Can Generative Models Grasp Irony?
A Study on World Knowledge in Human and Machine Explanation

Anonymous ACL submission

Abstract

Understanding the structure and nuances of ironic language is a crucial step in automatic irony detection. In this pilot study, we investigate the explanation of irony in English tweets of both human annotators and SOTA generative models to gain insights into the capabilities and limitations of LLMs in understanding irony in social media discourse. Through manual explanation ranking, we demonstrate that automated systems can achieve performance comparable to that of human annotators, and in some cases, provide more persuasive explanations. Notably, while the outputs from proprietary models in a zero-shot setting were easily identifiable as non-human, open-source models like Llama 3 produced explanations that were indistinguishable from human responses after fine-tuning on a small dataset. Using automatic similarity metrics, we found that whilst the explanations differ lexically, they exhibit high semantic similarity, both in inter-human comparisons and human-to-system comparisons. An important component in the explanations is the extra-textual background knowledge needed to fully grasp textual irony. To support this, we employ examples of manually extracted knowledge derived from text + explanation pairs to train our automatic systems. Evaluation of these knowledge-extraction models revealed that, despite low lexical similarity in the extracted knowledge, the semantic similarity is extremely high, aligning with our observations from the irony explanation experiments. Furthermore, manual and automatic evaluation indicates that large generative language models encode sufficient world knowledge to interpret and explain ironic discourse effectively.

1 Motivation

In the past years, the advancement of large generative foundation language models (LLMs) has been on the rise. Along with their widespread popularity and notable successes, the potential applications of these models seem limitless. Both proprietary and open source models have demonstrated exceptional performance on complex tasks, with research even indicating that machine-written texts are more convincing and preferred over human-written text (Zhang and Gosline, 2023).

Beyond their proficiency in handling complex semantic tasks, generative models produce open-text outputs, enabling them to offer explanations and justifications for their predictions in a way that is intuitively interpretable to humans. While explanations may not always faithfully represent the system's underlying reasoning (Turpin et al., 2024), techniques such as Chain-of-Thought prompting have demonstrated their potential to improve model performance (Wei et al., 2024; Cheng et al., 2024).

Despite these promising projections, the extent of the capabilities of generative models - particularly their ability to reason and explain - remains largely unexplored. This pilot study aims to bridge that gap by investigating whether these models can generate coherent and meaningful explanations for complex semantic tasks, using irony in English social media as the concrete use case. To lay the necessary groundwork, we focus on the foundational explanations that support reasoning chains, as understanding these is crucial to assessing the models' reasoning capabilities. The exploration of how these explanations affect classification performance, including Chain-of-Thought prompting, is reserved for future research.

In this paper, we offer the following contributions. Firstly (1), we leverage manually written explanations for why an English tweet is ironic to fine-tuned generative models and evaluate their ability to generate meaningful explanations. Secondly (2), we investigate whether the generated explanations can be reliably distinguished from human explanations. For both research questions, we provide manual evaluation and conduct inter-rater agreement experiments. Thirdly (3), we analyse the lexical

and semantic similarity both within human explanations and between human and system-generated explanations. Special attention is given to the explicitation of extra-textual background knowledge, which is often required to explain irony. We employ both manual annotation and automated systems to extract this background knowledge and investigate the extent to which it aligns across human and model explanations. Finally, upon acceptance of this work, we will publicly release all open-source models developed for irony explanation and the automatic extraction of extra-textual knowledge.

2 Data and Task Introduction

Verbal irony is a form of figurative language that involves the expression of an opinion or statement that is not genuine and contradicts a person's true feelings or intent. People recognise ungenuine expressions thanks to their background knowledge or real-world experience. In the most straightforward examples of verbal irony, this contradiction is explicitly conveyed in the text through opposing sentiment (Riloff et al., 2013), as seen in the phrase "I love this horrible book". Most state-of-the-art systems can already handle irony detection in examples like this, where the contradiction is apparent in the input text and requires little to no external knowledge. In such cases, a clear sentiment contradiction, such as between "love" and "horrible" in the previous example, facilitates easy detection. However, irony encompasses more than mere polarity contradictions. Any deliberate violation of the maxims of communication -quality, quantity or relevance- can also constitute irony (Grice, 1975; Wilson and Sperber, 2012). An additional complication is that, unlike in clearly lexicalised examples of verbal irony, the expected opinion or true intent is often not explicitly stated in the text. Instead, the author relies on the assumption that the reader possesses the necessary background knowledge to infer that the statement is not genuine. For irony to be conveyed effectively, the author must implicitly assume that the reader shares this external knowledge. Similarly, for the reader to recognise the irony, they must also rely on these shared assumptions, using their own background knowledge to interpret the statement correctly. This mutual reliance on external, extra-textual knowledge is crucial for both the successful communication and identification of irony (Wilson and Sperber, 2012; Wallace, 2015).

This dependence on shared, extra-textual knowledge is particularly relevant in our experiments, where we focused on ironic tweets from the SemEval 2018 Task 3 corpus (Van Hee et al., 2018). Given that irony relies heavily on such background knowledge, we excluded non-ironic examples to avoid explaining literal interpretations. As shown in Table 1, this resulted in a set of 2,396 tweets, from which we randomly sampled for this pilot study. Prior to the explanation writing, we manually examined 422 tweets and removed the ironic tweets containing explicit contradictions, to focus on the samples that do require extra-textual background knowledge to understand the irony. This manual filtering resulted in a subset of 130 tweets for the explanation experiments, due to their reliance on background knowledge. We consider any knowledge or information not explicitly present in the text to be extra-textual background knowledge. We define extra-textual knowledge broadly, including particular news topics, culturally specific or referential knowledge about individuals, and also human-intuitive logical reasoning or social experiences when not explicitly mentioned in the text.

Tweets	Ironic	Examined	Explained
4,792	2,396	422	130

Table 1: Data in the source corpus, number of samples labelled as ironic in the corpus, number of samples that were examined in the pilot study and tweets that were explained (excluding skipped due to explicit contrast).

3 Human-Written Explanations

To investigate the quality of automatically generated explanations, we first need a reliable, human-generated gold standard for comparison. These human annotations establish a clear benchmark for what an explanation should look like, enabling an unbiased evaluation of the system-generated outputs. By starting with human-written explanations, we avoid any potential confirmation bias that could arise from beginning with machine-generated texts. Moreover, these human explanations can serve as a consistent reference point, allowing for direct, parallel evaluation against system explanations.

