Cracking Cross-Cultural Comedy: Teaching BART to LOL!

Frank Lee and **Arthur Liang** and **Katherine Liu** and **Ella Wang**

Abstract

As human-AI interactions become increasingly commonplace, the ability for NLP models to authentically emulate human-like interactions becomes increasingly crucial. An important part of this capability is based on the model's capacity to decode and generate humor. In this work, we task state-of-the-art (SOTA) NLP models with two tasks: (1) matching a joke setup to its punchline, and (2) explaining why the joke is funny. These tasks, as presented in existing literature, assess an "understanding" of humor, whether it be identifying ironic scenarios or comprehending clever wordplay. We find that humans still significantly outperform language models in both tasks (BART for matching, GPTs for explanation). In addition to analyzing performance, we aim to understand if current models have the resolution to understand deep-rooted cultural difference in humor. Thus, we perform our analysis with two datasets of English and Chinese jokes and conduct quantitative and qualitative analyses between the two. Awareness of cultural nuances can enable future AI systems to engage in culturally sensitive conversations, improve cross-cultural communication, and foster cultural appreciation. This work illuminates the utility of such NLP models as human collaborators in tasks that not only require advanced semantic understanding, but also a sensible awareness of the cultural context.

1 Introduction

Humor is a complex and multifaceted aspect of human communication; the ability to decode and generate humor holds the key to developing NLP models that can truly mimic human-like interactions. This is a fundamental problem in NLP – the need for systems that not only comprehend and produce language but can do so with nuance and wit that resonates with human sensibilities (Anjum and Lieberum, 2023). With this understanding, we can bridge the gap between machine and human communication, enabling AI systems to engage in conversations that are not only informative but also entertaining and culturally sensitive. For example, if discussing a popular cultural phenomenon, an AI system should be able to make relevant and tasteful comments that resonate with its audience. Moreover, the significance of exploring cross-cultural humor lies in its capacity to promote intercultural understanding and appreciation. Humor in one culture might rely heavily on wordplay, while in another culture, it could be more centered around storytelling or satire (Raskin, 2012). Different societies have distinct values, traditions, and forms of humor, and by deciphering the intricacies of these comedic expressions, NLP can facilitate a more profound comprehension of various cultures.

Specifically, we want to assess the ability of SOTA NLP models to understand jokes across cultures and languages by employing a two-fold approach using English and Chinese jokes. First, models will be tasked with matching joke setups to punchlines given a set of possible alternative punchlines, offering a quantitative assessment of their humor understanding in both linguistic domains. Furthermore, we will challenge the models to generate an explanation for why a given joke is funny. This qualitative evaluation enables us to further probe models' ability to understand the subtleties of humor. By performing these tasks on both English and Chinese jokes, we identify commonalities and differences in how humor is understood by language models across cultural contexts. This awareness of cultural nuances can enable future AI systems to engage in culturally sensitive conversations and improve cross-cultural communication. Real-world applications of this research include enhancing virtual assistants and language translation services to better adapt to the preferences of users from different cultural backgrounds (Hershcovich et al., 2022).

2 Related Work

Past studies have found that humor can be created when there is a violation of a script or schema (Raskin, 2012), which can be accomplished through linguistic devices such as puns, metaphors, and hyperbole. Recent research in NLP has explored unconventional and context-specific word usages, encompassing similes (He et al., 2022), idioms (Dankers et al., 2022), and puns (Anjum and Lieberum, 2023), where wordplay often arises from the subversal of conventional language rules. However, the ability of LLMs to understand humor has received limited attention in NLP research. We want to contribute to this underexplored topic by assessing the degree to which language models can comprehend humor across cultures.

