# *I'm sure you're a real scholar yourself*: Exploring Ironic Content Generation by Large Language Models

**Pier Felice Balestrucci**[1]*, **Silvia Casola**[2]⋆, **Soda Marem Lo**[1]⋆,
**Valerio Basile**[1], **Alessandro Mazzei**[1]
[1]Computer Science Department, University of Turin, Italy
[2]MaiNLP & MCML, LMU Munich, Germany
{pierfelice.balestrucci, sodamarem.lo, valerio.basile, alessandro.mazzei}@unito.it
s.casola@lmu.de

## Abstract

Generating ironic content is challenging: it requires a nuanced understanding of context and implicit references and balancing seriousness and playfulness. Moreover, irony is highly subjective and can depend on various factors, such as social, cultural, or generational aspects. This paper explores whether Large Language Models (LLMs) can learn to generate ironic responses to social media posts. To do so, we fine-tune two models to generate ironic and non-ironic content and deeply analyze their outputs' linguistic characteristics, their connection to the original post, and their similarity to the human-written replies. We also conduct a large-scale human evaluation of the outputs. Additionally, we investigate whether LLMs can learn a form of irony tied to a generational perspective, with mixed results[1].

*Warning: Some examples shown in this paper contain offensive language, discriminatory remarks, and slurs.*

## 1 Introduction

Irony is a complex linguistic device that exploits semantic inversion, conveying the opposite of what is believed and what actually is. Irony is thus a complex phenomenon, the generation and recognition of which requires a nuanced understanding of the context, the tone, and the underlying meaning (Muecke, 1970). Moreover, recognizing irony involves various subjective factors and can be culture-specific and based on shared cultural references, norms, and world models (Gibbs and Colston, 2007).

Automatically generating ironic content is thus an interesting but challenging task, which is still difficult for state-of-the-art generative systems. This paper explores whether Large Language Models (LLMs) can generate content perceived as ironic by human readers. Specifically, we focus on contextual irony generation and generate ironic replies to a given social media post.

Leveraging human-written short post-reply conversations, we first fine-tune two models to generate ironic and not-ironic replies to a post. Then, we deeply analyze their outputs from a linguistic perspective, focusing on their intrinsic characteristics as well as on their relation to the original post and the similarity to the human-written replies. To explore whether such replies are perceived as ironic by human annotators, we perform a large-scale human evaluation. While less linguistically rich than human-written replies, ironic model outputs present interesting references to implicit context and are characterized by sarcastic remarks. Moreover, replies generated by the model fine-tuned with ironic post-reply pairs are perceived as more ironic by humans in the vast majority of cases when compared to those generated by a model trained on non-ironic data.

Inspired by previous work that identifies age as a highly polarizing dimension in the recognition of irony (Casola et al., 2024), we then try to create models that generated content specifically identified as ironic by old or young annotators, with mixed results.

In short, our contributions are the following:

- We explore contextual reply generation using Large Language Models (Section 4).
- Given a model trained on ironic post-reply pairs and one trained on non-ironic data, we perform a linguistic analysis showing the differences in their outputs and the relationship with the original post and the corresponding human-written counterparts (Section 4.3)
- We perform a large-scale human evaluation,

*Equal contribution. The work of SC was done while at the University of Turin.

[1]The data are released under the Creative Commons Attribution Non Commercial 4.0 International license and available at: https://github.com/DipInfo-Unito/IronicContentGeneration



*Findings of the Association for Computational Linguistics: EACL 2024*, pages 14480–14494

showing that outputs are, in fact, considered ironic in a large portion of cases (Section 4.4). The code, annotations, and the related post-reply pairs are available for future research.

- Inspired by previous work linking irony detection to annotators' sociodemographic characteristics, we explore whether we can build age-specific models targeted to a certain age group (Section 5).

## 2 Related Works

While humor and sarcasm generation is becoming increasingly explored in the NLG community (Amin and Burghardt, 2020; Chakrabarty et al., 2020; Oprea et al., 2022), few studies focus on irony (Zhu et al., 2019). As highlighted by Loakman et al. (2023), a significant amount of work in the field centers on puns generation, especially for phonetics and word senses (Tian et al., 2022). On the other hand, works in unsupervised sarcasm generation based on sarcasm theory are emerging (Zeng and Li, 2022; Mishra et al., 2019; Chakrabarty et al., 2020), where the linguistic phenomenon characteristics (e.g., reversal of valence, semantic or context incongruity) are used to produce sarcastic messages. Focusing on irony generation can be of high importance especially when thinking about the benefits of verbal irony as an instrument to mediate and negotiate boiling emotions (Pfeifer and Pexman, 2023).