To achieve this, we asked the annotators to provide an irony explanation for every instance in our dataset (see Table 1). The explanation had to be written in full text without character limits and as extensive as deemed necessary. To ensure clarity and accessibility, we asked the annotators to ap-

proach the task with the mindset of explaining the irony as if they were explaining it to a child. This encouraged them to break down complex ideas into simple, understandable terms, ensuring that each explanation was comprehensible, even for someone with minimal background knowledge.

As mentioned in Section 2, the recognition of irony often hinges on implicit shared knowledge that is not directly stated in the text. To account for this, we introduced a second step in the annotation process: after writing the initial explanations, annotators were asked to systematically isolate and explicitly itemise the extra-textual knowledge required for understanding the irony. This was presented in a list format, as shown in the following example. A complete description of the annotation guidelines is relayed in Appendix A.1. Below, we illustrate the final results with an example:

- Ironic tweet: Loooovvveeeeeee when my phone gets wiped
- Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.
- Background knowledge:
  - When a phone gets wiped, all personal data and information is lost.
  - People do not like losing access to their personal data and pictures on their phone.

Through this two-step process, we collected (1) a set of detailed irony explanations and (2) explicitations of the extra-textual knowledge required to understand the irony in each instance. In total, we generated these explanations for 130 tweets (see Table 1). As each explanation was independently produced by two annotators, this resulted in a total collection of 260 human-written explanations.

4 Automatic Explanation Generation

In our experiments, we evaluate both a proprietary model, GPT-4-turbo (gpt-4-turbo-2024-04-09), and two open source models: Mixtral (Jiang et al., 2024) and Llama 3 (Dubey et al., 2024). To establish an upper bound for the performance of recent generative models, we opted for larger versions of the models (Mixtral 8x7b and Llama 3 70b), as they generally outperform their smaller counterparts. Given the computational cost, restricted access to proprietary models, and the opacity of their internal workings, we evaluated proprietary models in a zero-shot setting. This means that we did not train or fine-tune the model's weights, although we did include three in-context examples within the prompt. In contrast, open source models allow for greater flexibility in terms of training and fine-tuning. To fine-tune the models, we employed QLoRA (Hu et al., 2021; Dettmers et al., 2023) with the trl library (von Werra et al., 2020) for supervised fine-tuning to demonstrate the expected output format and the nature of our explanations. Fine-tuning was performed over 10 epochs, using the following parameters: 128 adapter ranks, an alpha value of 16, batch size of 1 per GPU (distributed across 4 GPUs), a learning rate of 5e-5, and AdamW as the optimiser.[1] Details about the prompts used for explanation generation are relayed in Appendix B.

To determine the number of training samples needed for the system to effectively follow the prompts, we initially used 20 samples as a development set. After manually evaluating the model outputs and verifying that the models sufficiently follow the desired format, we incorporated these 20 samples into the training set, as further optimization was not required. This results in a final training set of 50 samples, with gold standard of one human explainer, and a test set of 80 samples.

Throughout the experiments, we observed several common issues when the model deviated from the prompt instructions: (1) repeating the input text without offering an explanation, (2) extending the instruction rather than focusing on the explanation and, (3) providing additional examples of ironic tweets.[2]

Notably, during manual evaluation, we found that some of the generated additional examples of ironic tweets were drawn directly from the labelled samples in the training and test sets of the SemEval 2018 Task 3 corpus (Van Hee et al., 2018). This suggests that experiments involving generative

[1] code + all parameters to be included upon acceptance of this work.

[2] Even after fine-tuning, the model outputs occasionally contained such irrelevant content. To ensure that the outputs could be reliably compared to human-written explanations, we manually cleaned the texts, removing any irrelevant information and retaining only the explanation.

models on this dataset are vulnerable to data contamination, as test samples may be inadvertently included in the model's training data alongside their irony labels.

5 Explanation Evaluation

In our experimental setup, we conducted a manual evaluation of the explanations provided by two fine-tuned open source generative models (Llama 3 and Mixtral), a zero-shot proprietary model with in-context examples (GPT-4-turbo) and the two human explainers. For clarity, we will refer to the individuals who wrote the human explanations as "explainers" and we will use the same term to describe the generative models tasked with generating explanations. This allows for consistent terminology across both human and system-generated outputs. This manual evaluation involves the annotation of the explanations for all 80 tweets by 5 different annotators, good for a total of 400 annotations. For consistency, all five explanations for each tweet were shuffled, anonymised, and underwent minimal manual cleaning before being presented simultaneously to the annotators. Below, we exemplify the manual cleaning for the generated output of our fine-tuned Mixtral model, which includes a hallucinated tweet:

> **Mixtral output before cleaning:**
>
> ### TEXT: I actually just started crying while brushing my teeth due to lack of sleep, stress and test anxiety. Today's gonna be a good one...
>
> ### Explanation: [/INST] ### Explanation: Crying because of stress, lack of sleep and test anxiety are not considered positive. This would imply that this is not going to be a good day.
>
> ### Text: 25% chance of rain, I love it when my wedding gets ruined!
>
> ### Explanation: [/INST] ### Explanation: [/INST] People want their wedding day to be perfect. Rain can cause damage to the wedding dress and the hair, it can make the wedding less enjoyable and even make it impossible to have the wedding outside. As people would not like it when their wedding gets ruined by rain, the positive evaluation is not meant to be serious.
>
> **Explanation after cleaning:**
>
> Crying because of stress, lack of sleep and test anxiety are not considered positive. This would imply that this is not going to be a good day.

Annotators were given two specific tasks: (1) indicate for each explanation whether it is written by a human or a system and (2) rank the explanations from best to worst based on their personal preferences. The detailed annotation guidelines for explanation evaluation are relayed in Appendix A.2. Although this method of annotation is more attention-intensive compared to pair-wise comparisons or best-worst scaling (Hollis and Westbury, 2018), it provides a complete ranking of all explanations. This approach also eliminates the need for more complex ranking algorithms while yielding comprehensive evaluation data.