A few methods of ascertaining an LLM's higher-level understanding of humor (via a cartoon caption contest) are laid out in Hessel et al. (2023). By tasking models to match captions to cartoons, identify winning captions, and explain the humor in each caption, the group concludes that while today's AI underperforms humans at these tasks, it still exhibits a significant capability in tasks that involve humor (Hessel et al., 2023). The tasks chosen in the paper "encapsulate progressively more sophisticated aspects of 'understanding' a cartoon; key elements are the complex, often surprising relationships between images and captions and the frequent inclusion of indirect and playful allusions to human experience and culture" (Hessel et al., 2023). We apply similar strategies to measure how well humor is understood by SOTA models by having them match punchlines to jokes and explain why the punchlines are humorous. Additionally, we expand our scope of analysis across cultures and languages by performing our benchmarking on both English and Chinese joke datasets, and identify discrepancies in performance.

Overall, this assessment of humor has an important place in the advancement of cross-cultural NLP. Hershcovich et al. (2022) provides several ways to improve diversity and cultural sensitivity in NLP models, highlighting that a current challenge is cultural bias. They specifically note that cultural bias can be reduced by balancing the preservation of cultural values while minimizing harmful cultural aspects. As culture-specific elements are a important component of any culture's humor, our efforts helps answer the larger question of whether LLMs are capable of understanding these cultural values, a first step in enabling culturally sensitive and inclusive communication.

3 Data

To realize our motivation of understanding wordplay and humor across different cultures, we identified two datasets of jokes in English and Chinese.

Our English dataset (Weller and Seppi, 2020) contains one million jokes from the r/jokes subreddit, annotated with the corresponding post scores [1], which are used to preprocess the dataset for the highest-quality examples. The majority of these jokes take the form of a question and answer, e.g. "Which animal has the softest bite? Gummy bears." The question-answer format of these data is easily sectioned into a joke setup (the "question") and its punchline (the "answer"), allowing us to task NLP models to match punchlines to setups.

The second dataset contains 3365 jokes in Chinese with their associated humor level [2] (Tseng et al., 2020). The humor levels are provided as part of the dataset on a scale of 1 to 5, with 5 being the funniest. We preprocessed the entries in our Chinese dataset for the top-scoring examples and converted suitable jokes into the same setup-punchline format. Although some examples have a multiple-line setup, we identified that the last line typically contains the punchline.

To further standardize our datasets, we sampled the top-scoring or highest humor level jokes from both datasets. For the HuggingFace dataset, we utilized the 7224 jokes with a score of over 1000, and for the Chinese dataset, we used 3002 jokes with a humor level of at least 2 out of 5. Although there may be reliability issues with the humor level scores due to lack of objective measures, we decided not to include the lowest scoring jokes since the processed data still has a large number of jokes that have been subjectively validated.

4 Experiments

To assess an LLM's understanding of humor, we pose two tasks.

1. *Matching*: Can a model recognize the appropriate punch line for a given joke? We provide four possible choices of punch lines, only one of which is the correct punch line. For example, for the joke "A guy named Bart walks

[1] number of upvotes minus the number of downvotes
[2] human-annotated

into a bar. He immediately gets shot and dies. Who killed him?", we provide the following possibilities:
(a) Caught her red handed
(b) Now he's in a pickle
(c) don't worry he's 0k
(d) **the bartender**
where the other three punchlines are randomly selected from *other* jokes in our dataset. This task was evaluated using metrics including accuracy, precision, recall, and F1-score.

2. *Explanation*: Given a joke, can a model explain why it is funny? For this task, team members wrote a ground truth explanation for 26 jokes, and we quantitatively evaluated the explanations given by our model using a cosine similarity to the ground truth explanation. Furthermore, explanations were qualitatively compared across cultures by team members to determine the nuances of what is considered funny in different cultural contexts. For example, humor may be derived from techniques such as clever wordplay, cultural references, satire and irony, or storytelling and situational humor–we assessed the comedic styles of English and Chinese jokes to determine the cultural contexts and preferences that shape how people find them funny.

5 Methods

Our overall approach to the matching and explanation tasks is depicted in Figure 1.

5.1 Matching Task

Baseline Model
For the matching task, we fine-tune Word2Vec to obtain our baseline results. After fine-tuning the Word2Vec model on our joke dataset, the "sentence embedding" for each of our setups and punchlines is calculated as an average of the word embeddings in each sentence. To match joke setups to punchlines, cosine similarity is calculated between the Word2Vec embeddings of the joke setup and potential punchlines, and the punchline with the highest cosine similarity to the setup is returned as the matched punchline.