Researchers in irony detection have recently demonstrated annotators' cultural backgrounds affect their data labeling choices (Casola et al., 2024). This body of work shows how taking into account the subjectivity of the annotators turned out to be effective in humor extraction (Bielaniewicz et al., 2022) and irony detection (Frenda et al., 2023; Casola et al., 2023).

Personalizing response generation has been extensively studied across various domains, evolving significantly with the advent of large-scale social media data and the success of sequence-to-sequence frameworks (Serban et al., 2016), followed by advancements in Large Language Models (Chen et al., 2024). Numerous models have been developed to incorporate user-specific information into dialogue systems, thereby improving their responsiveness and relevance (Wu et al., 2021).

Demographic characteristics and individual perception can play a key role, especially when generating highly subjective language phenomena, such as irony, humor, and sarcasm. Recent works have pointed out the influence of sociodemographic factors both in human annotation (Bender and Friedman, 2018) and evaluation (Loakman et al., 2023) of pragmatic phenomena, highlighting the importance of reporting them and taking into account their influence in the Machine Learning pipeline. A demographic-based approach to generating humor is found by Garimella et al. (2020) who proposed a location-specific humor framework, collected a dataset, and hired US and Indian annotators.

We noticed a general scarcity of datasets that allow us to consider demographic backgrounds in generating irony. Thus, we used the English Perspectivist Irony Corpus (Frenda et al., 2023), originally designed for irony detection, for irony generation, containing post-reply pairs suitable for a contextual irony generation.

| | | **ironic** | **not ironic** |
|---|---|---|---|
| Labels | | 31% | 69% |
| Annotations | | 4456 | 9716 |
| Tokens | Post | $33_{\pm 59}$ | $36_{\pm 59}$ |
| | Reply | $20_{\pm 24}$ | $23_{\pm 27}$ |

Table 1: EPIC Dataset Characteristics.

## 3 Dataset

EPIC (English Perspectivist Irony Corpus) (Frenda et al., 2023) is a disaggregated corpus of 3,000 short conversations evenly collected from Reddit and Twitter. The data was collected across five English-speaking countries, i.e., Australia, India, Ireland, the United Kingdom, and the United States. Each instance was annotated by $\sim 5$ crowd-sourced annotators, balanced across gender and nationality (chosen to match the data geographic origin). The annotators were asked to determine whether the reply was ironic in the context of the post. The resulting $14,172$ annotations were shared in a desegregated form and complemented with annotators' demographic metadata (including gender, age, ethnicity, student, and employment status). Table 1 summarizes the dataset characteristics.

## 4 Ironic Reply Generation

In this section, we aim to determine whether LLMs can generate replies perceived as ironic by human readers. Starting from the EPIC disaggregated dataset, we use instance-based majority vote to ag-

gregate the labels and exclude instances for which no majority is reached. Then, we select an LLM and generate ironic outputs. We examine the outputs to identify any linguistic patterns that provide insights into the irony of the sentences. Finally, we have humans evaluate the generated outputs.

We performed the bulk of the experiments using Mistral 7B (Jiang et al., 2023)[2], a state-of-the-art pretrained LLM for English, as our base model. Some preliminary experiments were also performed by using Llama2-7B (Touvron et al., 2023); output examples are reported in Appendix **??**. To collect enough ratings to ensure that the results were meaningful, we then decided to work with a single model, i.e., Mistral.

### 4.1 Zero-shot Reply Generation

We first tried to generate replies in a zero-shot manner. Specifically, we used the following prompt:

> **Instruction**: You are given in input (Input) a post extracted from social media conversations. Provide in output (Output) an ironic reply.
>
> **POST**: <Text of the post>
>
> **REPLY**:

where "ironic" is substituted with "serious" when generating non-ironic replies.

