AI + Creativity + Humor: What We Found?

Zhuolun Zhong, Lihao Hou, Alex Cheung

December 21, 2023

1 Background and project overview

A recent study has introduced the idea of using humor data to assess Large language model AI performance. While AI's performance falls short of humans in various tests, the results are still noteworthy. For instance, AI sometimes outperformed humans in tasks requiring the most sophisticated understanding (explaining why a joke is funny), even identifying references that human participants did not recognize (Hessel et al., 2023). This result suggests that although AI cannot surpass humans in complex mental tasks involving humor comprehension, there is a trend toward performance close to human levels rather than nonsensical outputs. Humor is a relatively understudied and mysterious topic because it is highly associated with spontaneous creativity, making it challenging to design ideal experiments involving human participants. In recent years, the rapid development of large language models has provided new perspectives and directions for researching humor.

Humor, with the laughter followed, is a universal human experience (Apte, 1987). The combined study of these aspects can reveal the nature of certain cognitive functions in humans. Graeme Ritchie, a Scottish linguist and AI researcher, suggested that AI investigations of humor can not only help to clarify theories of humor, but can also lead to important discoveries about human intelligence, language, problem-solving, and information processing more generally (Ritchie, 2001, 2004; Ritchie et al., 2007). In the past, research on humor and AI yielded little satisfactory results. However, with the advent of GPT-4, which can now pass the Turing test (Jones & Bergen, 2023), it is time to challenge the realm of humor once again using AI, aiming to unravel the uniqueness of human nature.

On the other hand, research suggests that humor plays an important and complex role at different stages of an interpersonal relationship (Caird & Martin, 2014). Making people think more humorously can bring about significant positive effects. Therefore, exploring humor's value is not confined solely to human cognition but extends to social and personal values. The challenge has always been the elusive nature of humor, making it difficult to grasp and define.

Our project is based on data from the classic humor activity, The New Yorker Cartoon Caption Contest (Jain, Jamieson, Mankoff, Nowak, & Sievert, 2020). Firstly, we aim to comprehend humor through various tests and models. It includes performance tests of artificial intelligence, feedback tests from human participants, predictions from word embedding models, predictions from neural networks, incongruity tests, visualizations of topic models, and more. Overall, these tests have failed to establish a reliable computational model theory for humor or accurately predict humor. However, our work holds potential value for future work in understanding humor. Secondly, we integrate AI to design interactive interfaces that assist participants in creating captions for The New Yorker Cartoon Caption Contest. We employed a specific prompt strategy, and the interactive experience in the tests yielded positive results. This interface can be directly applied for comprehensive humor caption creativity effectiveness testing in the future.

2 Data and Resource

2.1 The New Yorker Cartoon Caption Contest Dataset

The New Yorker Cartoon Caption Contest is one of the most popular features of The New Yorker magazine. The contest releases a cartoon each week, and readers are invited to submit their own captions for the cartoon with the goal of being as humorous as possible. Readers can also vote on the submitted captions for humor ratings.

The data we utilized consists of humor ratings from The New Yorker Cartoon Caption Contest, which involves readers voting on submitted captions (Jain et al., 2020). The dataset has over two million captions from cartoon contest 510 to 876. Readers have three choices for a caption: funny (3 points), somewhat funny (2 points), and unfunny (1 point). The mean for a caption is calculated by dividing the total score by the total number of votes, and precision is measured by the

standard deviation of the mean. Typically, each contest receives several thousand caption submissions, but the total votes increase rapidly over time. The contests closer to the present have more public votes than past contests.The following shows attributes in the dataset:

- **Rank:** the rank of a caption in one cartoon contest
- **Caption:** the caption a reader generated
- **Mean:** a score for assessing the funniness of a caption
- **Precision:** the standard deviation of the "mean" estimate
- **Total votes:** the number of votes for a caption
- **"Unfunny" votes:** the number of "unfunny" votes
- **"Somewhat funny" votes:** the number of "somewhat funny" votes
- **"Funny" votes:** the number of "funny" votes

Regarding the distribution of mean values (funniness) for all captions, only a minimal number of captions reached two or above, and many contests had no captions rated two or higher. For a single contest, there are usually dozens of captions above 1.5 (around 1 percent), hundreds between 1.5 and 1.3 (less than 10 percent), and the remaining thousands below 1.3 (more than 90 percent). Since mean values represent mass data, predicting numerical values is challenging even for humans; it becomes even more difficult for computational models and AI to do the same task. Therefore, we used 1.5 and 1.3 as thresholds to label captions as funny, somewhat funny, and unfunny. Models and AI performance in predicting humor will be evaluated based on label accuracy rather than variance.

2.2 Data from a Research

We also have a dataset from the research "Do Androids Laugh at Electric Sheep?" (Hessel et al., 2023). This research investigates AI's ability to understand humor, specifically in the context of The New Yorker Caption Contest. It utilizes various datasets, including The New Yorker Cartoon Caption Contest Dataset to examine three main tasks and gain corresponding datasets:

- **Matching task:** This dataset comprised cartoons from the contest, with the task of matching the correct caption to each cartoon.
- **Quality ranking task:** This dataset involved ranking the quality of captions for the cartoons.
- **Explanation generation task:** A smaller dataset was used for this task, where the goal was to generate explanations for why a given caption is humorous.

The study finds that AI models, including GPT-3 and vision-and-language models, still have significant gaps compared to human performance in understanding and explaining humor, highlighting areas for future improvement in AI comprehension of complex humor.

The research yielded a valuable dataset comprising descriptions for each cartoon. This dataset is characterized by three distinct attributes for every cartoon: the location where the cartoon is set, a 'canny description' detailing the literal, observable elements of the scene, and an 'uncanny description' that highlights the unusual or humorous aspects. These attributes offer a comprehensive view of the cartoons, aiding in the analysis of humor and its context.

2.3 Large Language and Vision Assistant

Large Language and Vision Assistant (LLaVA) is an advanced multimodal model designed to enhance AI's understanding and interaction capabilities through visual and language inputs (Liu, Li, Wu, & Lee, 2023; Liu, Li, Li, & Lee, 2023). It integrates a vision encoder with a language model, demonstrating impressive performance in tasks like multimodal chatting and Science QA. LLaVA's approach represents a significant advancement in AI, showcasing the potential for more sophisticated and nuanced AI interactions that can process and respond to a combination of visual and textual information.

2.4 Latent Dirichlet Allocation

The main resources that Alex used were the Gensim library documentation of its implementation of the Latent Dirichlet Allocation (LDA) model which is an unsupervised topic model algorithm and the SentenceTransformers library using a pre-trained Sentence-Bert (SBert) model for caption embeddings (Blei, Ng, & Jordan, 2003).

3 Database Development (Lihao)

3.1 Database initialization

Our data resources were previously scattered, which was a challenge for our project team to access the desired dataset conveniently. For example, data from caption contests were released individually for each cartoon. To obtain data across multiple contests, we had to download numerous CSV files and merge them manually. Additionally, extracting specific columns required repetitive efforts.

In response to this problem, I designed a structured MySQL database to centralize all our data from various sources. I meticulously organized and inserted the data into this database. Consequently, we eliminated the need to download files repeatedly. we can now precisely select the data we need, including specific attributes and contest numbers, from the database directly by using SQL queries. we don't need to suffer from managing scattered files.

Once the database was set up on my local machine, we encountered an issue when other team members needed access. They were unable to connect to the database created on my laptop, despite numerous attempts to deal with the problem. After exploring various solutions, I identified a viable workaround. We decided to use AWS (Amazon Web Services) as the server for our database. I initiated this by creating an empty MySQL database on AWS and then migrated our existing data into this cloud-based database. As a result, every member of our team can now freely connect to and interact with the database, ensuring smoother collaboration and data access.

The introduction of current database structure will be attach in the appendix.

3.2 Database updating

The ongoing nature of The New Yorker Cartoon Caption Contest determines that regular data maintenance, including cleaning and updating the database is needed. I perform these updates monthly, with the most recent contest entry in the database being number 863. Having new captions, I also create descriptions for the cartoons featured in the contests by specific models and integrate these into the database.

In addition to routine updates, I used to include results from my teammate into the database. Topic models and word embedding models were generated by my teammate for contest data. These models assign each sentence or word as a multidimensional vector. Vectors, as arrays, cannot be inserted directly into database, so I converted these vectors into blob type data to bypass this problem. Furthermore, I built up code functions to decipher blob-type of data, and revert it to vector form again.

3.3 Describing Cartoons by LLaVA

To enhance our analysis and develop our AI assistant interface, we have meticulously gathered descriptions for cartoon images from contests 2 through 763 from the research "Do Androids Laugh at Electric Sheep?". To expand our database with descriptions for cartoons beyond contest 763, I employed the Large Language and Vision Assistant – LLaVA. This model had good performance in producing precise and comprehensive descriptions for these more recent cartoons. I took the initiative to carefully create and refine the prompts used in LLaVA, guaranteeing that the resulting descriptions are of high quality before I inserted them into our database.

4 Experiment 1: GPT response on humor rating (Zhuolun + Lihao)

4.1 Introduction

Humor is closely related to implicit information, making it challenging to capture solely from literal expressions, such as sarcasm. Understanding humor can refer to two different scenarios. First is understanding why something is humorous and logically deducing the connection between implicit and literal information. Second is quickly perceiving implicit information

and emitting laughter. The former is structured and reflective, while the latter is intuitive and reactive (Levine & Redlich, 1955). Explaining why something is funny does not necessarily elicit laughter, or at least elicit lesser laughter compared to the extent of an intuitive response. Previous research has explored understanding humor in the first scenario, explaining why something is funny (Hessel et al., 2023), but whether this constitutes a genuine understanding of a joke is debatable. Understanding humor in the second scenario is more common for humans, and not all forms of humor involve language or can be accurately interpreted (such as performance-based humor). Even for the same humorous content, not all explanations are entirely consistent in their interpretation.

The first experiment tests the current AI's performance in direct humor judgment tasks. Considering that AI's performance in the context of why funny is not as good as that of humans, this experiment assumes that AI may perform even worse in humor judgment tasks.