The Word2Vec baseline serves as a benchmark for more advanced models, helping to gauge the complexity of the humor comprehension task. Word2Vec treats each word in isolation and does not consider the sequential nature of language limiting its ability to understand language. In the context of jokes, where context and phrasing play crucial roles, Word2Vec may lack the mechanisms to humor of a joke. We expect Word2Vec to perform decently with examples where setups and punchlines are semantically similar but otherwise fail to identify the the correct punchline, resulting in an accuracy slightly better than random guessing. This level of "understanding" might be comparable to a child's, which has close to no notion of wordplay. Moreover, Word2Vec uses a fixed-size context window to learn word embeddings. This limitation may lead to missing long-range dependencies, especially in jokes that rely on a story-telling structure for a setup. Due to these limitations in capturing certain linguistic nuances and context, using the Word2Vec model as a baseline model allows us to assess the understanding of more sophisticated models beyond the semantics of each joke.

Fine-Tuned BART
To benchmark an LLM's performance at this matching task, we fine-tune a pre-trained BART model, a denoising autoencoder capable of text generation. The model combines the bidirectional and language-agnostic encoder BERT with an autoregressive decoder, and the transformer architectures have a greater expression than our baseline model. With the capability to store a words' multiple meanings and keep track of longer dependencies, we posit that BART is better at detecting the humor in our punchlines.

For the English matching task, we used Facebook's BART base model with 149M parameters, and for the Chinese matching task, we used Fudan University's BART base model with 140M parameters. We fine-tuned the BART models on randomly selected jokes, using a 60%, 20%, 20% train/val/test split on both datasets. The models are given "context: joke setup. options 1: punchline 1, 2: punchline 2, 3: punchline 3, 4: punchline 4" as input and are fine-tuned to produce the correct choice out of 1, 2, 3, 4 as output.

Human Matching
To measure human performance on the matching task, 100 randomly selected jokes from our dataset were given to authors who had not previously seen them. The authors are similarly tasked to identify

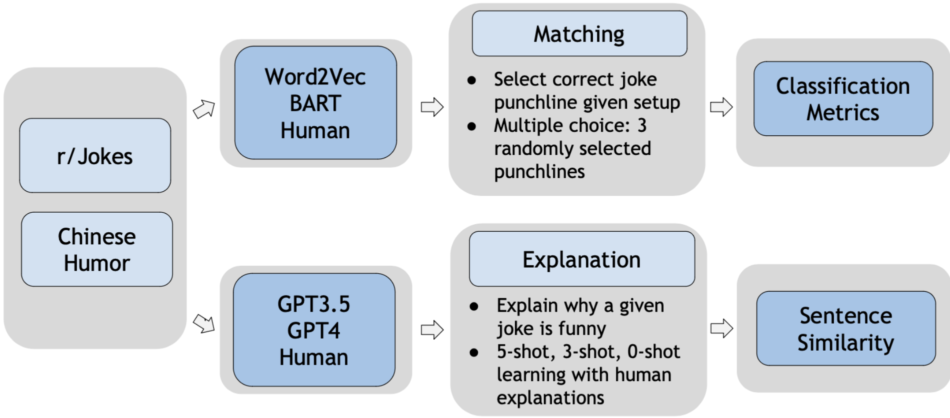

Figure 1: To analyze LLMs' ability to understand humor, we employ a two-task pipeline with a matching task and explanation task.

the punchline of these jokes from four possible options, including the correct punchline and three other randomly selected punchlines.

5.2 Explanation Task

GPT Models
To assess the ability of large language models to understand humor, we prompted GPT models to explain why jokes were funny. GPT-3.5-turbo and GPT-4 models were used as 5-shot, 3-shot, and 0-shot models. We provided the models with a joke setup and punchline, and for the 3-shot and 5-shot cases, we also provided examples of human explanations written by the authors. For Chinese joke explanations, the models were prompted to explain why the Chinese jokes were funny using English, implying that language translation was necessary. However, to ensure that the results were indicative of humor understanding only (as opposed to translation efficiency), the models were explicitly instructed not to describe any explanation of the jokes' translation or meaning.