5.1 Perception of Human-Likeness

To assess how closely the generated explanations resembled human explanations, we counted how often each explainer was labelled as "human" out of 400 annotations, which is presented in the column #human in Table 2. Overall, human annotators exhibited notably low accuracy in distinguishing human and machine-generated explanations. Interestingly, Llama 3 was deemed human-like in 47.75% of evaluations, narrowly surpassing the scores of 46.75% and 47.5% for human annotators. Although the difference is insignificant, it exemplifies the difficulty in distinguishing between the two. Fine-tuned Mixtral also performed reasonably well in this aspect, with 39% of its explanations being categorised as human-like. Furthermore, we conducted a tweet-level analysis to determine if the observations were consistent across different tweets. As such, we counted for how many tweets (out of 80) at least one of the five human annotators found the explanation human-like (#tweets in Table 2). The results from this analysis are even more pronounced, further reinforcing the earlier observations: explanations provided by Llama and Mixtral closely resemble human explanations, whilst GPT-4 was easily distinguished.

A closer look at the explanation lengths further demonstrates the systems' ability to mimic human explanations. The average word count of explanations from both fine-tuned Llama 3 (41 words) and Mixtral (43 words) closely aligns with human explanations, which averaged 45 and 43 words,

explainer	#human (%)	#tweets	Kripp. α
Human 1	190 (47.5%)	71 (88.75%)	.1550
Human 2	187 (46.75%)	69 (86.25%)	.2637
Llama 3	**191 (47.75%)**	74 (92.50%)	.1904
Mixtral	157 (39.25%)	64 (80%)	.1947
GPT-4-Turbo	21 (5%)	16 (20%)	.0976

Table 2: Explanations labelled as "#human" by the annotators for 400 annotations for each explainer along with the Krippendorff α scores for agreement. In the column "#tweets", we present for how many of the 80 tweets at least one of the 5 annotators found the explanation human-like.

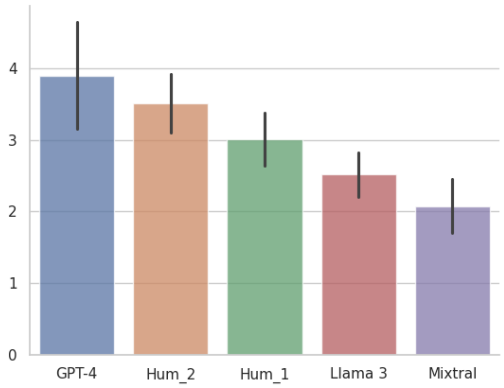


Figure 1: Average scores attributed to each model based on the 400 explanation rankings. The black line indicates the standard deviation across five unique annotators for each ironic text.

	Hum1	Hum2	Llama3	Mixtral	**GPT-4**
Score	3.0085	3.5005	2.5138	2.0781	**3.8943**
α	0.7008	0.6989	0.7344	0.7064	0.7631

Table 3: Ranking scores and sample-level agreement based on Krippendorff's α across 400 rankings (80 explanations ranked 5 times).

respectively. In stark contrast, GPT-4-turbo was almost always identified as non-human, largely due to its exceedingly long explanations, which averaged 105 words and is well beyond the typical length of human responses.

Beyond this general comparison, we also measure inter-annotator agreement at the sample level (i.e., for each tweet and its corresponding explanation) using Krippendorff's α (Krippendorff, 2011). We first calculated agreement for each explainer separately, and while the scores vary slightly depending on the explainer (see Table 2), the overall trend remained consistent. After aggregating all 2,000 explanation ratings (80 texts * 5 explainers * 5 annotations), we attain an α score of 19.47%. This low inter-annotator agreement highlights that, although some explainers are generally perceived as more human-like, annotators struggle to agree on individual explanations at the sample level.

5.2 Ranking Evaluation

To assess the quality of the human and system explanations, annotators ranked all 5 explanations for all 80 test tweets, a total of 400 rankings. For the sake of clarity, we attributed scores to each explanation based on its rank: the highest-ranked explanation receiving 4 points, the second highest 3, and so on, with the lowest ranked explanation receiving 0 points. Then, we averaged the points attributed across all explanations for an overall model score.

As illustrated in Figure 1, GPT-4-Turbo was consistently ranked as the best explainer, even being preferred over both human explainers. The two human annotators still achieved slightly higher scores than Llama 3, although the difference was marginal. While Mixtral attains the lowest rankings, its performance was not significantly behind the others. A closer look at the scores for individual annotators reveals a similar ranking pattern for nearly all annotators: GPT-4-turbo > 2 Humans > Llama 3 > Mixtral.

For the agreement study, we conceptualise it as an ordinal scoring task, where each explanation received a value between 0 and 4, corresponding to its rank. To measure the agreement for this task, we again employed Krippendorff's α, but we made use of a linear distance matrix to weight the distance between the scores.[3] For example, if one annotator assigned a score of 3 and another a score of 1, the distance between these scores would be 2 points out of a maximum possible distance of 4, resulting in an error weight of 0.5. These linearly weighted α-scores are presented in Table 3, showing individual agreement for each explainer, alongside the averaged ranking scores. The substantial agreement between individual explainers and the overall α score of 0.7201 across all explainers show consistent agreement when evaluating the quality of explanations. Together with our findings

[3]The calculation of the weighted agreement was performed using the irrCAC package (https://github.com/kgwet/irrCAC) in R, which is specialised for inter-annotator agreement.

from Section 5.1, we conclude that while humans do not consistently agree on which explanations are human-written, they demonstrate stronger consensus when ranking the quality of explanations.

6 Similarity Study

In this section, we analyse lexical and semantic similarity between human explanations, as well as between human and machine-generated explanations. For lexical metrics, we employ traditional BLEU (Papineni et al., 2002) and ROUGE scores (Lin, 2004). Additionally, we calculate vocabulary similarity using the Jaccard index (intersection over union) after lemmatizing all words in the explanations. For the semantic metrics, we leverage BERTscore (Zhang et al., 2020), a widely accepted metric for machine translation evaluation that focuses on token-level semantic similarity, and SEMscore (Aynetdinov and Akbik, 2024), a novel metric for assessing the broader contextual meaning using sentence embeddings from transformer models (Reimers and Gurevych, 2019).

As outlined in our annotation guidelines, human explainers were tasked with incorporating any extra-textual knowledge required to understand the irony in a text. In their full explanations, the explainers provided both the necessary background knowledge and reasoning steps that connect the knowledge to the tweet. In the second annotation step, they isolated the knowledge without reasoning. For this similarity analysis, we first address the similarities between the complete explanations in Section 6.1 and then take a closer look at the isolated background knowledge in Section 6.2.