4.2 Method

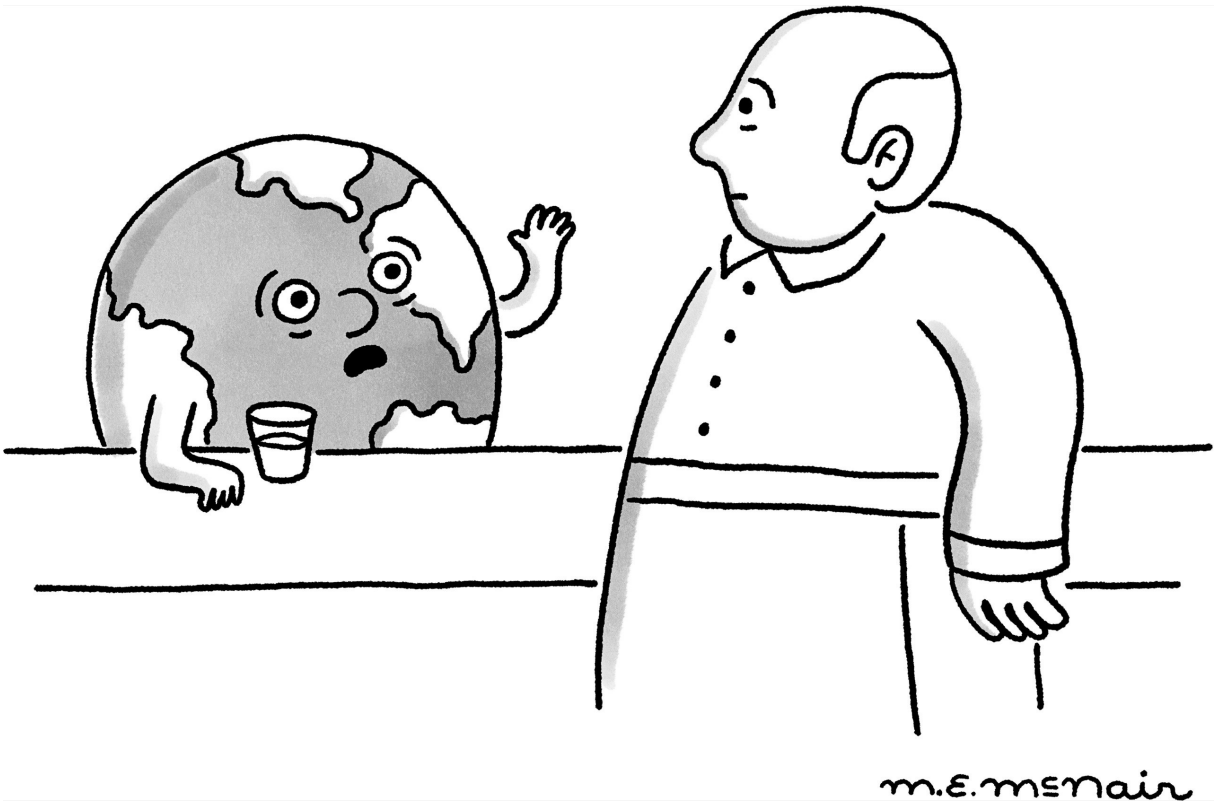

Figure 1: Cartoon contest 866

This experiment used data from contest number 866 (Figure 1). Each humor label was randomly split in a ratio of 2:8, with 20% forming the test set and 80% forming the training set. For example, ten of fifty funny captions would be placed in the test set, while the remaining forty would be placed in the training set. Although not every test would necessarily use the training set and the extent to which it was used varied, each test would employ the exact same test set.

The entire humor dataset is huge, so we chose the number 866 contest as the primary one for several reasons. Firstly, older matches may already be included in GPT's training data, so we opted for more recent ones. Secondly, computer vision was a bottleneck for model performance (Hessel et al., 2023), but today, the model's vision has made significant progress. We want to test the model's performance with original image inputs like real humans. The content of cartoon number 866 is concise, and in testing, GPT-4's descriptions perform exceptionally well, accurately covering all key elements. Third, the number 866 contest has a substantial number of rating votes (over one million), making the funniness data more reliable.

The experiment tested GPT-3.5 and GPT-4. The task involved requiring the models to make judgments based on selections from three funniness labels (funny, somewhat funny, unfunny). The textual descriptions of the cartoons were generated by the GPT-4, and human inspection confirmed the accuracy of the content.

For GPT-3.5, two approaches were taken: using the API with direct prompts and training a fine-tuned model then prompts. Several fine-tuned models are trained with different data. As for GPT-4, five approaches were taken (Figure 2): description input prompts, image input prompts, description input prompts with training data, image input prompts with training data, and prompts with no cartoon content. The website interface was used to upload data CSV files for prompting.

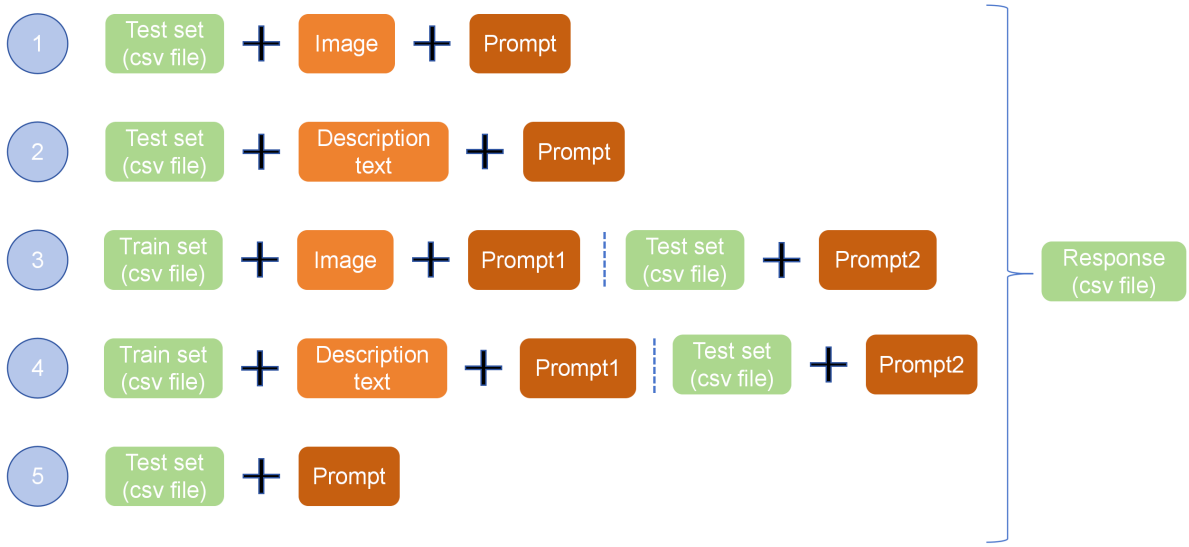

Figure 2: The processes of tasks

Comparing the generated results with the actual labels based on mean value will yield accuracy. The analysis will consider various scenarios to understand why accuracy might differ.

4.3 Results

GPT-3.5's direct response is always "somewhat funny". Due to the model's tendency to get stuck with repetitive prompts, the accuracy across the entire test set is inconclusive.

GPT-3.5 fine-tuning model response highly depends on the labeling ratio in training data. The different ratio combinations in the training set will be expanded and reflected in the results of the test set. For example, a ratio of 1:3:6 (funny, somewhat funny, unfunny) in the training set would lead to results in the test set of 0:0:1.

GPT-4 exhibits different outcomes under two input conditions. After repeated testing, the performance mean and standard deviation of text description input are significantly worse than those of image input (Table 1 and 2). No significant performance changes were observed in either input condition during the training condition compared to no training condition (Table 3 and 4). The control group, which had no cartoon input, performed the worst (Tbale 5). However, all GPT-4 response tests only show significant accuracy in labeling responses as unfunny, with GPT-4 rarely responses as funny or somewhat funny.

time	funny	somewhat funny	not funny	F1 score	method
1	0.20	0.12	0.89	0.85	heuristic
2	0.20	0.06	0.97	0.88	subjective
3	0.20	0.12	0.89	0.84	heuristic
4	0.40	0.04	0.84	0.81	subjective
5	0.20	0.10	0.89	0.84	heuristic

Table 1: The accuracy of image input prompts

GPT-4 also report its method on humor judgement task:

- **Heuristic model frame**

  In this frame, ChatGPT-4 employed a simple heuristic method based on the specific context and keywords relevant to the material I provided. ChatGPT-4 generated two lists of keywords: one for 'funny' and another for 'somewhat

time	funny	somewhat funny	not funny	F1 score	method
1	0.30	0.32	0.36	0.38	subjective
2	0.30	0.12	0.89	0.85	heuristic
3	0.30	0.34	0.31	0.43	subjective
4	0.30	0.07	0.98	0.89	heuristic
5	0.40	0.15	0.77	0.78	heuristic

Table 2: The accuracy of description input prompts

time	funny	somewhat funny	not funny	F1 score	method
1	0.10	0.16	0.84	0.82	heuristic
2	0.50	0.13	0.79	0.79	heuristic
3	0.40	0.04	0.84	0.82	heuristic
4	0.10	0.10	0.78	0.78	heuristic
5	0.40	0.00	0.89	0.84	heuristic

Table 3: The accuracy of image input prompts with training set

funny' terms. These two lists derived from the context provided. The model then evaluated the captions by checking how well the words or phrases in each caption aligned with the keywords in these two lists. A caption was considered "funny" if it contained a word from the funny keyword list or something closely related to it. Similarly, a caption was tagged as "somewhat_funny" if it included a word from the somewhat-funny keyword list or a closely related term. Captions that didn't contain or relate closely to any word from these two lists were classified as 'not_funny'.

- **Subjective model frame**
  This model frame added some methods including subjective factors, based on the heuristic model frame. These methods could be cultural analysis and linguistic factors analysis. While the specifics of cultural analysis are not clear, the linguistic factor analysis involves identifying language features typically linked with humor. This includes searching for linguistic traits like wordplay, irony, exaggeration, or the use of rhetorical questions.

4.4 Discussion

From the performance of GPT-3.5, it appears that it has no sense of humor. Its performance is even worse than random guessing because it responds in a fixed pattern (all somewhat funny or pattern from training data), ignoring the content of the task.

The performance of GPT-4 can only indicate that it knows what is unfunny. Unlike GPT-3.5, GPT-4 doesn't limit its responses to unfunny options or get influenced solely by the training set. It attempts to respond with funny options, showing a weak level of accuracy in doing so.

In any case, AI has not demonstrated an ability to judge humor close to humans, and there is considerable controversy over whether it even approaches human performance. These results align with expectations. Firstly, research suggests that humans have specific neural foundations for humor (Yamao et al., 2015). Secondly, information involving metaphors cannot be learned directly.

The most controversial aspect lies in whether GPT-4 truly understands what is unfunny or can predict what is unfunny. Knowing what is not does not guarantee an understanding of what is. Knowing what is not a unicorn doesn't ensure an understanding of what a unicorn is. Moreover, there are differences in the learning processes between large language models and humans. Therefore, after comparing the performances of GPT-3.5 and GPT-4, it is suggested that GPT-4 has developed the ability to know what is unfunny but still has no clues about what is funny. We believe that the critical aspects present in humans, which current AI designs have failed to consider, are responsible for these results. In the subsequent conclusion sections, I will explore this missing aspect: creativity.