Human Annotations
Human-written explanations served the ground-truth in this experiment. Explanations for 26 English jokes and 26 Chinese jokes, all randomly selected, were written by the authors. The Chinese jokes were read in Chinese with no translation, but the Chinese joke explanations were written in English, with no description of the translation process included.

6 Implementation

In terms of computational resources, we leveraged the GPU capabilities provided by Google Colab Pro to train our models. We obtained pre-trained models for the matching task through HuggingFace and tools from their transformers library for fine-tuning and evaluation. Additionally, we leverage OpenAI's GPT APIs for the explanation task. Finally, model performance was quantitatively evaluated using tools from Python package scikit-learn.

7 Results

7.1 Matching Task

Baseline Model
During evaluation, the model calculates the cosine similarity between the setup and each punchline, selecting the punchline with the highest similarity as the predicted punchline.

With accuracies of 29.8% and 38.1% on the English and Chinese jokes respectively, the Word2Vec model performs slightly better than random guessing at the matching task (Table 1).

Fine-Tuned BART
We fine-tuned the respective BART models on 4334 training examples from our English r/jokes dataset and 1224 training examples from our Chinese dataset on the matching task, achieving accuracies of 81.6% and 87.8% on the English and Chinese jokes respectively (Table 1).

Some of the examples that the BART model matches incorrectly demonstrate that the model is unable to fully grasp humor, particularly in semantically unrelated sentences. For example, the model is unable to understand the wordplay

Model	Accuracy	Precision	Recall	F1
Baseline Word2Vec				
English	0.2980	0.1788	0.2965	0.2079
Chinese	0.3807	0.2515	0.3807	0.2850
Fine-Tuned BART				
English	0.8159	0.8276	0.8159	0.8159
Chinese	0.8783	0.8796	0.8783	0.8781
Human				
English	0.9900	0.9850	0.9900	0.9867
Chinese	0.9800	0.9850	0.9850	0.9834

Table 1: Model results on matching task. When fine-tuned, BART performs far better than baseline measures for matching, but still significantly underperforms humans. Additionally, both Word2Vec and BART perform better at matching Chinese punchlines than matching English punchlines.

between "Beyonce" and "buoyancy" in the following example:

Setup: My best mate told me he was totally into Beyonce.

- I don't have the heart to tell her it's meant to be the bottom one. (**predicted by model**)

- I said "whatever floats your boat". He said "No, that's buoyancy". (**correct answer**)

Additionally, the Chinese Word2Vec and Chinese BART models perform better on this dataset compared to their English counterparts, which may be because the Chinese jokes in the dataset tended to be longer than the English Reddit jokes, giving more context to the language models and informing more accurate predictions.

Human Matching

The human matching task was evaluated by calculating how many punchline guesses were correct out of the 100 provided joke setups. Humans significantly outperform LLMs, with accuracies of 99% for English jokes and 98% for Chinese jokes. While the BART model has improved performance from the Word2Vec model, there is still a gap from how humans understand humor.

7.2 Explanation Task

The joke explanations generated by GPT-3.5-turbo and GPT-4 were evaluated using cosine similarity as a metric to quantify the semantic resemblance between the GPT-generated explanations and human-written explanations, which served as the ground

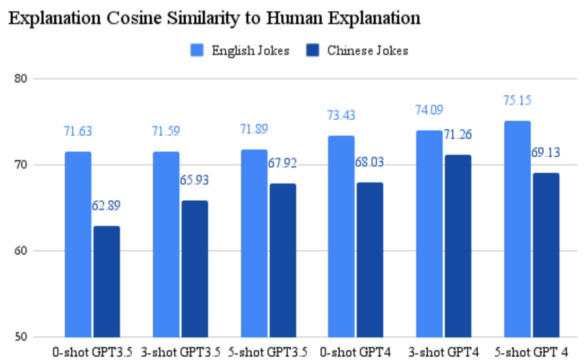

Figure 2: Cosine similarity between human-written and GPT-generated explanations for English and Chinese jokes. GPT-4 generally performs better than GPT-3.5-turbo, and the cosine similarity between English explanations is higher than between Chinese joke explanations.

