

# Do Electric Sheep Laugh at Androids? A Bidirectional Understanding of Contextual Information within Humor.

Nicholas Konovalenko, Shaolin Xie, Jeya Varshini Bharath

November 7, 2023

## Abstract

"Do Androids Laugh at Sheep" by Hessel et. Al [4] attempted to gain an understanding into whether vision and language models, as well as language only models, would be able to understand humor through a cartoon captioning contest. Their work focused on learning about humor through three separate tasks. First, the model had to choose the correct caption for an image out of several options. Second, the model had to rank the captions given an option of a finalist and non-finalist. The final task required the model to give a description of why the caption was funny. They found that while the models could achieve an accuracy higher than 60% on these tasks, they performed worse than humans. While this is important for humor, our team believes the image input to text output is only half the understanding required for true success with humor. Our team will be focusing on the other half. We will explore whether the model can extract contextual information from the jokes within the captions, such as location, subject, and other descriptors in order to generate an image that would make sense with the specified caption. We also wanted to explore two additional tasks related to idioms, the use of a specific type of phrase that is humorous when one understands the language well, but nonsensical when taken literally. We find that while language models such as GPT-3 have a thorough understanding of idioms, image models such as DALL-E2 still take these prompts very literally showing room for improvement. With this full two-way understanding, our team believes it'll be possible to make progress more quickly in understanding humor with machines.

## Introduction

The overall goal of Hessel et. Al [4] was to gain insight into whether models could learn to understand humor. While three experiments had been presented, we believe that these are incomplete. The experiments focus on learning features from the image to know which caption to select, where contextual clues will be very obvious. Their annotations include a 1-3 sentence literal explanation of the image for every cartoon, from which it would be easy to extract features. We believe that an additional experiment is necessary to explore the opposite, the ability to extract contextual information from the caption rather than the image, as text to image humor is largely unexplored. The models should be able to generate images that would be funny with the given caption. Additionally, we further delve into the exploration of humor within language by examining the models' abilities to understand idioms. The phrases slip into our usage of language without much thought behind the literal meaning, however does the same apply to GPT-3 and DALL-E2?

## Related Work

There are several previous works brought up by Hessel et. Al [4] which will be crucial to the success of our experiment. First, the original authors' experiments were based on incongruity as most cartoons from the New Yorker caption contest purposefully involve incongruous situations as described in Mulder and Nijholt[8].

Further, the authors reference Shahaf et al [10], Radev et al [9], and King et al [7] when setting up their experiments. These focus on covering more contests, using new data/resources/models to curate ranking pairs and making explicit evaluation distinctions between New Yorker editor preferences and crowd voting preferences.

Additionally, the data set includes ratings for captions, which Jamieson et al [5] studied modeling the average preferences of raters of the captions in the second task.

This will not be possible without multimodal interactions in humor. Jones et al [6] asked participants to judge cartoons with and without captions, concluding that "it is the dynamic interplay between picture and caption that describes the multiplicative relationship and the most humorous cartoons".

The goal is to automatically generate humorous content in various contexts as discussed in Amin and Burghardt. [1]

Finally, the paper by Chakrabarty, Choi and Shwartz shows that there is still ongoing research to see if language models truly understand the use of idioms and other figures of speech. [3]

## Methods

Our goal is to gain a deeper insight into how well textual models, image models, and the combination of both, understand humor. Since we wanted to establish benchmarks for humor understanding, we chose to use OpenAI GPT-3 Davinci model as well as DALL-E2, the most widely used text and image models currently. To accomplish this, we design a suite of three tasks that evaluate their grasp of humor.

In our first task, we examine how well computers extracts contexts from captions and interpret a picture for the given captions. We will first train the model by running few-shot learning on the first 50 entries of the cleaned NY-CCC dataset[4], which is done by feeding them to the model with the format of prompt: "Give a funny scenario for [a caption]", completion: "the expected corresponding descriptions". We then generate our testing results on entries from the rest of the dataset by asking the trained GPT-3 model to fill the "completion" by giving a prompt in the format as described above. We then pass the results, which are the funny

descriptions of the given captions, to the DALL-E2 machine to have it generate pictures. By conducting human evaluations on those pictures on a scale of 1 to 3, we will evaluate model's understanding of humor.

Our second task will evaluate individual model understanding on a list of idioms [2]. We query GPT-3 with a prompt of “Describe what it means to [IDIOM]”, to receive an explanation of the idiom. A human evaluator will rate whether the explanation is literal, or shows an understanding of the idiom. We also pass the same prompt to DALL-E2, and generate images of the idioms. A human evaluator will rate whether the image is a literal representation of the prompt, or whether it conveys the meaning of the idiom.

Our third task will combine the textual and image models to evaluate overall performance. We will take GPT-3 output from the second task, and use that as input for DALL-E2. A human evaluator will then rate whether the generated image conveys an understanding of the idiom’s explanation, or whether it is interpreted literally.

## Dataset

To accomplish these tasks, we use two datasets.

For our first task, we use the New Yorker Cartoon Caption Contest(NYCCC) dataset [4]. The NYCCC dataset contains 300 data entries, where each entry is data from weekly contests. Each contest consists of a captionless cartoons and a list of some thousands of the cartoon’s corresponding captions ranked by their quality.

To train our model, we will use descriptions for the cartoons paired with their captions. Given by Hessel et. Al [4], we are able to find a dataset for all NYCCC’s cartoons annotations that contains 1-3 sentence of descriptive phrases for each cartoon paired with their top three captions. The data cleaning for annotation dataset is done by first trimming the contests that don’t contain captions. Then we create a JSONL file that has the highest ranked caption as prompt and their descriptions as completion.

