mentions it. In six cases, we disagree with annotator consensus: the machine caption makes some correct references, but we believe it (sometimes subtly) misses the point.

G Aiding humor generation with system-assisted brainstorming

Our main experiments focus on three tasks that probe machine capacity for matching, ranking, and explaining caption contest entries. But, given prior interest in generating caption entries, to provide a pointer towards more creative use cases, we developed a curated prompt that re-frames the corpus in a manner that enables the production of cartoon descriptions, concept brainstorming, and, ultimately, caption generation. In contrast to our existing annotations, we are inspired by Wallace (2022)'s framing, where free associations are used to connect-the-dots from cartoons to concepts to captions. So, for 9 finalist captions, we reverse-engineer a plausible brainstorming process that could have given rise to the caption by listing several potential associations from the scene, and, from that list, selecting the 1-3 that underpin the basis for the winning joke. In the end, our prompt consists of 9 examples formatted as:

1. cartoon description (from our annotations)
2. uncanny description (from our annotations)
3. list of entities in scene (from our annotations)
4. list of concepts from the cartoon that could be used for a caption (hand-designed for the

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)
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)
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.

Scene/caption	Our opinion re: annotator decision	Human-authored explanation	Machine-authored explanation
Two teachers walking through a hallway of an elementary school while the children walk on the ceilings and walls. *Caption: Back in my day, kids had a little respect for the law of gravity.*	*Disagree:* the machine caption starts off well, but concludes with a number of non-sequiturs. Perhaps the annotators only skim the later parts of a long explanation.	(1/3) Normally "respecting laws" are human-written laws — but here, the kids have somehow opted-out of respecting physical laws of gravity, which means they can walk on walls/ceilings.	(2/3) The kids in this school are violating the laws of physics by walking on the walls and ceiling. The teachers is complaining that kids used to respect gravity, but now they don't. This is a funny complaint because whether or not kids respect gravity probably depends on the teacher's training, not whether they were a kid in the past.
A doctor is performing a physical examination on Superman. *Caption: They think now kryptonite is actually good for you, in small doses.*	*Disagree:* Both are okay, but the human-authored explanation is more specific in referencing the appropriate cultural phenomenon, and the machine-generated one hallucinates that kryptonite is medicinally helpful.	(1/3) A joke about modern medical findings sometimes being counter-intuitive: some studies claim to show that things harmful in big quantities like alcohol or chocolate are good in small doses. But here, the thing that's bad for superman in large doses is kryptonite: it's a funny parallel to the other studies, and funny to think that scientists are studying kryptonite's potential positive impacts.	(2/3) Kryptonite is a material that causes superman to become weak; however, the doctor is suggesting that kryptonite is good for him in small doses. This is funny because, while kryptonite might be helpful medicinally, it's difficult to explain how small doses of something that kills superman could possibly be helpful.
3 people are going to therapy in a 3-tiered bunk-bed. *Caption: And how do you feel about how she feels about how he feels?*	*Disagree:* The machine caption omits the oddity of having *three* patients involved in what appears to be couples therapy. Nor is there wordplay involved.	(1/3) Couples therapy is already complex – pairs of people may have feelings about their partner's feelings. Here, the joke is that if three people were getting therapy together, things would get even more complicated and meta, i.e., person 1 could feel strongly about person 2's feelings about person 3 (and any permutation of such). It's funny to think that these three would be so closely intertwined such that this meta-therapy could be a helpful exercise.	(2/3) A wordplay joke about feelings — a therapist might ask a question like: "how do you feel about what happened with your brother?"; in this case, the question is expanded to include the feelings of each party, e.g., "how do you feel about how he feels about what happened with your brother?"; this is a very complicated/deeper concern that requires a lot of reflection.
A board meeting is happening in an office, but the employees are birds. *Caption: We have to stop eating the seed money.*	*Agree:* the machine explanation is probably better.	(0/3) Seed money is cash provided to a startup company (like the ones the birds are running); but, the play on words implies that the birds spent the seed money literally on seeds to eat, because birds eat seeds.	(3/3) A play on the phrase "seed money" — in the investment world, "seed money" refers to the initial capital that is invested into a startup company. But here, "seed money" refers to bird seed that these birds might eat: the joke is that they need to stop eating bird seed because it's causing them not to be able to afford investments like a real office/chair/table.
A stack of cars with drivers in them. The driver on the topmost car is leaning out of the window yelling to the drivers below. *Caption: I hope this guy Knievel knows what he's doing.*	*Both have pros/cons:* The machine is more specific in referencing crashing and the implication for the drivers, but makes up "jumping the stack" meaning crashing.	(1/3) A reference to Evil Knievel, a stuntman who jumps motorcycles over cars. This pile of cars is going to be an obstacle in an Evil Knievel stunt, and the man in the top car is hoping that the stuntman coordinating this both knows how to jump over and also how to get the drivers out of this precarious situation.	(2/3) A reference to Evel Knievel, a famous stunt man that jumped cars and other objects. This is a joke about him jumping cars, but because the cars are stacked precariously, it's a reference to Evel Knievel "jumping the stack", i.e., crashing, which would be a bad outcome for the drivers.
A person wearing boxing gloves is attempting to play a piano on a stage. *Caption: I could have been a conductor.*	*Both bad:* the human caption gives a better mechanical explanation, but both miss that the caption is a play on a quote from "On the Waterfront" (1954), "I coulda been a contender."	(1/3) It would be mechanically difficult to play a piano with boxing gloves because the keys are too small. if this person must wear boxing gloves and still wants to do music, they humorously reason that they should be a conductor — it would be easier for them to hold the conductor batons vs. play piano keys.	(2/3) A joke about Muhammad Ali's skill as a boxer, which sometimes takes a metaphorical meaning when referring to his boxing career. Here, the boxer is playing the piano, and is presumably not very good at it. But, the joke is that, if he wasn't a boxer, he could have been a conductor, which requires a high level of skill. It's funny in part because it's not clear why he would have not been a boxer.
A father and a child are standing in an office building looking out the window onto a giant woman straddling a nearby building. *Caption: Son, we finally attracted the fifty-foot-and-over demographic!*	*Both bad:* The human-authored explanation misses the direct reference to the movie "Attack of the 50-Foot-Woman" (1958), and the machine focuses on non-sequiturs like age 18/viewers/etc.	(1/3) A play on the phrase "fifty and over demographic" — this phrase would usually refer to the group of people who are aged 50 or older, and would be of interest to people running businesses (like the father and son pictured) as a marketing target. But instead of the age 50+ demographic, they have attracted the height-of-fifty-foot+ demographic, of which this giant woman is a member.	(2/3) A play on the term " fifty-foot-and-over" and "over 18" — in the media, advertisers sometimes say that they want to attract 18+ year old viewers to their product, because 18 is the legal age of consent, and thus, to attract 18+ year olds, they will say they want to attract viewers "over 18". But here, the company is trying to attract viewers "50-feet-and-over" — the company is trying to attract the titular Attack of the 50-Foot-Woman (who is, indeed, over 50-feet-tall).