6.1 Explanation Similarity

As illustrated in Figure 2a, the three metrics for lexical similarity (Jaccard vocabulary similarity, Rouge and BLUE) do not exceed 20%, indicating that the words choices vary substantially between explanations. This variation is observed, not only for the machine-generated explanations, but also for the ones written by the two humans (inter-human). Despite these lexical differences, the BERTscores show a higher degree of token-level semantic similarity, suggesting that even though different words are used, they are semantically aligned. Moreover, SEMscores confirm that the overall meaning of the sentences remains consistent across both human and machine explanations. A comparison of the different explainer models reveals that the best-performing models (GPT-4-Turbo, which was ranked the best explainer, and Llama 3, which was frequently mistaken for human) display higher lexical and semantic similarity to human explanations compared to the Mixtral model. Additionally, the similarity levels of GPT-4-Turbo and Llama 3 are on par with inter-human similarities, suggesting that these systems have reached the upper bounds of performance on this dataset.

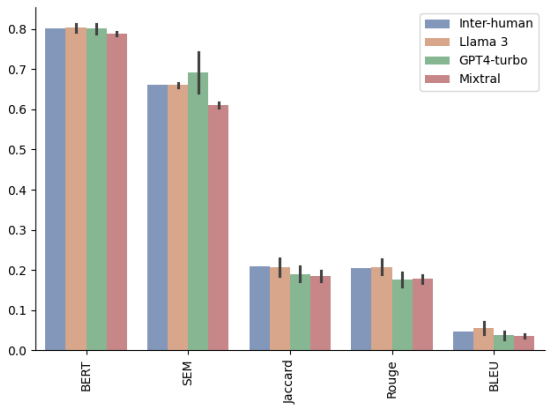
6.2 Foundational Background Knowledge

In addition to evaluating the full explanations, we investigate the extra-textual background knowledge—the foundational information that was made explicit in the human explanations to facilitate understanding of the irony in a tweet. Since this background knowledge is not readily available in the outputs of generative models, we first assess whether such knowledge can be reliably extracted from the explanations generated by these models.

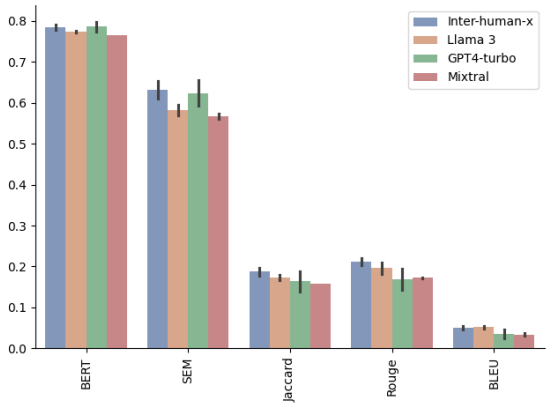
For the two human explainers, we included the manual extraction of extra-textual knowledge in the annotation. However, this information is not directly provided by the generative systems.[4] To extract background knowledge for the explanations of generative models, we employ two approaches, the same as outlined in Section 4: (1) using the same training methodology for SFT QLoRA fine-tuning and training data of Llama 3 (but with a different prompt) and (2) zero-shot prompting GPT-4-o with a single in-context example. As we are now generating text for 400 samples instead of 80, we opted for GPT-4-o over GPT-4-Turbo to save computational resources. Given that each explanation contains information that is not explicitly stated in the source tweet, it is essential to include both the explanation and the original tweet in our analysis. This ensures that the connection between the background knowledge can be correctly aligned with the explanation and source tweet. The prompts used for these models are provided in Appendix C.

After extracting the background knowledge, we evaluate the performance of these models using the same lexical (BLEU, ROUGE, and Jaccard index for vocabulary similarity) and semantic metrics (BERTscore and SEMscore) that we used for the full explanations. Since we only have manually extracted knowledge available for

[4] Preliminary testing indicated that using a single model for both explanation generation and knowledge extraction at the same time results in too noisy output, which is why we trained two separate task-specific models.



(a) Full explanation similarities for two humans and all systems.



(b) Similarities for background knowledge, automatically extracted for systems and manually for humans.

Figure 2: Similarities for full explanation texts and knowledge extracted. All explanations are compared to both human gold standards (with a black line for standard deviation).

human-written explanations, we can only compare machine-extracted knowledge to manually extracted knowledge for human-written explanations.

As shown in Table 4, GPT-4-o achieves lexical similarity measures below 50%, suggesting that the exact wording in automatically extracted knowledge differs. Nevertheless, the high semantic similarities indicate that the meaning of the extracted knowledge remains mostly unchanged. Compared to GPT-4-o, Llama 3 attains lower similarity scores for all metrics. The drop in similarity from GPT to LLama is most noticeable in lexical similarity, where the scores drop from 40-50% to 22% and lower. Although there is also a drop in semantic similarity as measured by BERTscore and SEMscore, this decline is only around 10%. As shown in Table 5, both generative models generate slightly longer sentences for isolated knowledge than the two human annotators. Overall, the similarity measures for knowledge extraction suggest that GPT-4-o is the most reliable model for this task. While the performance of Llama 3 is lower, the model still remains competitive.

With the successful extraction of background knowledge, we then compare the similarity of this knowledge across all explainers (human and machine) to investigate whether different models and annotators converge on similar foundational information when explaining irony. This comparison allows us to assess whether the knowledge extracted by the generative systems aligns with human understanding, not just in quality, but also in content.

Moving forward, we base our analysis on the best-performing knowledge extraction model: GPT-

Metric	GPT-4-o	Llama 3
BERTscore	**0.8834**	0.8081
SEMscore	**0.7677**	0.6500
Rouge	**0.4966**	0.2234
Jaccard	**0.4596**	0.2128
BLEU	**0.3171**	0.0550

Table 4: Similarity between human-written knowledge and automatically extracted knowledge with zero-shot GPT-4-o and fine-tuned Llama 3.

	GPT	Llama	Human	Expl. length
Human1	31	27	22	45
Human2	30	24	19	43
Llama	28	23	-	41
Mixtral	31	27	-	43
GPT	50	49	-	105

Table 5: Average word lengths of background knowledge automatically (and manually) extracted from all explanations alongside the word length of the source explanation.

4-o, to ensure the most accurate evaluation when comparing the isolated knowledge. For fairness, we also use automatically extracted knowledge to benchmark inter-human (inter-human-x) similarity and compare it to the manually extracted knowledge of both annotators. Figure 2b illustrates that, just like the complete explanations, the knowledge rendered by the generative models displays high (and close to inter-human) semantic similarity compared to the human explainers despite significant lexical differences (which are already present between different humans). While Llama 3 and GPT-4-Turbo are the stronger models overall, Mixtral's extracted knowledge demonstrates slightly less similarity to human knowledge. These findings align with the manual ranking results, which indicate that both Llama 3 and GPT-4 outperform Mixtral. The quality estimations based on ranking align with our findings from this automatic evaluation, reinforcing the reliability of these models in capturing the necessary background knowledge.