For our other two tasks, we download a list of idioms from [2]. We will parse this list, to generate the prompts that we will feed to GPT-3 and DALL-E2 in the form of “Describe what it means to [IDIOM]”. We then save the output from GPT-3 for task 2, as a json dictionary in the form of “idiom”: “explanation of idiom”, to use as input for task 3.

## Results

To gather preliminary results for task 1, we implement few-shot learning on GPT-3 with three sets of prompts and completions. For our example, we ask the model to give a funny setting for caption “Actually, this is my first rodeo”. Before few-shot learning, the GPT-3 model completes with the following description - “The rookie is in for a big surprise when he shows up to his first rodeo. He had no idea what to expect, and he’s completely overwhelmed by the noise, the crowd, and the animals. He’s so scared that he can’t even move.” After few-shot learning, the model gives the following result - ”A group of people are in a rodeo. But instead of horses, they’re riding dinosaurs.” After visualizing the results by passing them in DALL-E2, we generate images shown in Figure 1 and 2.



Figure 1: ”Give a funny scenario for Actually, this is my first rodeo” - Before few-shot learning



Figure 2: ”Give a funny scenario for Actually, this is my first rodeo” - After few-shot learning

As shown in images above, the GPT-3 model is able to learn some sense of humor after few-shot learning, and the descriptions that it generate are able to be understood by DALL-E2. Before the training, it interprets the caption then responds with a scenario that simply describes the caption. After the training, the machine grasps on a sense of humor and provides a setting that looks controversial and out of place to the caption, which makes the resulted setting more humorous.

For the second task, we found that GPT-3 is able to accurately explain the meaning of idioms. Given the prompt “Describe what it means to be full of hot air”, GPT-3 gave a response of “When someone is full of hot air, they are boastful and have an inflated sense of self-importance”.

However, for several idioms, GPT-3 must be given context as to whether it is describing humor or not. For example, when asked “Describe what it means to break a leg”, GPT-3 responded with the literal explanation of “A break is a medical term used to describe when bones are cracked or fractured”. When framed as “Explain why people tell others to break a leg”, GPT-3 correctly explained that “‘break a leg’ is used as a way to wish someone good luck, especially before a performance”.

While GPT-3 showed success with this task, DALL-E2 performed very poorly. When given the prompt “full of hot air”, the model generated images of hot air balloons as seen in Figure 5.



Figure 3: A literal interpretation of “You’re full of hot air”

Furthermore, when given the prompt “Explain why people tell others to break a leg”, the model attempted to generate images with the words “break a leg”. This shows an understanding that it is a saying, however no understanding that it is for good luck in performances. Additionally, the phrase “you’re a lemon” means that someone is bad or useless at the task they are performing. However, DALL-E generates images of lemons and people looking at lemons, which does not convey anything related to poor performance as seen in Figure 4.



Figure 4: A literal interpretation of “You’re a lemon”



Figure 5: DALL-E successfully depicts the emotion, while still showing a hot air balloon.

For our third task, DALL-E2’s performance did not improve much. When we fed GPT-3’s explanation of “full of hot air” into DALL-E2, the images do show people. The difficulty with these images came within the emotion they show. Most of the generated images, depict people that are distressed, as opposed to those that are boastful. There had been one image that came close, with a boastful face on a hot air balloon as seen in Figure ???. These images still contained the literal subject in the idiom though. While it makes sense for generative models to require a literal interpretation, it shows a short-coming in their ability to depict humor.

## Conclusion

Looking at results from task 1, we notice that GPT-3 is able to learn humor very quickly by providing just a few learning examples. In the rest of our experiment, we will further train the model with more examples and generate results. For few-shot learning, when given 3 prompts, it is able to generate a humorous scenario but the results are very similar to each other. We noticed, when increasing the number of prompts, the responses have more variety.

From the results of the idiom experiments, the strength of language models is apparent. GPT-3 was able to explain idioms with zero-shot learning, while DALL-E struggled to generate images that were more than literal interpretations of the prompt. This limits DALL-E’s ability to convey deeper details, such as complex emotions, within its generated images. However, we saw that with the explanations from GPT-3, DALL-E got closer to the desired outcome. We will attempt to get closer through generating different outputs with GPT-3. Rather than generating the explanation of an idiom alone, we will also ask GPT-3 to describe what a situation with that idiom may look like such as, “What does something that is full of hot air look like”. With this, we hope to make DALL-E generate images that convey humor. Addi-

tionally, we will have a human evaluator score the results, so that we can get interpretable data for our results.

## References

- [1] Miriam Amin and Manuel Burghardt. A survey on approaches to computational humor generation. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online, December 2020. International Committee on Computational Linguistics.
- [2] Jamie Brew. Idiomash. <https://github.com/jbrew/idiomash>, 2019.
- [3] Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. It’s not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606, 05 2022.
- [4] Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest, 2022.
- [5] Kevin G Jamieson, Lalit Jain, Chris Fernandez, Nicholas J. Glattard, and Rob Nowak. Next: A system for real-world development, evaluation, and application of active learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
- [6] James M. Jones, Gary Alan Fine, and Robert G. Brust. Interaction effects of picture and caption on humor ratings of cartoons. *The Journal of Social Psychology*, 108(2):193–198, 1979.
- [7] Ben King, Rahul Jha, Dragomir Radev, and Robert Mankoff. Random walk factoid annotation for collective discourse. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 249–254, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.
- [8] M. Mulder and Anton Nijholt. Humour research: State of the art. 11 2002.
- [9] Dragomir Radev, Amanda Stent, Joel Tetreault, Aashish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 475–479, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
- [10] Dafna Shahaf, Eric Horvitz, and Robert Mankoff. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’15, page 1065–1074, New York, NY, USA, 2015. Association for Computing Machinery.