Appendix A contains some examples of the output. The replies generated in this preliminary phase were largely unsatisfactory. We noticed that the outputs tended to be very long and typically contained numbered lists, repetitions, or a long sequence of post-reply pairs. This is partly expected since the model is not instruction-tuned nor fine-tuned on specific data and tasks and thus lacks the ability to follow instructions. Thus, we decided to fine-tune the model using the data available.

### 4.2 Fine-tuned Reply Generation

**Model fine-tuning** Starting from our base model, we performed Low-Rank Adaptation (LoRA) (Hu et al., 2022). We partitioned the aggregated data and used 80% of the instances for training and 20% for testing. To train the models, we relied on the original dataset annotations, using either the ironic post-reply pairs or the non-ironic ones. Thus, we trained two separate models: one to generate ironic responses to the original posts (referred to as *IRO* in the following) and another to generate non-ironic responses (referred to as *NIRO*). We trained the two models for five epochs (all training details and hyper-parameters are available in Appendix B). The prompt is identical to the one used for the zero-shot generation, and we added the human-written reply for training.

Since the dataset is extracted from social media conversations, we considered whether to anonymize or remove user mentions (in the form of "@usernickname"), which appear frequently, and whether to mask the input portion (instruction and post) to compute the training loss. Thus, we trained four distinct models to generate ironic content and four to generate non-ironical content, exploring all the combinations with masking and anonymization.

**Model selection and preliminary evaluation** To better explore the nature of the outputs and given that the task of contextual ironic and non-ironic reply generation is extremely open (i.e., a single post can lead to a variety of possible replies), we performed a preliminary human evaluation. This involved manually annotating a subset of 100 randomly selected post-reply pairs for each model. Specifically, out of these 100 pairs, 50 were originally annotated by humans as ironic and 50 as non-ironic. Three authors of this paper annotated the model outputs, assessing whether they had severe issues (e.g., being empty, containing textual loops, or being identical to the original post) or were satisfactory. For the satisfactory outputs, they evaluated whether the replies could plausibly have been generated by a person with an ironic intent (for an ironic model) or a non-ironic intent (for a non-ironic model). The annotators used a spreadsheet with different colors for the annotation. Table 2 shows an example of an annotation.

We collected 553 posts from the original dataset and manually removed those not in English or where no semantic content was found (for example, only containing mentions and hashtags). This left us with 523 posts, for which we generated a reply with IRO and NIRO.

To choose the two models, we opted for a trade-off between satisfactory and problematic replies, favoring the models that performed better. The models resulting in better performances were those where the input was masked for training, and the user mentions were removed. Appendix E shows the average percentage of correct, wrong, and problematic answers labeled by the three annotators,

[2] mistralai/Mistral-7B-v0.1

| **Post:** I'm trying to read reddit with one eye. | | |
|---|---|---|
| **Model** | **Model reply** | **Label** |
| $\text{NIRO}_{\text{Mask-noMentions}}$ | I'm trying to watch TV with one eye. | Match |
| $\text{NIRO}_{\text{mask-Mentions}}$ | r/perfecteye | No Match |
| $\text{NIRO}_{\text{noMask-noMentions}}$ | I'm trying to read this with one eye | Match |
| $\text{NIRO}_{\text{noMask-Mentions}}$ | | Problematic |
| $\text{IRO}_{\text{mask-noMentions}}$ | I'm trying to read this with one hand up my ass. | Match |
| $\text{IRO}_{\text{mask-Mentions}}$ | 2 eyes are better than 1 | No Match |
| $\text{IRO}_{\text{noMask-noMentions}}$ | I'm trying to read reddit with one eye and a broken finger | Match |
| $\text{IRO}_{\text{noMask-Mentions}}$ | 2 eyes are better than 1 | No Match |

Table 2: Examples of annotations for the model replies. The models are categorized based on irony (IRO for ironic, NIRO for non-ironic), masking (Mask for masked, noMask for non-masked inputs), and whether user mentions are removed (noMentions) or anonymized (Mentions). Replies consistent with the model's goal (e.g., an ironic reply from an ironic model) are labeled as "Match". Conversely, replies that do not fit the model's goal are labeled as "No Match". In cases where the replies were problematic (e.g., blank responses, or mere repetitions of the original post), they were labeled as "Problemic".)

who obtained an Inter Annotator Agreement (computed by Cohen's Kappa coefficient (Cohen, 1960)) of 0.59, a score comparable with several other studies on irony detection (Van Hee et al., 2018; Cignarella et al., 2018). From a qualitative point of view, the ironic model outputs tended to show sharp irony and sarcasm (in some cases, with offensive or racist replies), following the style of the data.