truth. Explanation strings were encoded into fixed-size vectors using the SentenceTransformer library, and the average cosine similarity was computed. Figure 2 reports the cosine similarity scores for both models using 5-shot, 3-shot, and 0-shot learning scenarios. GPT-4 consistently outperformed GPT-3.5-turbo; since GPT-4 can access updated information on current events, it may better understand popular cultural phenomena and references in jokes. Moreover, because GPT-4 is a much larger model than GPT-3.5, it may be more capable of understanding humorous content and discerning nuances in jokes, resulting in more coherent responses.

In contrast to the matching task results, the cosine similarities between English joke explanations were significantly higher than Chinese joke explanations. We hypothesize that these observations are likely due to two factors. Firstly, Chinese jokes tend to exhibit greater complexity and length than their English counterparts, often involving multi-line setups. This complexity may result in more diverse explanations of why the joke is humorous, introducing greater deviation from the ground-truth results. Additionally, despite the jokes being composed in Chinese, the models were instructed to explain their humor in English, necessitating a translation step that likely contributed to discrepancies in the joke explanations.

0-shot learning generally performed slightly worse than 3-shot and 5-shot learning, displaying an understanding of the content of a joke but not necessarily its humor. For example, below we have included a joke and the outputs from 0-shot and 5-shot GPT-3.5-turbo:

1. **Joke**: 'My dad is a social distancing champion. I haven't seen him since 2005.'

2. **GPT-3.5-turbo 0-shot explanation**: The joke is funny because it is a play on words. Social distancing is something that people have been doing since the outbreak of the COVID-19 pandemic, but the term can also be used to describe someone who is not very social. In this case, the joke is saying that the person's dad is so not social that the person hasn't seen him in 15 years.

3. **GPT-3.5-turbo 5-shot explanation**: This joke is funny because it is a play on words. The dad is a champion of social distancing, which means he is good at avoiding people, but in this case it also means that he has not seen his son in a long time.

For this joke, the 0-shot model is able to explain topics in the joke and recognize a play on words as well as the context of the COVID-19 pandemic (which is arguably irrelevant in the joke), However, it fails to accurately articulate the source of humor, incorrectly attributing it to the person's dad being 'not social.' In contrast, the 5-shot model correctly explains that the joke is funny by highlighting that the joke implies the dad 'has not seen his son in a long time'.

To quantify the differences in humor between English and Chinese jokes, sentence embeddings derived from joke explanations for these two languages were analyzed. Embeddings for the explanation strings were generated used the Sentence-Tranformer library, and the t-SNE algorithm was

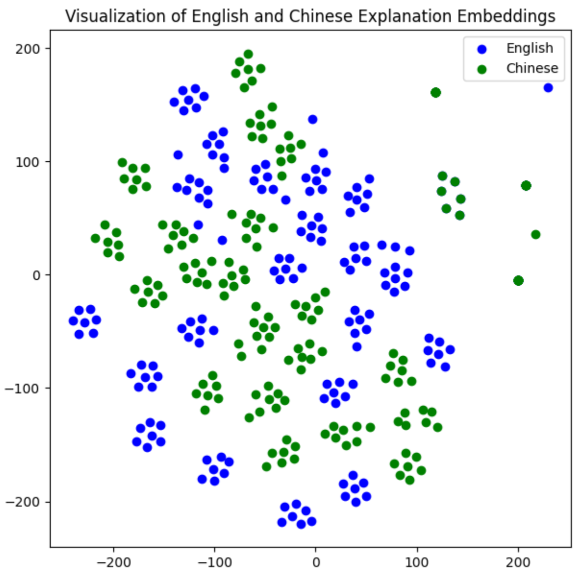

Figure 3: t-SNE visualization of English and Chinese joke explanation embeddings. 7 embeddings per joke are plotted, each generated from different GPT models and a human-written explanation. The embeddings appear clustered, suggesting differences in English joke and Chinese joke humor.