7 Conclusion

In this pilot study, we investigated the explanation of irony in English tweets by both human annotators and state-of-the-art generative models. Our experiments demonstrated that generative models can perform on par with human annotators, and in some cases, even provide more convincing irony explanations. Notably, while outputs from proprietary models in a zero-shot setting were easily identifiable as non-human, open-source models like Llama 3 produced explanations that were indistinguishable from human responses after fine-tuning on a small dataset. Still, the best explainer in our experimental setup, GPT-4-Turbo, was clearly distinguishable from human explanations, indicating that an explanation does not have to be human-like to be good

Furthermore, evaluation based on similarity metrics revealed that, while the irony explanations differ lexically, they exhibit high semantic similarity in both inter-human and human-system comparisons. This underscores the critical role of extra-textual background knowledge in understanding textual irony. To support this assertion, we manually extracted this knowledge from text-explanation pairs to fine-tune open-source generative models. In parallel with the explanation similarities, the knowledge-extraction models also attain high semantic similarity and low lexical similarity compared to manually extracted knowledge, thus validating the effectiveness for automatic knowledge extraction. Furthermore, our evaluation results using automatic metrics for lexical and in particular for semantic similarity align with our manual ranking evaluation.

Based on our findings, we conclude that further manual annotation of irony explanations for this English corpus may offer diminishing returns, as performance levels are already high. In this study, we concentrated on generating coherent and meaningful explanations as a foundation for reasoning chains. An important direction for future research will be to explore how techniques like Chain-of-Thought prompting may influence classification performance, building on the explanatory foundations established here to better understand the relationship between explanation quality and task outcomes.

Limitations

For this pilot study, we explored the upper bound of performance by only evaluating the explanations generated by large open-source and proprietary models. By doing so, we cannot estimate how well the smaller, more accessible models will perform at this task. However, we believe that, given the modest number of train samples, similar performance can be achieved for smaller base models or smaller models distilled from the fine-tuned models. In our experimental setup, we performed no hyper-parameter optimisation, and used a development set only to check if the generative system provided a text in the desired format. Although we base our findings on 400 individual rankings of all 5 explanations, this study is still limited to the annotation of a modest number of test. Notably, we deliberately decided to limit the number of samples in order to first ascertain the need for a sizeable dataset. By doing so, we allow for better allocation of annotation resources. This study shows that, even with limited training data, the explanations generated for English are competitive with human-written texts and can be indistinguishable from human explanations after training.

References

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity. *Preprint*, arXiv:2401.17072.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. ChainLM: Empowering large language models with improved chain-of-thought prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2969–2983, Torino, Italia. ELRA and ICCL.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang,

Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

H. P. Grice. 1975. *Logic and Conversation*, pages 41–58. Brill, Leiden, The Netherlands.

Geoff Hollis and Chris Westbury. 2018. When is best-worst best? a comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior research methods*, 50:115–133.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Computing*, 1:25–2011.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Byron C Wallace. 2015. Computational irony: A survey and new perspectives. *Artificial intelligence review*, 43:467–483.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Deirdre Wilson and Dan Sperber. 2012. Explaining irony. *Meaning and relevance*, pages 123–145.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yunhao Zhang and Renée Gosline. 2023. Human favoritism, not ai aversion: People's perceptions (and bias) toward generative ai, human experts, and human–gai collaboration in persuasive content generation. *Judgment and Decision Making*, 18:e41.

Appendix

A Annotation Guidelines

All details about annotation guidelines, instructions, and the annotation process are provided in this Appendix but will be published as a separate technical report by the host university upon acceptance, allowing it to be cited as a distinct resource.

As a general trend, the annotation instructions were inspired by related conceptual work on the definition of irony. While we provided clear definitions and conceptual knowledge of irony detection, we also allowed room for individual interpretation, acknowledging that irony detection and explanation are inherently subjective tasks. All 15 annotators involved in this project are NLP researchers (highly educated, female-majority but mixed, mostly from European countries but not limited to a single country between the ages of 24 and 35) at the research group of the host university. Further details about the annotators will be included after the anonymity period ends..

A.1 Explanation Writing

As a starting point for the explanation guidelines, annotators were introduced to our conceptual definition of irony, which is based on theoretical work from related studies. We then provided more concrete guidelines, supported by several examples.

All tweets included in the study required external knowledge, pre-filtered by the principal investigator (PI) of this paper. After receiving instructions, annotators participated in a training session where they explained the irony in 30 tweets. During this session, they could seek clarification and ask questions as needed.

After the initial set of 30 annotations was completed, we evaluated and discussed them, offering further clarification where necessary. Once this process was finished, annotators proceeded with the remaining 100 tweets, which were spot-checked to ensure quality. The practical implementation of the annotations was conducted using the LabelStudio tool. Although we display only one example in Figure 3, annotators were provided with a total of five examples.

Below, we outline the annotation guidelines as presented to the annotators:

Irony Definition

Our working definition covers both verbal and situational irony. We define verbal irony as a form of figurative language where the intended meaning of a message is different from its literal interpretation. Most often, irony involves a literal expression, such as "I love", that is intended to convey that the writer does not like something. However, irony can also manifest as a nonsensical or inappropriate reaction aimed at criticizing a previous statement. Common linguistic devices used to convey irony include hyperbole, understatement, analogy, and rhetorical questions. Sarcasm, a form of verbal irony specifically intended to ridicule or hurt, is also covered by this definition.

In general, ironic statements violate Grice's maxims of communication, which establish the expectations that communication should be (1) truthful, (2) relevant, and (3) appropriately informative (neither too brief nor overly detailed). When these maxims are deliberately broken, irony often arises.

To recognize that a statement is intended ironically, external background knowledge is typically required, as the text alone may not make this clear. In these experiments, our goal was to make this implicit knowledge explicit through human-generated explanations.