4.5 Future work

The progress of AI is advancing rapidly, making it worthwhile to regularly test its humor judgment capabilities. The funniness is calculated from voting data. For an individual human, the experiment of determining how much accuracy can be achieved in the same judgment task is worth exploring. In the future, recruiting participants for genuine human-AI

time	funny	somewhat funny	not funny	F1 score	method
1	0.10	0.40	0.40	0.52	subjective
2	0.40	0.40	0.35	0.48	subjective
3	0.10	0.09	0.97	0.88	heuristic
4	0.00	0.00	0.95	0.86	heuristic
5	0.10	0.12	0.85	0.82	heuristic

Table 4: The accuracy of description input prompts with training set

time	funny	somewhat funny	not funny	F1 score	method
1	0.00	0.00	0.99	0.88	heuristic
2	0.20	0.54	0.11	0.18	heuristic
3	0.00	0.29	0.59	0.66	subjective
4	0.00	0.82	0.13	0.22	subjective
5	0.00	0.24	0.51	0.60	heuristic

Table 5: The accuracy of no cartoon input

competitive testing is considerable.

There is a distinction between the textual descriptions of images and original images as information input for humans. Recruiting participants to test the funniness means voted by humans in situations where only textual descriptions are provided would be interesting. However, before committing to visual input, AI's visual capabilities still require further exploration and testing. Skepticism should be maintained when using complex hybrid models to ensure the model's performance and output reliability.

5 Experiment 2: Human response on humor theories (Zhuolun)

5.1 Introduction

From Experiment 1, it can be concluded that large language model AI still exhibits a significant gap compared to humans in humor judgment tasks. To comprehend this disparity, Experiment 2 shifted to humans, attempting to uncover clues from the cognitive information processes involved in humor. I delved into textbooks on humor psychology; there are three main contemporary theories of humor, each with its own explanatory power and limitations (Martin & Ford, 2018).

The first is reversal theory, which can be described as three hypotheses. (1) A person must be in a playful, paratelic motivational state. (2) The humor event must produce an increase in arousal experienced in the paratelic state as fun or excitement. (3) A person must experience a cognitive synergy in which the second interpretation of a stimulus or event involves diminishment from the first interpretation (Apter, 1982). When all three conditions are met, humor occurs.

The second is comprehension-elaboration theory, which delves into specific cognitive processes in minds. It argues that understanding humor is a three-step process: interpretation, incongruity, and reinterpretation. It proposes that the amount of humor one experiences as a result of these basic comprehension processes depends on (1) the degree to which reinterpretation of an event diminishes the importance or value of the event, (2) the type and amount of cognitive elaboration that one generates in response to the reinterpretation, and (3) the degree to which the humor event is difficult to comprehend (Wyer & Collins, 1992).

The third is the benign violation theory. The central proposition of benign violation theory is that to experience humor, one must (1) interpret a stimulus or event as a violation, (2) interpret the event as benign or harmless, and (3) hold these two interpretations simultaneously (McGraw & Warren, 2010). Essentially, a violation is anything that somehow threatens a person's view of how things should be (McGraw, Warren, & Kan, 2015).

Contemporary theories of humor remain controversial since neither is universal enough for humor. For example, empirical studies of the benign violation theory only prove that jokes meeting benign violation conditions are funnier (McGraw & Warren, 2010). Besides, the theory is vulnerable to the circular reasoning of post hoc (after-the-fact) explanations (Martin & Ford, 2018). While the comprehension-elaboration theory lacks strong empirical research, another study suggests that the easier a joke was to understand, the funnier it was rated to be (Derks, Staley, & Haselton, 1998). The reversal theory

is relatively conservative; although it has support from neuroscience (Yamao et al., 2015), it lacks an in-depth exploration of cognitive processes and cannot explain the semantic understanding involved in humor (Wyer & Collins, 1992). Overall, no ideal theory can be directly followed and applied to improve AI or model performance. Therefore, the question posed in Experiment 2 is: Do people need to explicitly undergo the processes and conditions described by the theory to experience humor from The New Yorker cartoon caption contest?

5.2 Method

Experiment 2 adopted an online survey format, collecting data through Prolific. Participants were asked to watch a series of jokes created by combining cartoons with captions, and for each joke, they responded to a series of questions related to humor theory. Each participant watched a total of 5 jokes. Due to many captions in the original dataset being not funny, this experiment randomly selected captions based on humor-level labels to create a dedicated pool of jokes. This pool had an equal number of funny, somewhat funny, and unfunny captions. The jokes encountered by participants were randomly selected from this pool.

I created five questions based on humor theories. The first and second questions are related to the hypotheses of the reversal theory, corresponding to the states and arousal. The third and fourth questions relate to the hypotheses of the benign violation theory, corresponding to the violation and benign (accept). For violation, six aspects were measured: physical threat, identity threat, illogical, contrary to expectation, bad, and incorrect. The last question involves whether participants experienced two different interpretations (jump), a point emphasized by all three theories. No specific questions targeted the comprehension-elaboration theory because the processes described by the comprehension-elaboration theory are challenging to incorporate into a survey. In pre-testing before the survey was released, feedback indicated that the question created by the comprehension-elaboration theory was too complex and confusing for participants. Participants were required to evaluate the subjective funniness they experienced on a scale of 1 to 100.

After collecting the survey data, linear regression was employed to test theories.

5.3 Results

A total of 30 participants were recruited, but two had severe data missing issues and were excluded. In the end, feedback for a total of 140 jokes was obtained.

The first model for the reversal theory: lm(funny˜state*arousal*jump). The only significant coefficient is arousal.

The second model for the benign violation theory: lm(funny˜violation*accept*jump). The only significant coefficient is accept.

The third model for part of the comprehension-elaboration theory: lm(funny˜jump). The coefficient for 'jump' is significant.

5.4 Discussion

The reversal theory did not receive confirmation in the experiment. All three predictors in the first model should have been significant, but only arousal showed significance. Furthermore, there are doubts about the significance of arousal because arousal is a physiological concept that cannot be directly measured in a questionnaire survey. As a substitute, subjective responses with a simple question were employed. The failure of the reversal theory may be attributed to criticisms of the comprehension-elaboration theory: the reversal theory lacks an assessment of funniness level (Wyer & Collins, 1992). In other words, the reversal theory did not account for differences in the funniness levels of different jokes. Therefore, if the survey's method of measuring humor is limited to choosing between funny and unfunny, the data results may vary. Overall, the shortcomings of the reversal theory, which lacks an in-depth exploration of cognitive processes, have been supported, as evidenced by the lack of significance in the data concerning humor levels.

The benign violation theory did not receive confirmation in the experiment. All three predictors in the second model should have been significant, but only accept showed significance. The survey did not directly inquire about the term "benign" but instead used "acceptable" to avoid confusion. For jokes, a bias toward "acceptable" with a high rating is predictable. The lack of significance in violation as a predictor might be because cartoons and captions in The New Yorker Cartoon Caption Contest rarely include elements of violation. As a mainstream magazine with cartoon content, caption creators may try to

avoid aggressive elements. This result supports criticisms of the benign violation theory, suggesting that benign violation is an additional, unnecessary prerequisite for humor.

The comprehension-elaboration theory asks for a more complicated response on humor rating, and the "jump" predictor from the third model failed this requirement. Therefore, as criticized, the comprehension-elaboration theory lacks empirical research evidence and is challenging to investigate in experimental studies (Derks et al., 1998). The significance of the "jump" predictor in the third model suggests that the cognitive processes proposed by the comprehension-elaboration theory may exist. However, the specific cognitive processes might happen too quickly to be consciously realized.

In summary, the results of Experiment 2 support criticisms of the three contemporary humor theories. Regarding the research question, the results answer with no. Concerning the initial research motivation, the results do not offer an ideal next step. The mystery and difficulty of humor, revealed in Experiment 1, are reinforced in Experiment 2.

5.5 Future work

Based on the research question posed by Experiment 2, more detailed questions and surveys could be conducted in the future. For example, a test of the reversal theory using only funny and unfunny options. Another possibility is a content analysis-based test of the comprehension-elaboration theory, relying on descriptive responses.

Additionally, having AI attempt a similar survey could be intriguing, but the reason for doing so needs to be clarified. Overall, the most crucial aspect of Experiment 2 lies in the extensive literature review, providing an overarching understanding of existing humor research and outcomes. Humor lacks a unified theory; the so-called contemporary theories seem antiquated from a psychological perspective. In the future, attention should be focused more on conceptualizing new theories.

6 Experiment 3: Machine learning humor level prediction (Zhuolun)

6.1 Introduction

According to Ritchie, most of the existing humor theories are too vague and imprecise for computational application(Ritchie et al., 2007). This provides some support for the failure in Experiment 2. The results of Experiment 1 also failed to capture any clues to humor successfully. Therefore, the new question is: Can machine learning methods accurately predict humor solely from captions? The curiosity for Experiment 3 arises from the question of whether there is some pattern of humor in linguistic information when cognitive processes are ignored.

6.2 Method

Considering that there are too few funny captions in a single contest to provide sufficient data for model training, data from contests 850 to 870 were used. Subsequently, captions with different labels were extracted in a 1:1:1 ratio and separated into the training and test sets.

Two machine learning methods were employed: word embeddings and neural networks. Since the original data is linguistic information, the models utilized the training set directly for word embeddings. For the neural network method, a tool called LIWC-22 (Tausczik & Pennebaker, 2010) was used to transform captions into over a hundred features, and the original captions were removed. The neural network model was trained using these features.

Model parameters were adjusted, generating multiple models for both methods. The results will be evaluated based on the accuracy of each label and the overall accuracy.

6.3 Results

Despite noticeable variations in the accuracy of the three labels under certain parameter settings, the overall accuracy hovers around the level of random guessing (33%) for both methods in all models.

6.4 Discussion

Figure 3: Cartoon contest 523

From the results, it can be deduced that machine learning has failed to learn any underlying patterns from the data. At least linguistically, there is no discernible pattern of humor in captions.

Upon closer inspection of the data, it is observed that similar captions exhibit different levels of humor (Table 6). The following examples with their funniness mean are from Contest 523 (Figure 3):

Table 6: Captions with mean

captions	mean
You ordered from the wrong Amazon.	1.8912
I think I ordered from the wrong Amazon.	1.8098
we might have ordered from the wrong Amazon.	1.6794
I believe you ordered from the wrong Amazon.	1.6656
Looks like I ordered from the wrong Amazon.	1.5959
honey i think you ordered from the wrong Amazon.	1.5263

These six captions are nearly identical in content, with the remaining differences mostly consisting of some stop words (common words considered for removal in word embeddings). However, the humor level is significantly different among these six captions, and human intuition aligns with their respective rankings. For example, "I believe" is redundant and boring, "You" sets a better context than "I think". Clearly, neither word embeddings nor neural networks can capture the significant impact of these subtle differences. Because the captions in a single contest are always centered around the cartoon, there will inevitably be content similarities among captions. Considering that in Experiment 1, GPT-4's performance was the worst when there was no cartoon input, the results of Experiment 3 aligned with expectations. The cartoon content as the context plays an important role in the sense of humor.

6.5 Future work

Continuing similar explorations in machine learning in the future is not recommended.