applied for dimensionality reduction, creating 2D representations of the embeddings. Figure 3 displays the spatial distribution of explanation embeddings for English and Chinese jokes. Each cluster of 7 points represents the embedding for the human-written explanation and explanations generated by the 5-shot, 3-shot, and 0-shot GPT-3.5-turbo and GPT-4 models, totaling 7 embeddings per joke. The embeddings do not appear to be randomly distributed; rather, the green points appear clustered in one area, and the blue points appear clustered around them. These results suggest that the large language models have learned some difference between the humor in English and Chinese jokes, thus causing the explanation embeddings to exhibit cluster patterns in their spatial distribution.

We further sought to identify these differences by conducting a qualitative review of the English and Chinese joke embeddings. The authors reviewed the explanations generated for 100 English jokes and 100 Chinese jokes to determine trends in their use of humor.

The comparison between English and Chinese jokes in Table 2 highlights cultural distinctions in humor expression. English jokes, often characterized by vulgarity and political themes, reflect a culture of open discourse and satire, while Chi-

English Jokes	Chinese Jokes
Reddit content is often vulgar or relevant to politics/religion	Derive humor from offensive/satirical comments, but not “controversial” topics
Many references to popular cultural phenomenon, events, and celebrities	Jokes tend to set up made-up situations, more isolated from current events
High use of puns and wordplay	More situational humor, rather than employing figures of speech

Table 2: Qualitative analysis of English and Chinese jokes suggest differences in their content and humor techniques.

nese jokes, avoiding controversy and favoring situational humor, suggest a cultural inclination towards subtlety and avoidance of sensitive topics. These differences underscore the influence of overarching cultural values on the manifestation of humor in linguistic contexts. We note that these differences may be influenced by the choice of datasets in our experiments; specifically, the nature of Reddit, known for its user-generated content and diverse communities, might introduce a bias towards content that is more open, edgy, and politically charged. Thus, the observed themes in English jokes may be more reflective of Reddit's user demographics and community norms. However, due to Reddit's large user base with over 430 million active users and its frequent usage as training data for large language models, this dataset may still serve as a significant and relatively comprehensive portrayal of English humor across diverse contexts.

8 Discussion

We demonstrate that SOTA NLP models currently fall short of identifying and explaining humor as effectively as humans. However, the significant improvement observed between these models and baseline Word2Vec models signifies the utility of these models as collaborators, perhaps most directly as brainstorming, benchmarking, and feedback aids for comedians. The non-trivial understanding of our jokes demonstrated in matching and explanation also supports the injection of humor into human-AI interactions. Models that can adjust to different cultural values in their understanding and generation of humor may find use in any setting that requires advanced semantic understanding and application of cultural context.

Our findings also suggest that Chinese culture exhibits a more conservative humor style than American culture, emphasizing the need for models that can adapt to diverse cultural contexts to ensure cultural sensitivity and appreciation in user-facing products. For example, if telling a joke in Chinese, we may intend for models to demonstrate an awareness of the cultural nuances that contribute to the conservative nature of humor in Chinese. Given the subjectivity of humor, the development of models that can account for individual variances and avoid generating content that may be perceived as offensive is imperative.

The results of this study also call for a more in-depth analysis of language agnostic LLMs, models with a vocabulary that spans multiple languages. A matching experiment with such models might reveal further biases in the detection or generation of humor.

To address the limitations of our evaluation schema, we aim to develop more representative metrics, independent of proxies such as sentence embeddings. Having an independent metric corroborate those results will help ensure that our observations are not simply due to noise introduced by our proxy choices. Specifically, we would like to incorporate human validation in our results for the explanation task by presenting GPT-generated explanations and human-written explanations for a joke to humans, and then prompting them to choose their preferred explanation in a blind experiment. The results of this experiment may further elucidate the effectiveness of large language models in understanding and articulating humor.