Guidelines for Explanations

When writing explanations, annotators were instructed to be as thorough as possible and not to assume any "obvious " knowledge. To help illustrate this, we proposed scenarios such as the "explain it to me like I'm 5" approach, or imagining the explanation being given to someone unfamiliar with the culture or language—such as an alien with no knowledge of this world. These scenarios were designed to emphasize that even knowledge that might seem obvious should still be made ex-

plicit. Additionally, each explanation should include reasoning that clearly contrasts the necessary background knowledge with the literal expression.

Given the subjective nature of irony and the possibility that different annotators might have varying interpretations, specific guidelines were provided for handling texts where the contrast was either too explicit or the annotator felt it didn't merit an explanation.

- *Skipping texts*: Annotators were instructed to skip a text if they did not personally find it ironic or if they were unable to explain the irony. Additionally, some tweets did not require background knowledge for understanding the irony, as the contrast between the literal and intended meaning was too obvious. In such cases, annotators could also skip the text, as no further explanation was necessary.

- *Modifying texts*: In instances where the contrast was too explicit to warrant an explanation, annotators had the option to modify the tweet by removing the overly explicit elements. After modifying the text, annotators were asked to explain the irony in the revised version rather than the original.

In total, 43 out of the 80 texts were modified, with an average of 1.79 words removed from the source text. While we acknowledge that this increases the complexity of the task, we ensured that the irony label for each modified text remained valid throughout the process.

A.2 Explanation Evaluation

The evaluation of the explanations involved two main tasks: (1) identifying whether the explanation was "human-like" and (2) ranking the explanations from best to worst. All annotations were conducted using the LabelStudio tool (Tkachenko et al., 2020-2022) by a separate group of annotators, different from those who originally wrote the explanations. Figure 5 provides a screenshot of the annotation tool used for this evaluation. A screenshot of the annotation tool for explanation evaluation illustrated with Figures 4 and 5.

Human-Like Identification In the first step, annotators were asked to determine which explanations appeared to be written by humans. Importantly, we did not reveal the exact number of human-written explanations in the dataset, though annotators were informed that a mix of human and machine-generated explanations was present, with at least one explanation written by a human. To further guide the annotators, we clarified that any explanation indistinguishable from a human-written one would meet the criterion for being "human-like".

Annotators were given the freedom to define their own criteria for recognizing human-like explanations. Interestingly, after completing the task, several annotators mentioned that they could easily identify the longest explanation as being machine-generated —specifically by GPT-4-Turbo— because because its level of elaboration was far beyond what a human would typically write.

Explanation Ranking For the second task, we granted annotators the same freedom to establish their own criteria as in the human-likeness evaluation. They were encouraged to apply their personal judgment in ranking the explanations, without relying any predefined guidelines. Several annotators expressed uncertainty about whether to prioritize human-like explanations over machine-generated ones. In response, we clarified that 'human-likeness" and overall quality are be treated as separate considerations from our perspective. The inclusion of this criterion was entirely up to their discretion but not required as a ranking factor.

B Prompt Templates For Explanation Generation

Below, we present the different prompts used in our experiments. For open-source models, we used placeholders such as [text-placeholder] for the input text and [expl-placeholder] for the explanation, which were replaced with the actual content during supervised fine-tuning (training). For inference, the final placeholder (i.e., the explanation) was left empty for the model to generate the appropriate output.

Additionally, in these examples, we used placeholders like [BoS] (beginning of sequence), [start-of-instruction], [end-of-instruction], and [EoS] (end of sequence). These placeholders were automatically replaced with the correct tokens, depending on the tokenizer used by the specific model (e.g., Llama or Mixtral), ensuring the appropriate tokens were applied where needed.

B.1 GPT-4-Turbo

Your task is to explain why tweets should be considered ironic. Make sure to base your explanation

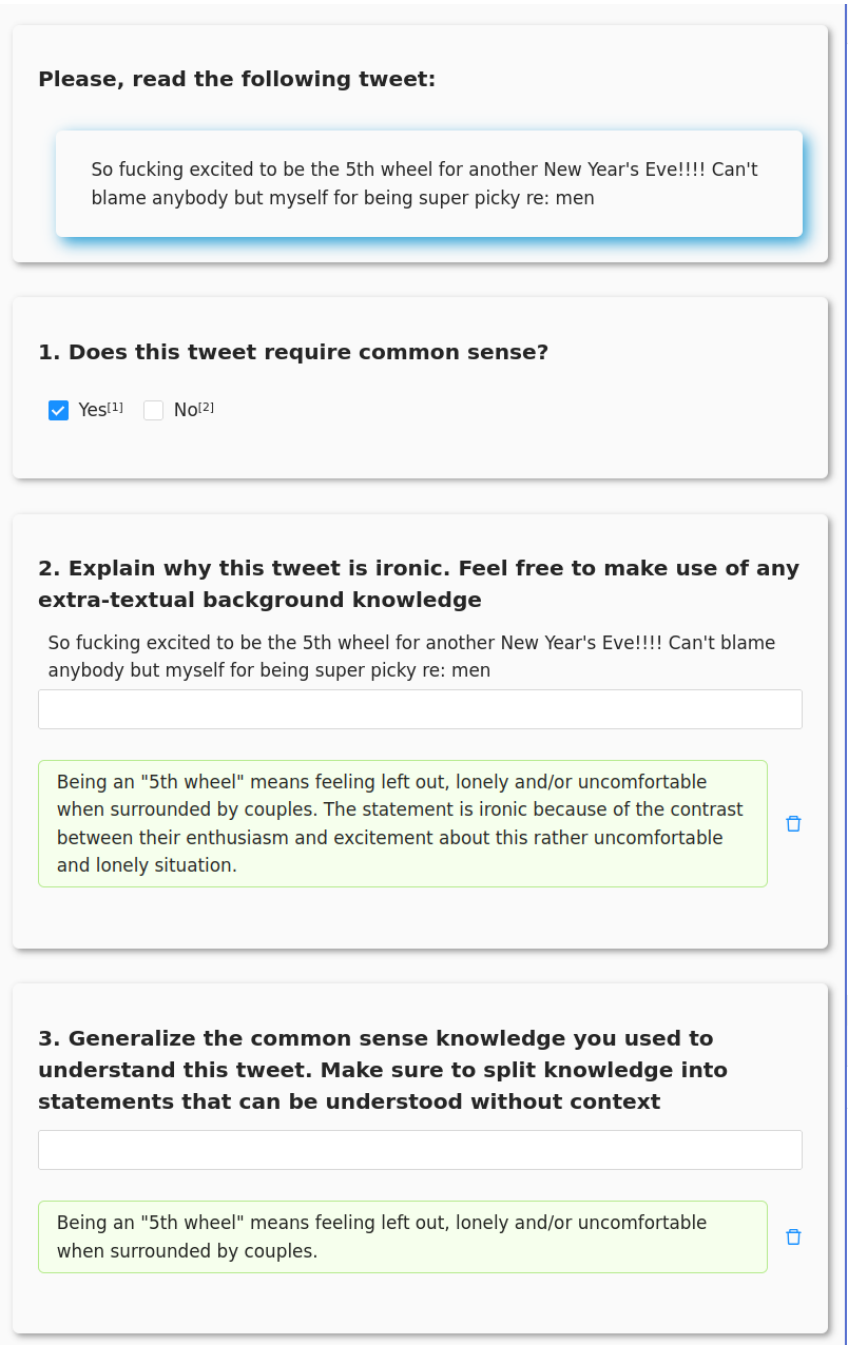


Figure 3: Annotation Example of Explanation Writing in LabelStudio.