7 Experiment 4: Incongruity test (Zhuolun)

7.1 Introduction

Back to humor theories, the experience of humor appears to be predicated on two cognitive-perceptual processes activated by characteristics of a humor stimulus and the social context in which it is encountered: (1) perception of incongruity and (2) appraisal of incongruity in a nonserious humor mindset (Martin & Ford, 2018). While the first characteristic is confirmed by all three theories and supported by the third model from experiment 2, the second characteristic becomes controversial for different theories and did not receive support from experiment 2. The reversal theory describes the humor mindset as a mental state. The comprehension-elaboration theory describes the humor mindset as a complex understanding process. The benign violation theory describes the humor mindset as a benign attitude towards violations. I argue that the humor mindset should not be treated independently for two reasons. Firstly, mindset is a concept that lacks a precise definition and cannot be easily quantitatively measured. Secondly, both the perception of incongruity and the humor mindset are subjective to humans; a subjective experience (humor mindset) based on another subjective experience (perceptive of incongruity) is redundant. Therefore, after conducting the previous experiments and reviewing the literature, I believe a fusion of the two characteristics would be worthwhile. In other words, a nuanced humor mindset may not be necessary, and perceiving incongruity in specific patterns may be sufficient to trigger a humor experience.

The discovery and navigation of possibilities involve more than the simple, separate representation of particular future or action-outcome consideration (Poulsen & DeDeo, 2023). We are often called upon to hold multiple, incompatible possibilities in mind and prepare for them in action, an ability that appears early in childhood and may well be unique to our species (Redshaw & Suddendorf, 2016). This ability is also described as representing a "matrix of maybe" (Baumeister, Maranges, & Sjåstad, 2018). Works have been dedicated to understanding how humans navigate and constrain the possible futures into this matrix (Phillips, Morris, & Cushman, 2019; Kvavilashvili & Rummel, 2020; Cole & Kvavilashvili, 2021; Sjåstad & Baumeister, 2023). When there is a noticeable disparity between reality and future matrix, incongruity experience arises. Incongruity can trigger various responses, including fear, confusion, surprise, and more, with humor being just one of the possibilities. Under the view of the future matrix, a fusion of the two humor characteristics becomes possible in a new direction. While reactions to incongruity are diverse, various responses may exhibit distinct matrix patterns, such as differences in predicting probability ranges or variations in predicted content.

Given large language models' ability to produce human-like text (Bail, 2023; Dillion, Tandon, Gu, & Gray, 2023; Binz & Schulz, 2023), they can serve as tools to check the future matrix to address incongruity patterns for humor in a given context (cartoon). The situation of different levels of humor produced by similar texts mentioned in Experiment 3 could potentially yield significant results when using large language models as tools. Experiment 4 involves using a large language model to generate token probabilities and includes three tasks: detecting specific probability ranges for punch lines in captions; comparing the changes in probabilities with and without context (description of cartoons); finding a computational model that can be used to predict humor.

7.2 Method

Due to limitations in technology, resources, and time, GPT-2 was utilized for the experiment. GPT-2 processes token IDs and generates token IDs, and it is convenient to get its prediction probabilities on all token IDs. Perplexity is calculated based on the token's actual probability in the context and GPT-2's prediction probabilities for all toke IDs in that position.

The method involves manually examining the perplexity in captions. First, the focus is on whether the perplexity generated by GPT-2 aligns with human perception. Second, the focus is on the changes in perplexity when there is no context. Third, the focus is on attempting to summarize patterns through observation.

7.3 Results

In the case of the simplest humorous captions, the perplexity generated by GPT-2 broadly aligns with human perception. However, the specific values do not accurately predict differences in humor levels, as mentioned example in Experiment 3 (Table 6). For instance, the perplexity of the punch line "Amazon" does not reflect the correct ranking of humor level, let

alone consider a linear relationship.

Additionally, GPT-2 sometimes shows greater perplexity for predicates than punch lines. For humans, the predicate should be the part that sets up the expected context, generating a noticeable but not overly pronounced level of perplexity. In the case of unfunny captions, GPT-2 sometimes displays perplexity patterns similar to those seen in funny captions but entirely inconsistent with human perception. Most importantly, due to the instability in perplexity with GPT-2, observing any pattern differences between unfunny and funny captions is impossible.

A similar situation arises in context detection. Context can sometimes enhance the alignment between GPT-2 perplexity and human perception when dealing with simple cautions. However, this result is inconsistent, and the data becomes uncontrollable when facing slightly more complex captions.

The experiment did not extend to a larger dataset because the performance of GPT-2 could not help yield a computational model suitable for the hypothesis.

7.4 Discussion

Experiment 4 did not produce results that highly align with human perception, supporting the viewpoint that new ways to explore the latent spaces of human possibility need to be tempered with a certain conservatism (Poulsen & DeDeo, 2023). LLMs cannot replace ordinary psychological experiments (Dillion et al., 2023) or correct their faults (Poulsen & DeDeo, 2023). While the results of Experiment 4 can be largely attributed to the performance limitations of GPT-2, it is important to note that large language models and humans explore the space of possibilities in different ways (Poulsen & DeDeo, 2023). Caution should be exercised when using large language models to develop computational models for humor.

7.5 Future work

Considering the explanatory power of the future matrix and the partially aligned results with human perception in Experiment 4, much work still needs to be done in further researching humor from this new perspective. However, the emphasis should return to human participants. It must be emphasized that in psychology, the focus always remains on humans, and humor is a uniquely human phenomenon. Future work should remember the subject of study and the centrality of humans in understanding humor.

It's possible to revisit perplexity detection tasks using more advanced large language models in the future. However, as mentioned in the discussion section, investing excessive effort in this direction may not be necessary.

Subsequent tasks could involve similar testing on humans, having participants provide perplexity ratings word by word, followed by an overall assessment of humor and recording descriptive psychological processes.

8 Conclusion of Experiments (Zhuolun)

8.1 Background

Experiments 1 and 2 showed me that solely focusing on humor or relying on AI has limitations. Therefore, I chose to take a cognitive psychology course. In the course, my emphasis has been on exploring cognitive architectures and unique human cognitive abilities to find the cognitive foundations of humor. During this process, I narrowed down the topic to creativity. In the conclusion section, I will integrate findings from four experiments to discuss the relationship between humor and creativity and potential future research paths.

8.2 The puzzle of creativity

Although the conceptual discussion of creativity should ideally fall within the realm of philosophy, creativity has yet to receive widespread attention from philosophers, even within aesthetics, the field most closely related to creativity (Gaut, 2010). On the other hand, in response to J. P. Guilford's claim in his 1950 APA presidential address (Simonton, 2000), psychology has generated substantial research on creativity. However, these efforts tend to be discrete, specialized subfield studies, resulting in a fragmented landscape and a lack of coordinated, unified knowledge construction (Glăveanu, 2014).

This situation arises because creativity is an exceptionally challenging topic. What is creativity? The most commonly used definition in psychological experiments is "the production of effective novelty" (Cropley, 1999; Lubart, 2001; Mumford,

2003; Plucker, Esping, Kaufman, & Avitia, 2015). However, defining creativity in terms of novelty and value (Weisberg, 1993) needs to be revised. Firstly, novelty is a relative concept over time; something novel at one point will not be eternal, as novelty tends to diminish over time. If novelty is adopted as the defining criterion for creativity, it would classify almost everything as creative (Hausman, 1979). Secondly, substituting originality for novelty is problematic because creative productions are not generated out of thin air but have strong connections to existing known entities, making the line between original and pre-existing content blurry (Runco, 2023). Lastly, value is a retrospective judgment, and the value of a creative production cannot be assessed before its existence in time.

I argue that the premise for discussing human creativity involves accepting indeterminism, meaning that causes do not constrain the future to a single path. It implies uncertainty and probability, which ensure the novelty. In contrast, determinism implies that humans are not creating but merely producing. Considering the uncertainty principle in quantum mechanics, indeterminism in the physical world might also be true, but I will focus on human cognitive indeterminism to avoid unnecessary discussion. Human creativity is an incremental process, introducing a certain degree of novelty based on existing entities. Therefore, I define creativity as "the production of randomness against predictive model at range from 10 percent to 30 percent". The numerical range is speculative and resolves the blurry boundary issues; falling below it suggests a lack of originality, while exceeding it implies insufficient association with existing entities. The term "randomness" emphasizes the universality of creativity, resolving the timeliness issues of value and novelty and excluding the limitation of context. The term "predictive model" emphasizes the relativity of creativity, which can be relative to subjective cognition or relative to existing entities.

The definition of creativity is controversial; hence, it is reasonable to argue that AI lacks genuine and significant creativity. As mentioned by Simon (1996), there are two assertions about computers: A simulation is no better than the assumptions built into it; A computer can do only what it is programmed to do. AI has not taken creativity into account in its cognitive architecture, so it lacks creativity.

8.3 From creativity to humor

Humor is immediate, highly context-dependent, and closely tied to human creative abilities (Martin & Ford, 2018). It is obvious that producing humor requires a high level of creativity. However, I argue that appreciating and experiencing humor also relies on humans' significant creativity. This is why humor is a unique phenomenon exclusive to humans.

Experiments 2 and 4 show that human subjective cognition plays a significant role in humor judgment. Similarly, this applies to creativity. People's judgments about creativity are based on subjective predictive models. What truly associates creativity and humor is the probability pattern behind the prediction. For creativity, too much randomness beyond predictive models is considered unrelated to existing entities. In the case of humor, excessive incongruity beyond expectations is seen as perplexing. Both creativity and humor necessitate the subjective experience of new stimuli in a certain range of probability. Both creativity and humor also experience decay in sensitivity (novelty and funniness) over time. While experiencing creativity does not guarantee the elicitation of humor, appreciating humor does require the perception of creativity. This is why using logic to understand humor does not necessarily evoke laughter, as logic seeks confirmed and reasoned outcomes incompatible with creativity's randomness. This also explains the results of Experiments 1 and 3. AI lacks the perception of creativity, making it unable to discern funny captions and rarely make funny judgments. Humor relies on novel content; therefore, machine learning has no definite pattern to master. From an evolutionary perspective, the effects of humor may serve as a foundation for further enhancing human creativity, providing positive feedback for the release of more creativity in humans. Future experiments could measure the extent of the association between an individual's subjective sense of humor and creativity. However, great caution and attention must be exercised in experimental design, as both humor and creativity have controversial definitions, making it challenging to ensure that subjective surveys collect the desired data.

How does humor differentiate itself from other responses to incongruity and creativity in terms of the future matrix? Early studies often considered aggression crucial to humor (Martin & Ford, 2018), and the benign violation theory emerged as a response to aggression (McGraw & Warren, 2010). However, the role of aggression in humor is likely to be social context-dependent rather than essential. Imagine in a desperate context where someone suddenly makes absurdly optimistic remarks, triggering a sense of humor among those around. In this thought experiment, humor arises from a situation opposite to benign violation. Optimism is considered morally encouraged, but in a desperate context, people cannot truly accept such encouragement, leading to laughter as a response. However, humor is not solely about reversal; forms like puns do not involve extreme reversal relationships. Humor takes on various forms, so it is advisable to differentiate between different forms of humor before further dissecting the cognitive processes behind humor. I believe that important factors in controlling a sense of humor include response time and attention. For instance, in appreciating the creativity of aesthetics, humans need highly focused attention, retrieving known entities from memory and expressing a sense of creativity after a certain period. In

contrast, humor occurs rapidly and only requires humans to invest a little attention in the context. These factors could be considered in future human participant tests to gain a deeper understanding.