### 4.3 Linguistic Analysis

To better understand the models' output characteristics, we performed a linguistic analysis of their generated replies. We aimed to study the differences between the two models' outputs and how the provided context interacts with the generated answers. To do so, we analyze the human-written and generated replies in the test set, composed of 410 non-ironic and 130 ironic instances.

As illustrated in Table 3, both IRO and NIRO tend to produce shorter replies when compared to the human-written gold standard. Moreover, as measured using the type/token ratio, models' replies have a lower lexical variation. Intuitively, the results could show that the generated replies are more stereotyped and less creative than the original human-written ones. Furthermore, we looked into three types of irony markers (Karoui et al., 2017), i.e., interjections, negations (as linked to context-shift and rhetorical questions strategies), and named entities, using SpaCy and the SpaCy-udpipe English models. Both interjections and negations are consistently more frequent in generated replies.

As irony tends to rely on the interlocutor's knowledge of specific events or topics, we also looked into the number of named entities in the original post (656) and the human-written (433) and generated replies (248 and 315, respectively, for IRO and NIRO). Although the latter shows a lower number than human-written replies, in 83% (IRO) and 70% (NIRO) of the cases, when a named entity appeared in the post, the same was true for the generated reply (Example 1).

| | **Post** | **Human replies** | | **Model replies** | |
|---|---|---|---|---|---|
| | | ironic | non-ironic | IRO | NIRO |
| Tokens | 28 | 17 | 20 | 15 | 14 |
| TTR | .25 | .42 | .31 | .19 | .18 |
| Interjections | 56 | 18 | 45 | 37 | 74 |
| Negations | 186 | 31 | 122 | 168 | 303 |

Table 3: Number of tokens, type/token ratio, and average number of tokens interjections and negations for the original post and the human-written labeled as ironic and non-ironic and replies generated by IRO and NIRO models.

(1) [Post] I'd fire him for going to Greggs period

[IRO reply] He's already been sacked from Greggs for stealing the sausage rolls

To explore whether the generated replies resembled short punch lines, we studied the use of nominal utterances and the overall syntactic complexity of the replies generated by the models. Nominal utterances were rare in both models' replies, for a total of 9 and 26, respectively, for IRO and NIRO models. Looking at the height of the syntactic tree of the generated sentences, we observe how ironic replies tend to have a higher tree than non-ironic ones, oscillating between 3 and 6 (Figure 2). This

| | Human replies | | Model replies | |
|---|---|---|---|---|
| | ironic | non-ironic | IRO | NIRO |
| Post sim | 0.584 ±0.206 | 0.585 ±0.234 | 0.614 ±0.211 | 0.578 ±0.228 |

Table 4: Average text similarity and standard deviation with the original post of the human-written and generated replies labeled as ironic and non-ironic, and replies generated by IRO and NIRO models.

might indicate that the ironic replies tend to be syntactically richer and more complex when compared to their non-ironic counterparts.

Finally, we were interested in exploring the relationship between the post and generated replies to understand whether the models tended to mimic the original message. Thus, we computed the text similarity between the posts and the corresponding human-written and generated replies. To do so, we used a Word2Vec model (Mikolov et al., 2013) to obtain 300-dimensional vectors of the tokens and averaged them. Then, we computed the vector similarity by using cosine similarity. Looking at the distribution in Figure 1, both IRO and NIRO behave analogously to human replies. Ironic replies tend to be slightly more similar to the post, and non-ironic ones tend to be less similar with respect to human texts. To explore this insight further, we also computed the average similarity for human replies labeled as ironic and non-ironic in the original dataset (Table 4); generated replies follow the same pattern for both cases.