In addition, future work involves collecting larger datasets of human-written explanations to validate our results and mitigate potential bias. Incorporating sentiment and text analysis methods will inform fine-tuning strategies and help quantify qualitative observations, helping identify areas where current large language models perform poorly.

Additionally, from just this work alone, we cannot claim that the human-machine "humor understanding gap" is closed. Another experiment that might yield interesting insights is, within the matching task, prompting our models with not only the correct punchline and incorrect punchlines from other jokes but also human-crafted, syntactically similar, but nonsensical punchlines. If the model performs similarly to how it has performed in our paradigm, we can be more confident that the model is making more than just an embedding association based on vocabulary. These sentences with low-distance embedding similarity can be generated by masking or switching out words from the correct punchline and rearranging the sentence structure to "eliminate" the humor of the punchline.

Lastly, NLP models alone still do not have any sense of timing and delivery of not only jokes but also speech. In verbal or stand-up humor, timing and delivery are crucial. Because we are simply prompting our models using different corpuses without this information, they cannot mimic the timing and vocal inflections that often make verbal jokes funny. Furthermore, much of humor in human interaction relies on non-verbal cues such as facial expressions, body language, and tone of voice. NLP models, which primarily operate on text, cannot interpret or utilize these non-verbal cues. These aspects of humor are still absent in text-based models, limiting the ability for large language models to fully understand the nuances and subtleties involved in humor.

9 Impact Statement

The two datasets our work is based on represent a narrow slice of humor, originating from a particular language and culture. Therefore, the results of our study do not represent all types of humor and our models may be biased towards just the ones represented. For example, our models fine-tuned to jokes from r/jokes, especially when filtered by ratings, may include a partiality or bias towards a humor style that conforms to the characteristics of the average Reddit user (just under half of Reddit's users are American). We do not claim that our datasets represent the ground truth for humor and these models capable of understanding just a few aspects of humor might also inadvertently perpetuate stereotypes or cultural insensitivities if not carefully designed and monitored.

Another failure mode of our project is that there is potential for joke generation to be used for malicious purposes like cyberbullying. In the wrong hands, humor-proficient models can be prompted to generate witty yet offensive comments. With how accessible social media is today, the bar to operationalize these comments is low. Unfortunately, this may cause even further harm to disproportionately marginalized communities who already face such behavior. For example, in a world run by polarizing news, generating humorous or satirical propaganda that may have sway in the political and thus policy-making scene needs to be controlled. Humor can be a powerful tool for political commentary and persuasion. Humor-proficient models could be exploited to craft persuasive political messages, potentially influencing public opinion and elections. Thus, as large language models become increasingly advanced, it is critical to establish clear ethical guidelines for the development and use of humor-generating models, as well as integrate algorithmic safeguards to detect and filter potentially harmful content when these models are deployed in the real world.

Finally, the explanations in our annotated corpus were written by authors who themselves have individual cultural backgrounds and thus biases. It would be sensible in the future to analyze person-to-person variance in explaining why particular jokes are funny and obtain a larger diversity of perspectives.

References

A. Anjum and N. Lieberum. 2023. Exploring humor in natural language processing: a comprehensive review of joker tasks at clef symposium 2023.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human?

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

V. Raskin. 2012. Semantic mechanisms of humor. Vol. 24.

Yuen-Hsien Tseng, Wun-Syuan Wu, Chia-Yueh Chang, Hsueh-Chih Chen, and Wei-Lun Hsu. 2020. Development and validation of a corpus for machine humor comprehension. *"12th Language Resources and Evaluation Conference (LREC 2020)"*.

Orion Weller and Kevin Seppi. 2020. The r/jokes dataset: a large scale humor collection. *"Proceedings of the 2020 Conference of Language Resources and Evaluation"*.

Acknowledgements

We would like to thank Professors Jacob Andreas, Yoon Kim, and Chris Tanner as well as Michael Maune and Marco Nocito for their support and guidance throughout this semester and project.