The psychological functions of humor can be classified into three broad categories: (1) emotional and interpersonal benefits of mirth, (2) tension relief and coping, and (3) social functions in group contexts (Martin & Ford, 2018). I suggest that considering the relationship between humor and creativity, humor plays a role in enhancing human creative potential. Creativity involving higher-level cognitive abilities may not necessarily overlap; for instance, aesthetic creativity may not apply to scientific creativity. However, appreciating and producing humor are foundational creative activities, and they may have predictive value for various forms of higher-level cognitive creativity. Furthermore, I advocate that attempting a detailed predictive model for humor is futile, much like how humans cannot predict what humans will create, or similar to the fading novelty of creativity over time, theories of humor are destined to maintain a certain level of ambiguity to adapt to future developments. The evidence of humor observed from human experience is likely contingent on social environments and cannot serve as a permanent assertion for humor.

8.4 The thought experiment of twin

Twin Earth is a thought experiment meant to serve as an illustration of semantic externalism. It is proposed by philosopher Hilary Putnam in his papers "Meaning and Reference" (1973) and "The Meaning of 'Meaning'" (1975). Semantic externalism is the view that the meaning is determined, in whole or in part, by external factors to the speaker. Here is the detail of Twin Earth thought experiment: If there were a twin Earth in the universe where everything is identical to our known Earth, including its inhabitants, the only difference would be that the water on this twin Earth would have a different chemical composition from the water on our Earth. Now, if a person on Earth and his identical counterpart on the twin Earth simultaneously say "water", do they mean the same thing?

Twin Earth implies the answer is no to support the semantic externalism. Here is the counter version of the thought experiment: If there is a pair of genetically identical twins in a room, with their current physical states also being identical, and after hearing a joke, one laughs while the other doesn't, is the meaning of the joke the same for this pair of twins?

I do not intend to engage in philosophical discussions, but I would like to use this example to emphasize the significance of subjective mental activities in humor. Without considering the individuals' perspectives, the humor analysis is incomplete.

9 Caption Embeddings and Topic Modeling (Alex)

9.1 Topic Modeling

When starting my part for this project, I first watched some video tutorials on how to perform NLP techniques in Python. At this point in time, I only knew how to do NLP in R. After watching the videos, I then researched what topic models were and how to build one. What is a topic model? Imagine you have a large collection of documents—these could be articles, captions, or any written content. Now, the goal is to understand what each document is about without reading all of them one by one. A topic model is like a smart assistant that helps you automatically discover the main themes or topics hidden in your pile of documents. It does this by looking at the words used in the documents and figuring out which words often appear together. It is also an unsupervised machine learning algorithm which looks at a pile of data and tries to find interesting connections or groups without having someone guide it step by step. It's about letting the model explore and discover patterns on its own. In essence, it tries to find patterns in the words used across all documents.

An example of topics in a collection of documents would be if your documents are about animals, the topic model might identify groups of words like "lion," "tiger," "jungle," and "roar" that frequently appear together. It would then say, "These words seem to form a 'wildlife' topic or 'animal' topic. The most useful aspect of a topic model is that it can discover multiple topics within your documents, even if you didn't know they were there.

After finishing all the tutorials and documentation, I decided to build a topics model using the LDA algorithm because I felt that it was the simplest model while also being the most interpretable. A LDA model is a type of topic model that allows a document to be about multiple topics. So, a document on a safari might have words from both the "wildlife" and "travel" topics. Additionally, it assigns probabilities. It says, "This document is 60% about wildlife, 30% about travel, and 10% about Africa." Some topic models only assign one document per topic which may hinder the understanding of what the topics are truly about, which is why I chose the LDA model to work with.

I also decided to use only a single contest which contained 7500 captions for this model to save time, but the steps for building the model can be applied when using a larger data set. In order to build an LDA topics model, the raw text has to be preprocessed and tokenized first to be used as its corpus otherwise there is too much noise getting in the way of creating meaningful topics. Before preprocessing the text, I had to decide on how to remove my stop words using either NLTK or Spacy's library of stop words. I ultimately chose to use Spacy's version because it has a larger collection of common stop words which can remove more noise for the model. After choosing my stop words library, I pulled the data from our SQL database using the mysql-connector python library, which consisted of the caption text column and the ranking column into a Pandas dataframe. Since I only want my data to be text, I dropped the ranking column which contains numbers that are not meaningful in NLP.

Before building the LDA model, the most important step is to preprocess the raw text data. This is because the model is an unsupervised learning approach that takes in the data that it is offered and sends back results based on that input. What this means is that the model is inherently dependent on the quality of data that it is given and that the preprocessing steps beforehand are very important in extracting good quality of topics that are clear, segregated and meaningful.

After making my data frame only contain text, I began to prepreprocess my text data by first removing punctuations and lowercasing all words to make a standardized list of words. Afterwards, I tokenized and deaccented my words to ensure that the same words have the same meaning and are not interpreted incorrectly by the model. Before moving on to the next steps of preprocessing, I removed stop words using the Spacy library since I did not want my next steps to account for these words. Within the data, there are bigrams which are essentially pairs of consecutively written words which I need to account for. To account for bigram and trigrams, I created a function that found repetitive pairs of consecutive words that appeared at least five times in the dataset. I chose the threshold of five because I don't want words that coincidentally appear next to each other or rare pairings to be designated as brigram phrases by the function.

Now that the pre preprocessing steps have been completed, I started to create the model's id2word dictionary and corpus. The id2word dictionary is a collection of captions where each unique word in each individual caption is assigned a unique numeric ID. This is used to map between words and their corresponding IDs. The corpus represents the captions as a list of (word_id, word_frequency) tuples, where word_id is the numerical ID assigned to the word by the id2word dictionary, and word_frequency is the number of times the word appears in the caption.

Once I had the id2word dictionary and corpus created, I used the multicore version of the model because it uses all available CPUs to parallelize and speed up training. I built the LDA model using several parameters. The first parameter is the number of topics of which I put as 20 since the number does not matter as later on I will find the best number of topics in a more precise manner. I also put a random seed in the model in order to have the same results reproduced as the model uses randomness in its training and inference steps. When deciding the chunksize and passes parameters which are how many documents are used in each training chunk and the number of times the model passes through the corpus during training, I chose 100 and 10 respectively. This is because having a smaller training size and number of passes allows the model to be trained more thoroughly when searching for coherent topics. Additionally, when tuning these parameters, I found that increasing both the chunksize and passes parameters or either one made the model's performance worse overall.

How do you know that a topic model is good? I can judge my model's performance by its coherence score. The model's coherence score is a measure of the interpretability of topics and evaluates the degree of semantic similarity between words in the same topic by comparing word co-occurrence within the documents. Higher coherence scores generally indicate more coherent and interpretable topics. I used the "C_v" coherence measure since it is the default measurement of Gensim's LDA model and is the most interpretable.

As mentioned previously, there is no correct number of topics in a LDA model unless we generate multiple models with a different number of topics in each model. My approach to finding the optimal number of topics is to build many LDA models with different values of the number of topics (k) and pick the one that gives the highest coherence value. Choosing a 'k' that marks the end of a rapid growth of topic coherence usually offers meaningful and interpretable topics. Picking an even higher value can sometimes provide more granular sub-topics. To illustrate this approach, I created a graph that plots the growth of the coherence values on the y-axis and the increasing number of topics on the x-axis. In the graph, the line peaks at a certain coherence value and decreases afterwards indicating the best number of topics. With the optimal number of topics found, I created an interactive visualization of the LDA model that shows topic bubbles containing the top 20 keywords found in the topics thus we can start interpreting meaningful and relevant topics.

Now that I had my best LDA model, I went on to explore the topics more in-depth by looking at what topics are the

most dominant in each individual caption, what caption best captures its respective topic, and how the topics are distributed across the entire data set. By doing this exploration, I can see what topics impact certain captions and the probability of each caption showing up in each topic since I want to find how captions are distributed across all topics. I can get a better understanding of how the topics are structured in the data that allows me to find interesting groupings in which some captions may be connected to one another through high word frequencies.

After finishing the LDA model, I briefly explored an alternative technique of building a topic model called Top2Vec. Top2Vec is another unsupervised model that automatically detects topics present in text and generates jointly embedded topic, document and word vectors. The difference between LDA and Top2Vec was that I did not have to do any preprocessing steps because words that appear in multiple documents cannot be assigned to one single document. While testing Top2Vec, the model created some meaningful topics but resulted in too many topics that had to be hand-checked. Ultimately, there was not a clear visualization of the topics which made me pivot to the LDA model for topic modeling.

9.2 Caption embeddings

The second piece of my part in the overall project was to create sentence embeddings for the contests' captions. The embedding model that I am using is called Bert which is a type of language model that's really good at understanding the context of words in a sentence. It's like a language understanding powerhouse trained on a massive amount of text data. Sentence-Bert is a version of the Bert model that has been fine-tuned on not just individual words but whole sentences. It can take a sentence and turn it into a numerical representation, or "embedding," that captures the meaning of the entire sentence. The reason for using a pretrained model instead of a self-built model is that the former is excellent for tasks where understanding the meaning or similarity between sentences is crucial, such as finding cluster similarities between captions. Sentence embeddings can be thought of as a follow up to the topics model because while the model gives me information on the latent topics, it does not give me information on the semantic relationship between captions. This is because the model does not account for contextual meaning in the captions when finding captions, however the embedding model does as it takes the whole caption when it encodes them.

By using a pretrained model, I am able to look at semantic connections between individual captions by using the context clues from the entire caption to find clusters among the whole dataset. The SBert model that I chose is called "all-MiniLM-L12-v2" which is a general purpose model meant for sentence encoding and semantic search. I chose this model because while it sacrifices a bit of performance compared to the best model (68.70 vs 69.67), it is three times faster when computing embeddings (7500 sentences/sec vs 2800 sentences/sec). I was willing to trade off some performance for faster speed because the performance difference from the best model was marginal at best. Once I chose my model, I proceeded to fetch the data from our SQL database and did not do any preprocessing steps unlike the LDA model. The Sbert model learns to compute text representations in context. This means that the representations computed for a word in a specific sentence would be different from the representations for the same word in a different sentence. This context also comprises stopwords, which can very much change the meaning of a sentence. The same goes for punctuation: a question mark can certainly change the overall meaning of a sentence. Thus no preprocessing steps are needed when using the pretrained model.