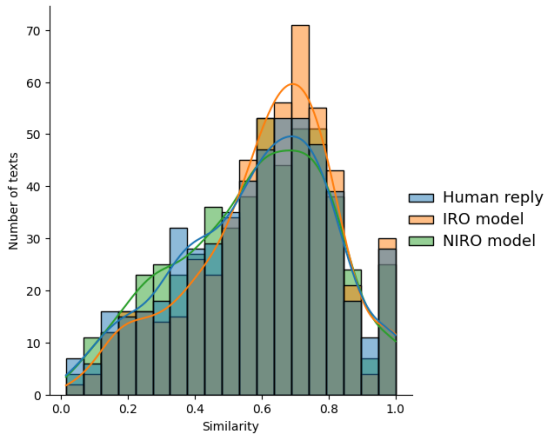


Figure 1: Similarity between post and replies (human written, and generated by IRO and NIRO models).

### 4.4 Human evaluation

We conducted an extensive human evaluation campaign to determine if the model fine-tuned with

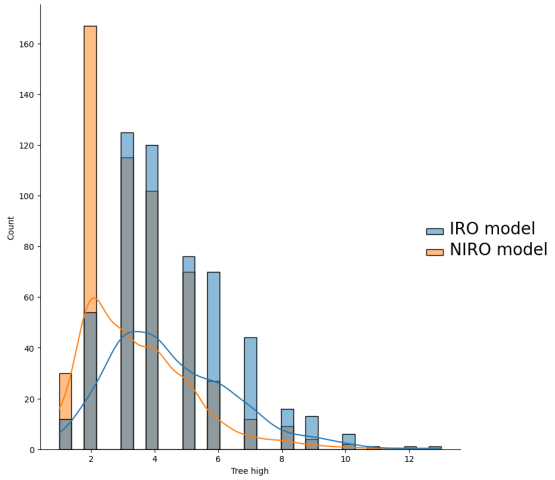


Figure 2: Tree height for IRO and NIRO replies.

ironic data generates responses perceived as more ironic by human annotators than the model fine-tuned with non-ironic data. To do so, given a post, we showed human annotators the replies generated with both IRO and NIRO. Each annotator was asked to indicate on a 5-point Likert scale (Likert, 1932) whether they found the reply ironic. Figure 3 shows the interface used for the evaluation. For each post, replies generated by both models were shown to users in random order. Each annotator was asked to annotate up to 30 pairs; annotators had a 10% chance to be prompted with attention questions in the form of "Please select [choice]".

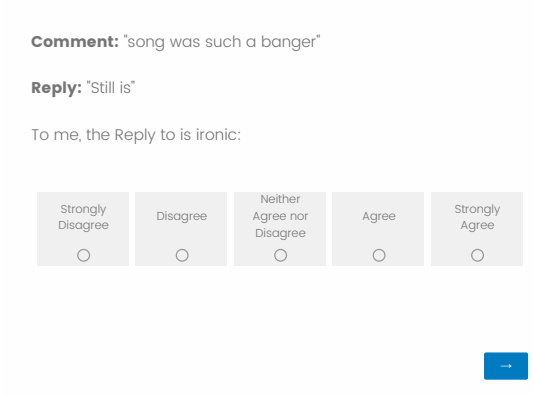


Figure 3: Qualtics human annotation interface for the irony generation task.

We used Prolific[3] to hire native English-speaking annotators and Qualtrics[4] to distribute the questionnaire, also collecting their demographics (reported in Appendix G). We set a rate of £9/hour, however since workers were slightly faster than expected

[3] https://www.prolific.com
[4] https://www.qualtrics.com

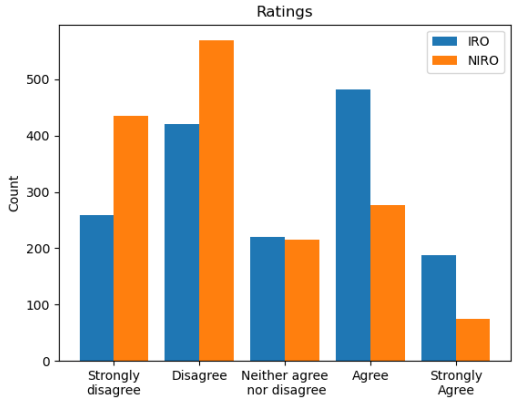


Figure 4: Comparison of the human ratings for IRO and NIRO.