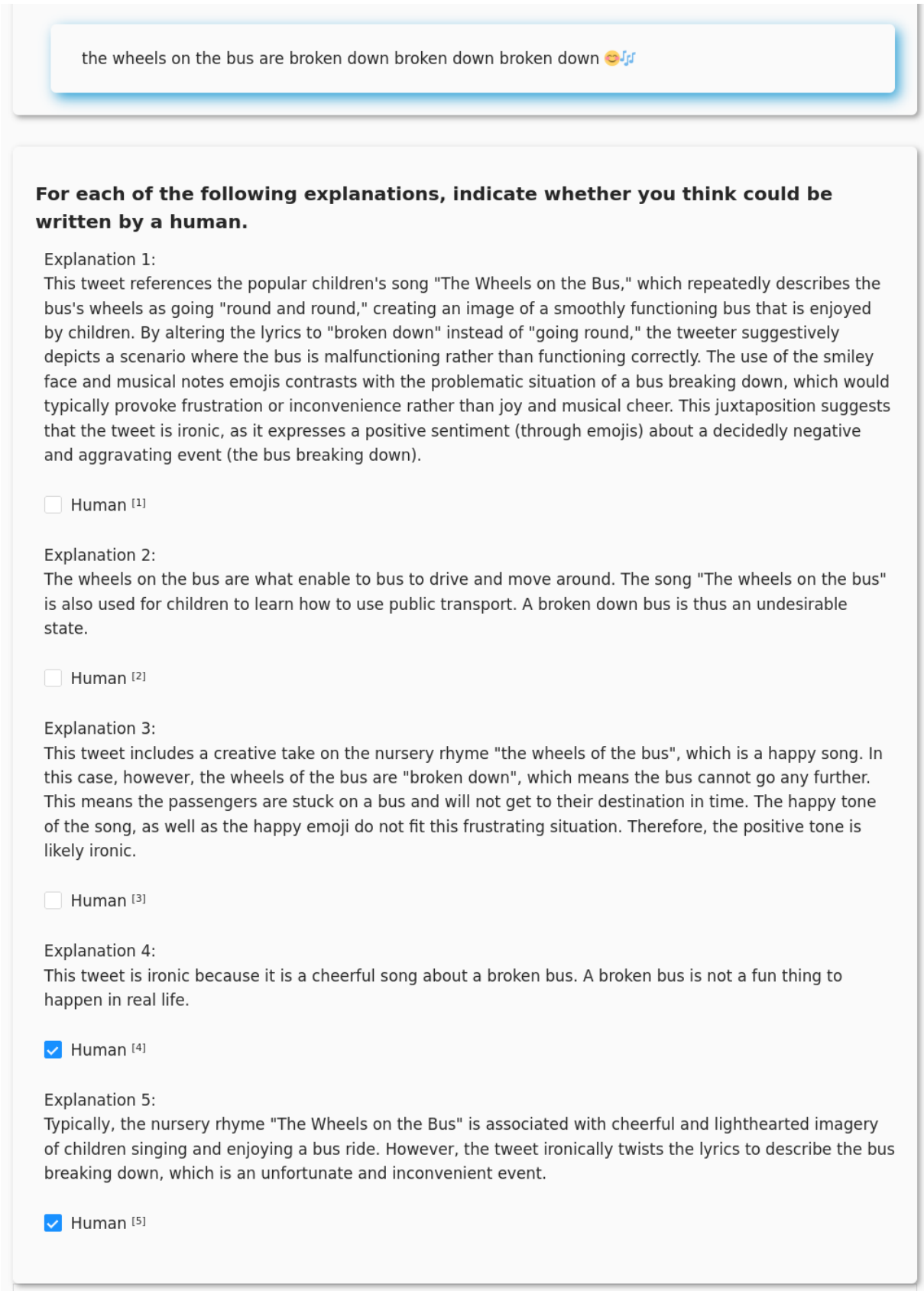


Figure 4: Example of the human-likeness annotation (before ranking) in LabelStudio.