Afterwards, I created the model embeddings using the SentenceTransformers library and saved them in a compressed numpy file with each caption having its own embedding. Once I had these embeddings, I started to work on creating an interactive visualization to show how they were semantically connected to one another. I decided to use k-means clustering to visualize the embeddings because I wanted to see how the embeddings clustered together. K-means is a way to group similar items together. In our context, these items are sentence embeddings, which are numerical representations of sentences. The "K" in K-means represents the number of clusters or groups you want to create. K-means looks at the similarity between sentence embeddings and tries to form clusters where sentences within the same cluster are more similar to each other than to sentences in other clusters.

I created a function that found the optimal amount of k-means clusters for each contest using silhouette scores. I then put the embeddings into a numpy array and fit them to a t-SNE plot. A t-SNE plot allows me to use a dimensionality reduction technique to visualize high-dimensional data in lower-dimensional spaces, in this case 2D space. It offers insights into the distribution and relationships between different sentences in the embedding space since it preserves the pairwise similarities between sentences. Additionally, the plot reveals semantic structures or relationships between different groups of sentences. Overall, t-SNE helps by mapping these embeddings into a space where similar sentences are close together, making patterns more visible For example, sentences with similar sentiments might form distinct groups. I made the plot interactive using plotly so when users view the plot, they can hover over each individual dot and see the caption, its mean, ranking, and contest number. This is to give users the ability to see how the captions are clustered in groups through semantic similarity between them. Ultimately, the embeddings plot is supposed to show a different aspect of the captions compared to the LDA

topics model in which the former showcases the semantic network between captions and the latter exploring hidden topics within the data.

I displayed the interactive caption embeddings visualization through Streamlit. Streamlit is a tool in Python that turns data scripts into interactive web apps without a steep learning curve. You can add interactive widgets like sliders, buttons, or text input to your app with minimal effort. This makes it easy for users to play with the parameters or input data and see the results instantly on the web page. I made my Streamlit app display the interactive plot and had a legend on the right hand side detailing what cluster the captions were in. Users can double click on the legend icons to showcase only a single contest at a time. Alternatively, I created a drop down menu on the left hand side that enables users to select one contest to view at a time. At first glance, all the caption embeddings from every contest will be displayed to give users an overview of how the plots will look like for every contest. In every contest, there might be different amounts of clusters shown as each contest is calculated differently, thus having dissimilar clusters.

In addition to using a pretrained model to create caption embeddings, I also used the Doc2Vec algorithm from Gensim to create embeddings. During a meeting, it was brought up that there could be alternative methods to create embeddings. I researched other ways for embeddings and found Gensim's Doc2Vec method. Doc2Vec is an algorithm designed to represent documents as vectors. It was introduced as an extension from Word2Vec which represents words as numerical vectors. So while Word2Vec is used to learn word embeddings, Doc2Vec is used to learn document embeddings or in this case; caption embeddings. Doc2Vec is an unsupervised learning technique that maps each document to a fixed-length vector in a high-dimensional space, similar to the SBert model. There are two variants of the Doc2Vec approach: Distributed Memory (DM) or Distributed Bag of Words (DBOW). I chose the former approach because this approach learns a vector representation for each piece of text data such as a sentence by taking into account the context in which it appears compared to the latter approach that focuses on understanding how words are distributed in a text, rather than their meaning. DM architecture considers both the word order and document context, making it more powerful for capturing the semantic meaning of captions.

After I chose which variant to use, I simply loaded in the data as before and created two versions of the Doc2Vec model, one with lemmatization of the captions and one without lemmatization. I thought it would be best to see how the vectors differed in meaning between the two versions of the model. For both versions, I loaded in the data and did preprocessing on the text except for doing lemmatization on the no lemmatization version. I then created a unique id for each word in a caption using the tagged_document function because the model expects document representation in the form of a word list and tags list. Afterwards, I built the two Doc2Vec models and trained the models on their respective corpora to generate embeddings. Finally, I saved the caption embeddings into their own numpy files for future usage. All of these vectors and embeddings serve to help me understand how the captions are related to one another either through latent topics and their probabilities of occurring in those topics or semantic network relationships visualized interactively.

9.3 Visualizations

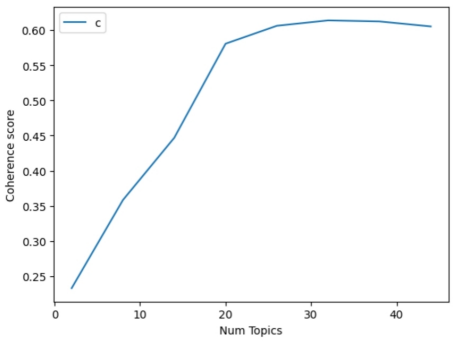

Figure 4: Graph of K for finding Optimal Number of Topics

9.4 Results

The results of my part in the project culminated in me understanding how topic models and sentence embeddings function and their roles in helping understand what makes a caption funny. Through the topic model I created, I found some meaningful topics in the dataset such as topic 23 which is about art and topic 30 which is about stick figures. The captions found in these topics reflect their probability showing up in each topic. For instance, the caption "I'm sorry, but our ad

	caption_text	ranking	0	1	2	3	4	5	6	7
0	First, let me fill you in.	0	0.343750	0.010417	0.010417	0.010417	0.010417	0.010417	0.343749	0.010417
1	Unfortunately, we're looking for someone of more substance	1	0.007813	0.007813	0.007813	0.207150	0.007813	0.007813	0.007813	0.007813
2	"Coworkers have been complaining about your constant lack of detail."	2	0.005209	0.005209	0.005209	0.005209	0.005209	0.171886	0.171887	0.005209
3	We now recognize that Hangman was an insensitive choice for the team building exercise.	3	0.003906	0.003906	0.003906	0.003906	0.003906	0.003906	0.003906	0.003906
4	Your lack of a mind, body, and soul is exactly what this corporation is looking for	4	0.003906	0.003906	0.003906	0.003906	0.003906	0.253909	0.003906	0.503903

Figure 5: Table of Topic Probabilities

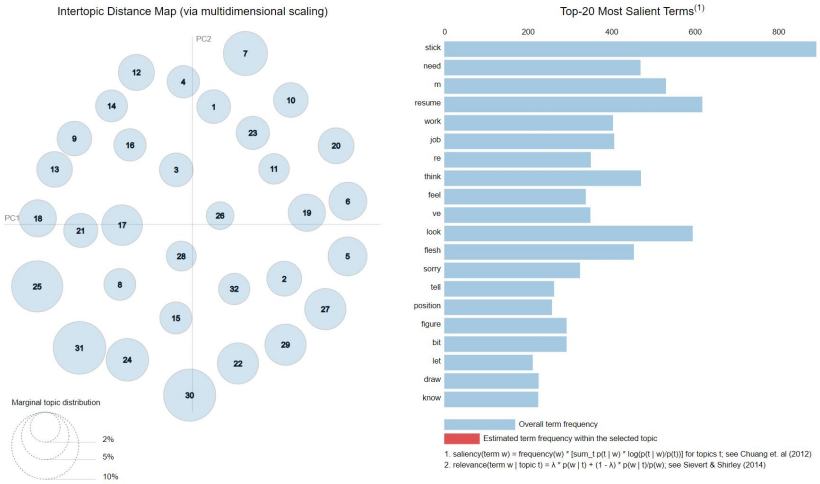

Figure 6: Interactive Plot of LDA Topic Model

says 'six-figure' salary, not 'stick-figure' salary" is primarily dominated by topic 30 (.50) and does not load well onto topic 23 (.003). On the other hand, the caption "Of course, you're Mr. Jones! I recognized you from your son's drawings" loads more on topic 23 (.25) compared to topic 30 (.003). These vectors show me the probability of captions being present in each topic, as some captions may load more dominantly on a single topic which may show insight into what the topic is about. Additionally, the interactive plot of the topic model allows users to explore what words make up the topics. In the plot, there is a bubble for each topic and users can hover each bubble to see what words comprise the topic and interpret the grouping of words to come up with a meaningful and interpretable topic.

As for caption embeddings, the results of the plot allowed me to explore how the captions were semantically related to one another regardless of what contest or cluster they were in. Some clusters that were identifiable included a cluster about Ozempic, a cluster about dieting, and a cluster about resumes. The clusters about Ozempic and dieting are related to one another and are close to each other in proximity on the plot as expected, while the cluster about resumes is seen on the other side of the plot. This example of clusters shows me how they are close together or separated based on their semantic similarity.

What I found in my vectors and plots indicate that there are meaningful topics found in the data even from such a small sample size which means that the model is working well to discover hidden topics. Around 8 topics were able to be interpreted in some sort of way out of 26 topics which is excellent considering I only had 7500 captions at my disposal for this part. The caption embeddings also showcased some potential in their semantic clustering as the interactive showed many clusters of interesting topics such as AI, bathroom, and pronouns. The t-SNE plot also showed the distance between clusters which is useful in understanding the dissimilarity between them.

9.5 Conclusion and Future work

Some limitations of my part in this overarching project include having a small sample size for my data which impacted the topic model's performance in finding good topics. Another limitation is the parameters I used for the model; I used a small amount of documents in each training set and iterated over them five times. A higher amount of documents in each training set and doubling the amount of iterations may increase the model's coherence score resulting in better topics.

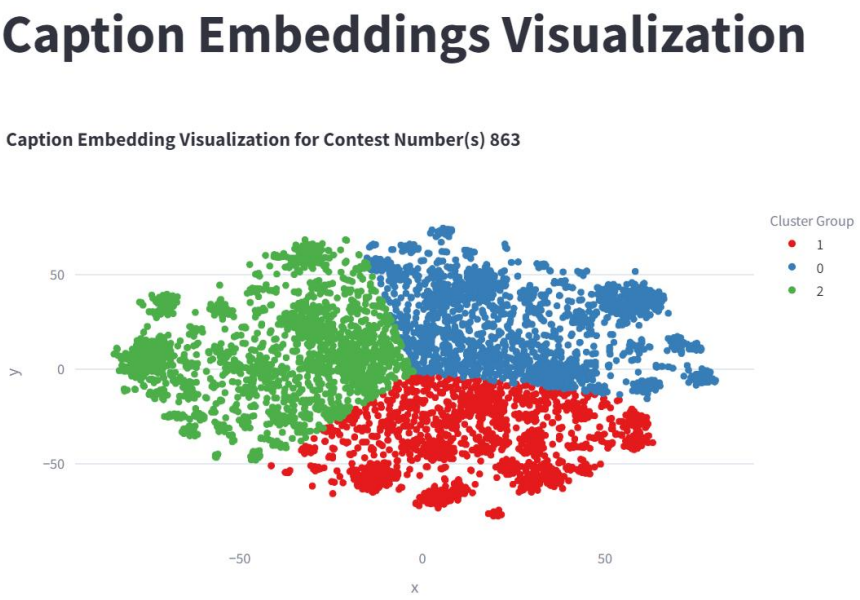

Figure 7: Interactive Plot of Caption Embeddings

For future steps, I would like to incorporate my topic vectors and caption embeddings into our SQL database so that other people can pull down this data without recreating the models again. I would like to see how using another pretrained model would affect the accuracy of creating good embeddings that captured semantic similarity. For instance, how would the performance of "all-mpnet-base-v2" compare to "all-MiniLM-L12-v2" which is the model that I used when creating caption embeddings. How different would the k-means clusters look like in the t-SNE plot? Would they be more close together or far apart? Future scientists can also test different topic model algorithms and compare them to the LDA model for topic model coherence scores. Additionally, these vectors and embeddings are different methods future scientists may be able to use to help predict future captions' funniness ratings. It would be interesting to see the difference between the usage of topic vectors versus caption embeddings when predicting a caption's funniness. In the future, the goal for my part should be to include my interactive visualizations into the artificial intelligence interface with Streamlit so that users can have the possibility of playing around with the interactive plots.