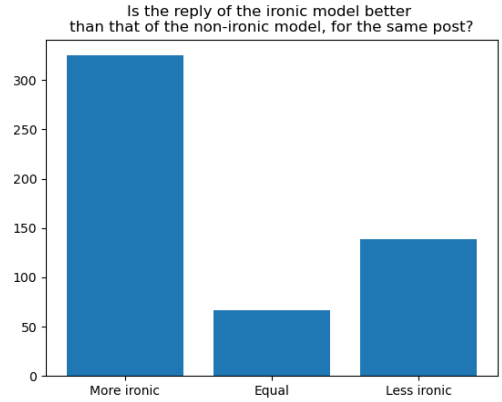


Figure 5: Direct comparison between instance-level aggregated scores obtained by IRO and NIRO. "More (Less) ironic" means that, for the same post, the reply of IRO received a higher (lower) score than that of NIRO.

the average compensation was £13, and they took around seven minutes to complete the task. We collected 119 responses to the questionnaire. Only one annotator failed the attention questions (all others were corrected in 100% of the cases), and we removed their annotations. This left us with a total of 3,242 instances.

Figure 4 illustrates the disaggregated rankings provided by annotators for replies generated by models fine-tuned on ironic and non-ironic data. As expected, IRO tends to generate replies perceived as more ironic by human annotators, who tend not to perceive irony, in the majority of cases, for replies generated by NIRO.

To directly compare scores in an instance-based manner, we mapped the Likert Scale to a $-2$ to $+2$ numerical scale (where $-2$ stands for "Strongly disagree" and $+2$ for "Strongly agree") and aggregated results by summing the individual rating. Figure 5 shows the number of instances for which the aggregated rating was higher, equal, or lower for IRO. Specifically, of all the annotated posts, 325 of the replies generated by IRO were marked as more ironic, 67 as equally ironic, and 130 as less ironic compared to those generated by NIRO, given the same post. The responses of the model trained on ironic responses are thus perceived as more ironic, on average, than those generated by NIRO. This result indicates that our fine-tuned model can generate content that is significantly perceived as ironic by humans.

We further explored the analysis by investigating instances where NIRO-generated responses were deemed equally or more ironic than those from IRO. To do this, we employed TextBlob[5] to compute the sentiment polarity of each sentence, including the original post text, replies from IRO, and replies from NIRO. This allowed us to assess whether sentiment polarity, ranging from $-1$ (negative) to 1 (positive), potentially influences the ironic nature of responses produced by both models.

Figure 6 presents three violin plots depicting different scenarios based on annotator judgments. In the scenario where annotators found IRO replies to be more ironic, the mean sentiment polarity tends to be higher than post and NIRO replies. Moreover, the sentiment polarity distribution for IRO responses suggests a tendency towards more positive sentiment. Conversely, when annotators judged NIRO replies as more ironic, the three violins are very similar and the sentiment polarity of NIRO replies shows more outliers, suggesting greater variability or inconsistency in sentiment. In cases where annotators rated both models as equally ironic, the sentiment polarity mean of the original post and IRO replies are similar. However, the sentiment polarity distribution of NIRO replies tends to peak around neutral values, indicating a tendency towards less extreme sentiment expressions. These observations indicate that when the sentiment of the original post is more positive, IRO tends to generate responses with higher positive sentiment than NIRO.