Figure 15: Examples of the small number of cases (16/131) where the machine-generated caption (fine-tuned GPT-3 175B) was preferred to the human-authored explanation by at least 2/3 annotators.

First, you will see a description of a scene from a New Yorker cartoon that has an unusual and/or funny element. Our goal is to brainstorm a caption that we can enter into the caption contest. The captions should be funny, and relate to the image. As part of this process, we will brainstorm about potential concepts from the cartoon, and then combine those concepts to make the caption. Finally, we will explain the caption, and how it relates to the scene.
===
this scene takes place in/at/on a kennel
description of scene: Two men are sitting in a dog cage next to a dog. One of them is reading a book while saying something to the other man who is visibly upset.
unusual part of scene: these cages are for dogs in shelters not for humans.
entities in scene: animal shelter, doghouse.
potential concepts to form a funny caption based on: men as dogs, chew toys, being adopted at the pound, spaying/neutering pets
here is a caption that combines these concepts: men as dogs, spaying/neutering pets
funny caption: Last time I'll book a discount vasectomy, for sure.
explanation of how the caption relates to the scene: Spaying/neutering pets is a commonly performed operation for animal sterilization performed at animal shelters; the human equivalent sterilization procedure is called a vasectomy and is usually more expensive due to higher quality, human medical requirements. But here, the two men are trying to save money by having their operations done at an animal shelter. It's funny in part not only because this is an unusual way to save money, but also, because vasectomies only are performed once per person, so this is literally the last time the person will do this procedure, even if they liked it; the quote implies they aren't enjoying being locked in an animal cage.
===

... (8 more examples formated as above) ...

===
This scene takes place in/at/on

`text-davinci-002 (unconditional sample)`

a garden
description of scene: A group of people are playing croquet. One of the players is a very large chicken.
unusual part of scene: Chickens are not usually intelligent enough to play croquet and they are not usually that big.
entities in scene: chicken, croquet.
potential concepts to form a funny caption based on: farm, clucking (noise chickens make), free-range chicken, backyard game, playing chicken (game, metaphor), cockfighting
here is a caption that combines these concepts: free-range chicken, backyard game
funny caption: I'm not sure this is what they meant by free-range.
explanation of how the caption relates to the scene: A play on the term "free-range chicken," which is a type of chicken raised for meat that is allowed to roam freely outdoors, as opposed to being confined in a cage. Here, the chicken is not only free-range, but also playing a backyard game. The joke is funny because it's a play on words (the chicken is both free-range and playing a game) and because it's unexpected (chickens are not usually this big or this intelligent).