In conclusion, the two pieces of my part showcased different aspects of the caption dataset. The topic model was designed to find hidden and interesting topics in the dataset to potentially see if the topics can be used for future inferences on captions. The model was capable of producing quality topics which is a good sign for reproducibility in the future. As for the caption embeddings, they illustrated semantic structures of the captions and their relationship with semantic similarity through an interactive t-SNE plot shown by Streamlit. The pretrained model that I used worked well in producing good embeddings for the plot and showed that it is a good indicator of discovering connections between sentences.

10 Humor creativity assistant interface (Zhuolun + Lihao)

10.1 Function 1: AI assistant

10.1.1 Design

Some may argue if AI does not comprehend humor and creativity, how can it assist humans in tasks related to humor and creativity? Artificial tools do not need to possess the ability to autonomously solve problems; their mission is fulfilled by adapting to the external environment under human use, such as a paintbrush being used for painting. However, it is evident that directly asking AI for assistance is not a purposeful way of using the task of humor creativity. Because that would simply mean handing over the task to AI for autonomous completion. Therefore, what truly matters is the specific process through which creativity comes into play and the extent to which AI can assist in this process.

Creativity does not occur at the physical level; instead, it takes place at the conceptual level. Therefore, descriptions of its process are abstract and inferential. The conceptual process in which creativity occurs is often described as inspiration. In the process of literature review, Zhuolun found that research suggests that inspiration stems from imitating and observing

subjectively novel entities (Okada & Ishibashi, 2017). After delving into cognitive architecture and certain theories in computer science, Zhuolun conceptualizes inspiration as the result of parallel interactions of mental representations. Specifically, mental representations are mental imagery humans use to reconcile the internal and external environments (Simon, 1996). They take specific forms or content, portraying reality or abstraction, but are not bound by the rules of the physical world. For example, decimal representation can be a mental representation that humans use to solve mathematical problems in the external environment. However, humans can also adopt binary or hexadecimal representations in the internal environment. The physical world does not limit mathematics to the decimal system, allowing humans the freedom to flexibly and abstractly adjust their representations. Although the limitations of working memory can quickly force humans to focus on a single mental representation in a serial manner, humans can maintain multiple mental representations simultaneously in a parallel manner. Due to the adoption of different forms and content in various mental representations for different purposes, the interaction of these mental representations can trigger inspiration. For instance, reasoning logic and unicorns are entirely different mental representations. When both mental representations are given relatively balanced weight and considered simultaneously, results emerge without aligning with the predictions of either single representation.

Creativity requires inspiration, and inspiration relies on the parallel interaction of mental representations. Therefore, as a tool, AI can provide the mental representations necessary for such interactions. This is also the psychology perspective in utilizing AI: Large language models are perhaps best thought of, not as intelligences in any usual sense but as vast summaries of human behavior that take a particularly compelling form (Mitchell & Krakauer, 2023). In other words, AI can serve as an ideal repository of mental representations, offering materials for inspiration tailored to specific needs.

Our AI assistance function has established a fixed template to extract relevant information feedback from the large language model, providing users with mental representations to spark inspiration. This template primarily focuses on three design principles: strategy, randomness, and recording. Firstly, we referenced funny captions to compile dozens of strategies incorporated into the prompt for obtaining feedback from the AI. This means leveraging AI's flexible conversational abilities to convey successful humorous mental representations to the user's mind. Secondly, these strategies are randomly selected when users attempt to seek assistance from AI. This is aligned with the definition of creativity discussed in the conclusion section of the experiments. If creativity relies on randomness, it is preferable to emphasize randomness in the inspiration process. Finally, our program records captions generated by users, using labels to distinguish between those already conceived and newly created ones. Both types of records will be prompted to AI with distinction when seeking AI feedback. This approach aids AI in more quickly and accurately pinpointing relevant mental representations from the broad repository for specific scenarios.

10.1.2 Details

Figure 8 offers an overview of the AI assistant function. Each number in the figure corresponding to different parts of the interface. (1) This is a choice box for function options (AI assistant or inspiration); (2) This is the block designed for connecting to ChatGPT, where users are required to enter their OpenAI API key. Users have the options to choose between two AI models: GPT-3.5 or GPT-4; (3) This is the place to post The New Yorker Cartoon of the most recent caption contest; (4) This is the block for users' inputting. Users can input their own descriptions for the cartoon or select default descriptions. They are required to input thier captions that they want to seek suggestions. After inputting, users can click the "ask for assistant" button to gain the suggestion; (5) This is a "free scratch paper" tool that automatically records all captions requested by users for suggestions. Additionally, there's a "record to draft" button that allows users to manually insert the current caption from the caption box into the scratch paper. This functionality provides a convenient way for users to track and manage their caption ideas and revisions; (6) This expandable area allows users to review their entire suggestion history.

10.2 Function 2: inspiration

10.2.1 Design

The primary limitation of the AI assistance function lies in the requirement for users to provide content to receive feedback and gradually improve through continuous interaction. This implies that there may be significant challenges in the early stages of use. If a user cannot come up with any captions, the AI assistance cannot commence. Additionally, suppose a user does not find inspiration after receiving some feedback. In that case, the AI assistance may struggle to pinpoint suitable mental representations accurately, remaining in a stage of generating vague feedback. To address this issue, our interface introduces a second function to complement the shortcomings of AI assistance function.

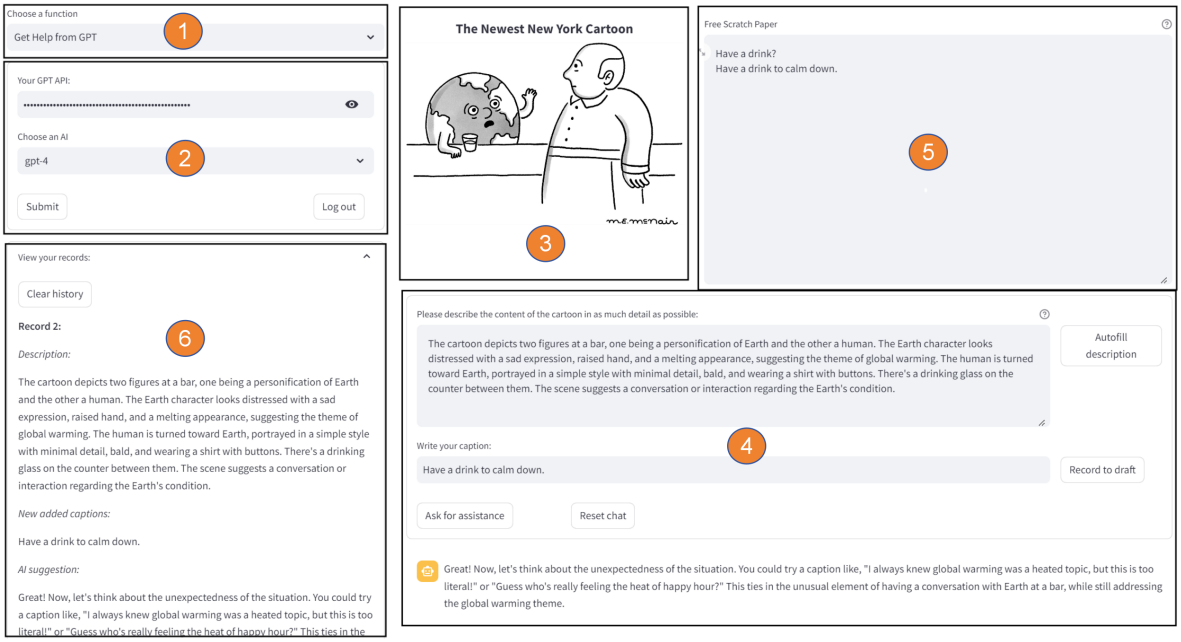

Figure 8: Interface for AI assistant

The second function follows the same theoretical framework as the first one, so it won't go into unnecessary details. The key difference is that the second function operates independently of AI, doesn't require caption input from users, and utilizes a different template to provide inspiration. In terms of strategy, the inspiration function is simple. It selects a cartoon contest and funny captions corresponding to it from the database. This strategy aligns more closely with the conclusions drawn from inspiration research (Okada & Ishibashi, 2017). Similarly, the inspiration function incorporates randomness. Users can randomly draw cartoon contests or choose them manually. However, the three corresponding captions are randomly selected from the top 20 in the funny ranking. Despite indicating in previous experiments and discussions that appreciating humor does not require a specific mindset, specific mental representations can help create humor. These mental representations are part of one's personality and are not obtainable through direct conscious thinking. In the inspiration function, appreciating captions from cartoon contests unrelated to the task can be more beneficial for activating these underlying mental representations.

10.2.2 Details

Figure 9 simply displays inspiration function. (1) In this choice box, users can either select a contest randomly or specify a particular contest number; (2) This place displays a cartoon along with three selected funny captions.

10.3 Data collection

In addition to providing valuable functions for users, Lihao has also created a robust login system and a user data collection mechanism for our interface. In the user authentication of this system, for signing up, users are not required to set a password. Instead, they use their email address as their username. This system allows our project partners to collect and utilize data for future research purposes. Whenever users interact with our interface to seek assistance or inspiration, the system automatically captures and inserts important data into the database. This includes the descriptions users provide for the current cartoon, the initial captions they generate, the suggestions they receive and so on. Figure 10 is an example of collecting one record from one user. Every time a user clicks the "ask for assistance" button, the system automatically records this action by inserting a record into the database. Each set of data is associated with the user's account username, ensuring that the information is securely attributed to the correct user. By combining this user-generated data with the data from each caption contest, we can gain insights into whether the top-rated captions originate from human creativity or AI assistance. This comparative analysis is helpful in evaluating and contrasting creative abilities between humans and AI. Ultimately, it enables us to better understand the contributions of both human users and the AI assistant in the captioning process, further enriching our research capabilities.

The detailed attributes of collected data is as below.