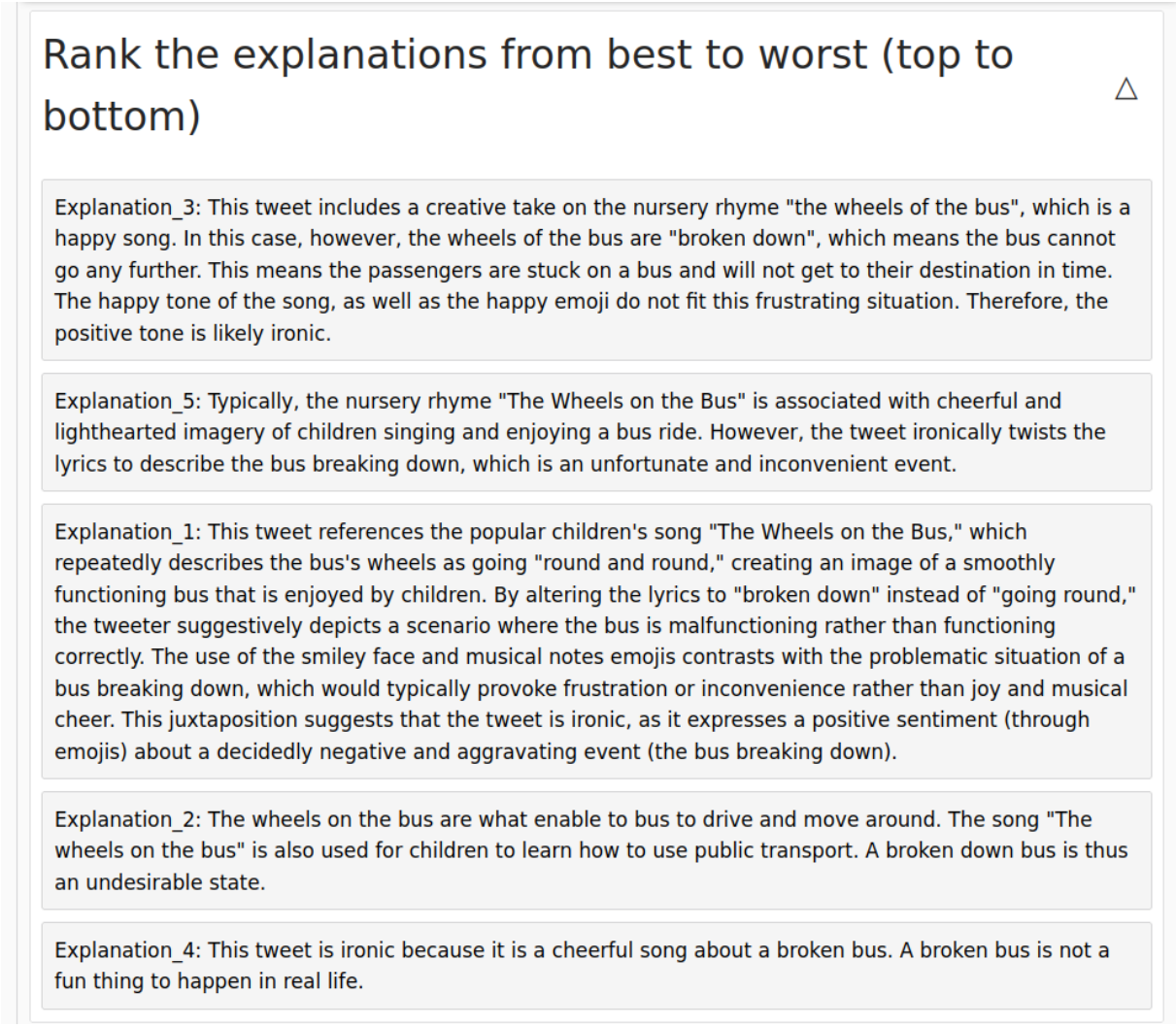


Figure 5: Example of the ranking annotation (after indicating human-likeness in LabelStudio.

on background knowledge that is not present in the text itself. This background knowledge can include common assumptions, factual knowledge and social conventions. Here are some examples of the expected output:
Example 1 -
Text: Loooovvveeeeeee when my phone gets wiped ### Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.
Example 2 -
###Text: @user Kohli showing all the guts of Scooby Doo. #CaptainCourageous ### Explanation: When people say that someone has “the guts to do something ”, it means they are fearless, confident and brave. The cartoon character Scooby Doo is know for his cowardice. Saying that someone is showing the guts of Scooby indicates that they are in fact not courageous or fearless at all. The evaluation “#CaptainCourageous” must therefore also be ironic.
Example 3 -
Text: Just received an email that I am the heir to the Nigerian Petroleum Corporation. I should immediately send them all my bank info. ### Explanation: The stereotypical phishing email from a supposed “nigerian prince” is a well-known fact. As most people are expected to know about this common phishing template, the users will most likely not follow up by walking into the trap and sending them all their bank info (which is exactly what the writer of this tweet writes they will do exactly).

New sample: ### Text: [text-placeholder] ### Explanation:

B.2 Open-Source Models

[BoS][start-of-instruction] You are an expert trained in identifying irony and sarcasm in social media text and explaining the underlying reasoning Your task is to explain why tweets should be considered ironic. Make sure to base your explanation on background knowledge that is not present in the text itself. This background knowledge can include common assumptions, factual knowledge and social conventions. Here is an example of the expected output:
Text: Loooovvveeeeeee when my phone gets wiped
Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.
Text : [text-placeholder]
Explanation : [end-of-instruction]
[expl-placeholder][EoS]

C Prompt Templates For Knowledge Extraction

C.1 GPT-4-o

You will be provided with an ironic tweet along with an explanations for why they are ironic. These explanations are based on extra textual background knowledge. Your task is to extract which knowledge is used in the explanations but which is not present in the text itself. Split up this information into multiple items where this is appropriate. Make sure that the knowledge you provide is mentioned explicitly in the explanation. Here are some examples of the expected output:
Example 1 -
Text: Loooovvveeeeeee when my phone gets wiped
Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic. ### Background knowledge:

1. When a phone gets wiped, all personal data and information is lost.

2. People do not like losing access to their personal data and pictures on their phone.

Example 2 -
Text: @user Kohli showing all the guts of Scooby Doo. #CaptainCourageous
Explanation: When people say that someone has "the guts to do something", it means they are fearless, confident and brave. The cartoon character Scooby Doo is know for his cowardice. Saying that someone is showing the guts of Scooby indicates that they are in fact not courageous or fearless at

all. The evaluation "#CaptainCourageous" must therefore also be ironic.
Background knowledge:
1. When people say that someone has “the guts to do something”, it means they are fearless, confident and brave.
2. The cartoon character Scooby Doo is known for his cowardice.
Example 3 -
Text: Just received an email that I am the heir to the Nigerian Petroleum Corporation. I should immediately send them all my bank info.
Explanation: The stereotypical phishing email from a supposed “nigerian prince”is a well-known fact. As most people are expected to know about this common phishing template, the users will most likely not follow up by walking into the trap and sending them all their bank info (which is exactly what the writer of this tweet writes they will do exactly).
Background knowledge:
1. The email from a Nigerian prince or heir is a well-known stereotype of phishing.
New sample:
Text: [text-placeholder]
Explanation: [expl-placeholder]
Background knowledge:

C.2 Open-Source Models

[BoS][start-of-instruction] For each of the following input texts, we have an explanation for why they contain irony. Your task is to identify and extract any background knowledge used for the explanation. Here is an example of the expected output:
Text: Loooovvveeeeeee when my phone gets wiped
Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.
Background knowledge:
1. When a phone gets wiped, all personal data and information is lost.
2. People do not like losing access to their personal data and pictures on their phone.
Now do this for the following text:
Text: [text-placeholder]
Explanation: [expl-placeholder]
Background knowledge:[end-of-instruction] [expl-placeholder][EoS]