`text-davinci-002 (conditional sample, given the italics)`

a living room
description of scene: A person is sitting in their living room, but is looking towards the door as an entire circus, complete with all performers, is walking through the door.
unusual part of scene: It is unlikely and disruptive for an entire set of circus acts to be intruding on a quiet living room.
entities in scene: circus, bystander, performers, circus acts.
potential concepts to form a funny caption based on: unannounced visitors, salespeople, clowns, big top
here is a caption that combines these concepts: unannounced visitors, salespeople
funny caption: I'm never buying a timeshare again.
explanation of how the caption relates to the scene: The circus is a metaphor for an unannounced group of salespeople who are trying to sell a timeshare. The joke is funny because it's an extreme example of an unannounced group of salespeople, and also, because it's disruptive and intrusive.

Figure 16: A portion of a 2,407 token prompt that re-formulates various annotations within our corpus in a format conducive for creative collaborations with a language model. The full prompt is available here. Generating line-by-line from this prompt could help to facilitate brainstorming for: unusual cartoon situations (first 4 lines), concepts about real or generated contests that could serve as a basis for a humorous caption (line 5), and, captions themselves (lines 6-8). As a demonstration, we present an unconditional sample, in which the model describes a garden party where a chicken is playing croquet (cherry picked from 3 outputs; temperature=.8, top p=.9, frequency penalty=.2, presence penalty=.05), and also, a conditional sample, given a basic description of Contest #818's scene, which ran in mid-September 2022 (cherry picked from 5 outputs; same sampling parameters): the caption is arguably funny, but the explanation is not correct.

prompt)
5. a selected set of 1-3 ideas (selected from (4))
6. caption (a finalist)
7. explanation of the caption (from our annotations)

A portion of our prompt is given in Figure 16, along with an unconditional generation (where the cartoon concept and caption are generated) and a conditional generation. Within 5 samples, GPT-3 invents a scene where a large chicken is playing croquet in a yard, and the caption: “I’m not sure this is what they meant by free range.” Also, when conditioned on a basic description of a real contest which depicts a large group of circus performers intruding on an unsuspecting person in their living room (Contest #818), it generates “I’m never buying a timeshare again.” Looking forward, we expect the matching/quality ranking models could be used in conjunction with this prompt to automatically filter for scene-specific generations with style similar to previous finalists.

H Related work beyond peer reviewed AI venues

Outside of peer-reviewed NLP venues, several projects have used computational techniques to analyze the contest, usually with the goal of generating AI-assisted entries:

- **The Pudding:** Mishkin et al. (2022) collaborated with GPT-3 (Brown et al., 2020) to generate entries.
- **coolposts:** Wilson (2019) used topic models to condition an RNN caption generator.
- **LILY Lab @ Yale’s** Spring 2017 projects include a number of caption contest efforts, including work by Prince, Friedman, Zucker, Anbarasu, and Dohrn.
- **The Verge:** Zelenko and Bi (2015) trained a Markov language model on previous winning entries.

I Some of our favorite New Yorker cartoons

We list our favorite captions below. The corresponding images can be seen by clicking on the cartoonist/author names.

YC: “The doctor said it might help me quit.”

— Vince Conitzer/Jeffrey Adam Katzenstein

JD: “You are so smart. You look amazing. You inspire me. [Complimentary bread].”

— Seth Fleishman

JMH: “Thanks, I’ll write that down.”

— Victoria Roberts

JDH: “They’re from Earth. I wonder if they know Dan.”

— Benjamin Schwartz

LL: “I want to be feared as a tyrant, loved as a father, and revered as a god, but I also want them to think I’m funny.”

— Zachary Kanin

AM: “I can’t believe I’d been carrying them in my mouth.”

— Amy Hwang

RZ: “Well, there’s your problem.”

— Edward Koren

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section 6

☑ A2. Did you discuss any potential risks of your work?
Limitations section 6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The abstract

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

yes, our new corpus/tasks. Section 2 describes them.

☑ B1. Did you cite the creators of artifacts you used?
Yes, section 2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Yes, we discussed the distribution of our dataset, which have made public under Creative Commons Attribution 4.0.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, section 2

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 2 and appendix C

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3 and appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3 and Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 2

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 2, Appendix A

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 2, Appendix A

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
appendix A

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We don't know many specifics, other than country of IP: which we discuss in appendix A