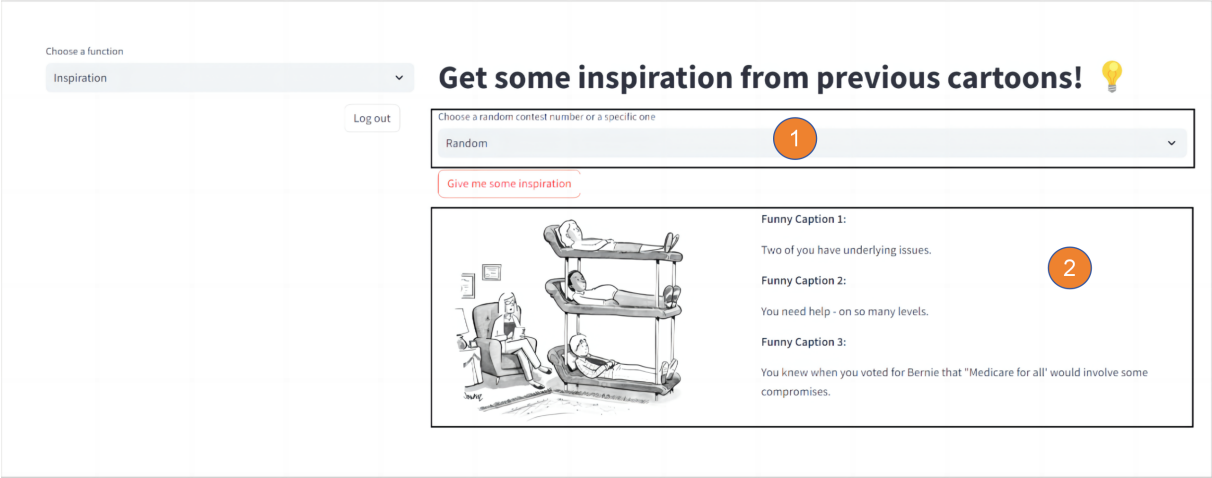

Figure 9: Interface for inspiration

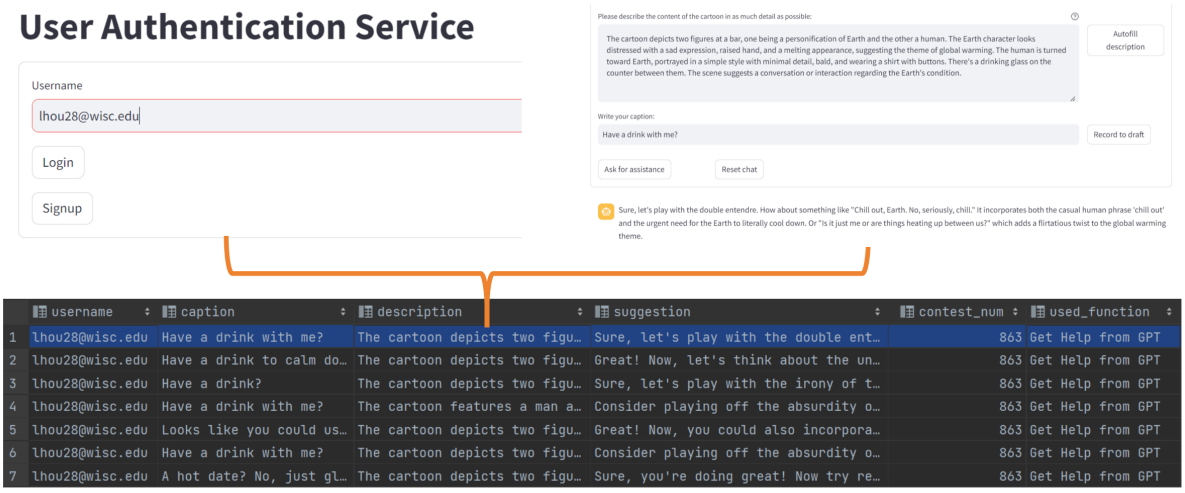

Figure 10: User data collection

- **Basic information**
  - username
  - used_function (Inspiration / Get Help from GPT): the function that the user used in this record
  - contest_num: the meaning of 'contest_num' varies based on the 'used_function'. If the used function is "Inspiration", the 'contest_num' corresponds to the contest number selected in the inspiration function. Conversely, if the used function is "Get Help from GPT", the 'contest_num' refers to the most recent The New Yorker Cartoon contest number.
  - time: the time of records creating
- **Information of the function of AI assistant**
  - description: the description a user inputs
  - description_type (0 = inputting by a user own description; 1 = inputting default description 1 generated by ChatGPT-4; 2 = inputting default description 2 generated by LLaVA)
  - caption: the caption a user inputs
  - method: the strategy used to ask ChatGPT-4 for a suggestion
  - model (gpt-3.5-turbo / gpt-4)
  - suggestion: the AI-generated suggestion

- **Information of the function of inspiration**
  - random_select (0 = selecting a specific contest number; 1 = randomly selecting)

10.4 Future work

The interface has not undergone rigorous human participant experiments yet, so future testing is desired. Ideally, creating an independent cartoon contest and dividing participants into three groups—AI, humans using the interface, and humans not using the interface—could be a suitable approach. After collecting captions, classic voting methods involving other human participants could be employed to assess humor levels. However, it is important to note that individuals exhibit significant differences in humor creativity from a personality perspective, and judgments of humor levels are highly subjective. This experiment may require a substantial number of human participants to yield meaningful results.

What is more worth investigating is the developmental changes that occur in humans after long-term use of the interface. Since humor and creativity exhibit individual differences in humans, and these abilities are not static, conducting smaller-scale longitudinal tests to assess skill development over time may be better. In conclusion, there is much valuable work to be done with human participants in the future.

References

Apte, M. L. (1987). Ethnic humor versus "sense of humor" an american sociocultural dilemma. *American Behavioral Scientist*, *30*(3), 27–41.

Apter, M. J. (1982). " fawlty towers": A reversal theory analysis of a popular television comedy series. *Journal of Popular Culture*, *16*(3), 128.

Bail, C. A. (2023). Can generative ai improve social science?

Baumeister, R. F., Maranges, H. M., & Sjåstad, H. (2018). Consciousness of the future as a matrix of maybe: Pragmatic prospection and the simulation of alternative possibilities. *Psychology of Consciousness: Theory, Research, and Practice*, *5*(3), 223.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Caird, S., & Martin, R. A. (2014). Relationship-focused humor styles and relationship satisfaction in dating couples: A repeated-measures design. *Humor*, *27*(2), 227–247.

Cole, S., & Kvavilashvili, L. (2021). Spontaneous and deliberate future thinking: a dual process account. *Psychological research*, *85*, 464–479.

Cropley, A. J. (1999). Creativity and cognition: Producing effective novelty. *Roeper review*, *21*(4), 253–260.

Derks, P., Staley, R. E., & Haselton, M. G. (1998). " sense" of humor: Perception, intelligence, or expertise?

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Gaut, B. (2010). The philosophy of creativity. *Philosophy Compass*, *5*(12), 1034–1046.

Glăveanu, V. P. (2014). The psychology of creativity: A critical reading. *Creativity. Theories–Research–Applications*, *1*(1), 10–32.

Hausman, C. R. (1979). Philosophy of creativity. *Ultimate Reality and Meaning*, *2*(2), 143–162.

Hessel, J., Marasović, A., Hwang, J. D., Lee, L., Da, J., Zellers, R., ... Choi, Y. (2023). *Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest.*

Jain, L., Jamieson, K., Mankoff, R., Nowak, R., & Sievert, S. (2020). *The new yorker cartoon caption contest dataset.* `https://nextml.github.io/caption-contest-data/`.

Jones, C., & Bergen, B. (2023). Does gpt-4 pass the turing test? *arXiv preprint arXiv:2310.20216*.

Kvavilashvili, L., & Rummel, J. (2020). On the nature of everyday prospection: A review and theoretical integration of research on mind-wandering, future thinking, and prospective memory. *Review of General Psychology*, *24*(3), 210–237.

Levine, J., & Redlich, F. C. (1955). Failure to understand humor. *The Psychoanalytic Quarterly*, *24*(4), 560–572.

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). *Improved baselines with visual instruction tuning.*

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). *Visual instruction tuning.*

Lubart, T. I. (2001). Models of the creative process: Past, present and future. *Creativity research journal*, *13*(3-4), 295–308.

Martin, R. A., & Ford, T. (2018). *The psychology of humor: An integrative approach.* Academic press.

McGraw, A. P., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological science*, *21*(8), 1141–1149.

McGraw, A. P., Warren, C., & Kan, C. (2015). Humorous complaining. *Journal of Consumer Research*, *41*(5), 1153–1171.
Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.
Mumford, M. D. (2003). Where have we been, where are we going? taking stock in creativity research. *Creativity research journal*, *15*(2-3), 107–120.
Okada, T., & Ishibashi, K. (2017). Imitation, inspiration, and creation: Cognitive process of creative drawing by copying others' artworks. *Cognitive science*, *41*(7), 1804–1837.
Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, *23*(12), 1026–1040.
Plucker, J. A., Esping, A., Kaufman, J. C., & Avitia, M. J. (2015). Creativity and intelligence. In S. Goldstein, D. Princiotta, & J. A. Naglieri (Eds.), *Handbook of intelligence: Evolutionary theory, historical perspective, and current concepts* (pp. 283–291). New York, NY: Springer New York. doi: 10.1007/978-1-4939-1562-$0_1$9
Poulsen, V. M., & DeDeo, S. (2023). *Large language models in the labyrinth: Possibility spaces and moral constraints* (Vol. 1) (No. 4). SAGE Publications Sage UK: London, England.
Putnam, H. (1973). Meaning and reference. *The journal of philosophy*, *70*(19), 699–711.
Putnam, H. (1975). The meaning of" meaning".
Redshaw, J., & Suddendorf, T. (2016). Children's and apes' preparatory responses to two mutually exclusive possibilities. *Current Biology*, *26*(13), 1758–1762.
Ritchie, G. (2001). Current directions in computational humour. *Artificial Intelligence Review*, *16*(2), 119–135. doi: 10.1023/A:1011610210506
Ritchie, G. (2004).
The linguistic analysis of jokes. doi: 10.4324/9780203406953
Ritchie, G., Manurung, R., Pain, H., Waller, A., Black, R., & O'Mara, D. (2007, December 1). A practical application of computational humour. In (pp. 91–98). (4th International Joint Workshop on Computational Creativity, IJWCC 2007 ; Conference date: 17-06-2007 Through 19-06-2007)
Runco, M. A. (2023). *Creativity: Research, development, and practice*. Academic Press.
Simon, H. A. (1996). *The sciences of the artificial*. MIT press.
Simonton, D. K. (2000). Creativity: Cognitive, personal, developmental, and social aspects. *American psychologist*, *55*(1), 151.
Sjåstad, H., & Baumeister, R. F. (2023). Fast optimism, slow realism? causal evidence for a two-step model of future thinking. *Cognition*, *236*, 105447.
Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.
Weisberg, R. W. (1993). Creativity: Beyond the myth of genius.
Wyer, R. S., & Collins, J. E. (1992). A theory of humor elicitation. *Psychological review*, *99*(4), 663.
Yamao, Y., Matsumoto, R., Kunieda, T., Shibata, S., Shimotake, A., Kikuchi, T., ... others (2015). Neural correlates of mirth and laughter: a direct electrical cortical stimulation study. *Cortex*, *66*, 134–140.