[5] https://textblob.readthedocs.io

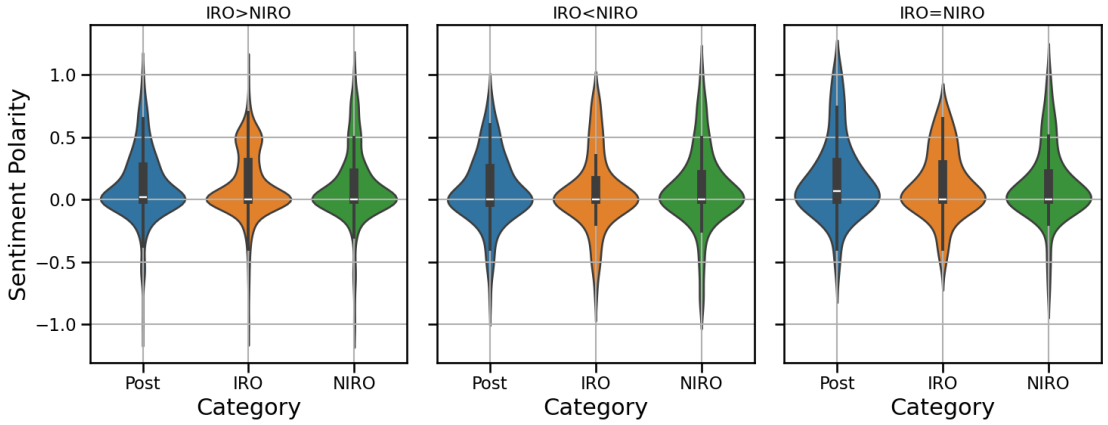


Figure 6: Sentiment polarity distributions. Left: IRO replies rated as more ironic; Middle: NIRO model replies rated as more ironic; Right: replies rated as equally ironic.

## 5 Building age-specific models

Previous work has shown that irony detection is a highly subjective task (Reyes et al., 2013), where the annotators' background plays a crucial role. Specifically, Casola et al. (2024) highlighted the importance of considering annotators' sociodemographic characteristics. In their analysis, they found the annotators' age is one of the sociodemographic dimensions corresponding to the highest polarization in the recognition of irony. Example 2 shows a case from the EPIC dataset where GenY and Boomer's annotators labeled the reply as ironic and non-ironic respectively.

(2) [Post] When you're young, work to learn don't work to earn. You should prioritise study over work. Go full time uni and part time work.

[Human Reply] work to learn don't work to earn What kind of boomer shit is this? Young people still have rent to pay.

To understand whether we could encode an age-specific perception of irony into an LLM, we trained two age-specific models to generate ironic replies. Starting from the disaggregated annotations available in the original dataset, we partitioned the original annotators into young (i.e., < 42 years old) and old (i.e., >= 42 years old), using the available metadata[6], then we constructed two separate semi-aggregated gold standards. We discarded instances for with no majority was reached and trained the models on ironic instances only. Given both gold standards, we fine-tuned two separate models on the age-specific ironic instances, using the same method discussed in Section 4.2[7]. We will refer to these models as *Y-model* and *O-model*. Given 455 test posts, we generated the corresponding replies using both models. After manually excluding 17 post-reply pairs not in English and removing 21 identical generated replies from both models, we obtained a new corpus of 417 sentences to be analyzed and annotated.

### 5.1 Linguistic Analysis

As done in Section 4, we extracted information about the number of tokens, type/token ratio, interjections, negations, and named entities. We compared human replies rated as ironic by young annotators (98 instances) and old annotators (85 instances) from EPIC against replies generated by Y-model and O-model. Results in Table 5 confirm the previous consideration about human ironic answers tending to have a lower number of tokens, closer to generated replies, and these latter having a consistently lower linguistic variation expressed in terms of type/token ratio. In both human and model replies, the average number of interjections is higher for the older generation, which gives insights into linguistic differences linked to age variation. On the other hand, almost no differences are present regarding the use of negations, but they are consistently higher in generated replies. We find a similar pattern with named entities corroborating the hypothesis presented in Section 4 about how the post tends to trigger references to contextual

[6] In EPIC, the annotators were classified as Boomers (age >= 58 ), GenX (age between 42 and 58), GenY (age between 26 and 42), and GenZ (age < 26). We simplify this original division. Note that any hard split of the age label is somewhat arbitrary and a reductionist approach to true variation; however, our analysis is limited by the metadata provided in the original dataset.

[7] We used the same base model, masked the input when computing the loss, and removed the user mentions.

knowledge in generated answers. Finally, we computed the average similarity between post and both human and generated replies. Results in Table 6 show that human and model replies tend to have the same average similarity, confirming the results in Section 4.3, while the two age-specific models do not seem to show significant differences.

### 5.2 Human evaluation

We conducted a human evaluation campaign to determine whether younger (older) individuals found the replies generated by their age-correspondent model more ironic. We recruited native English-speakers using Prolific and we administered the questionnaires via Qualtrics. A total of 102 annotators were recruited, half with an age above 42 and half under 42. We collected all their demographics, reported in Appendix G. Each annotator was asked to evaluate 25 post-reply pairs, having to choose which answer they found more ironic between the Y-model and O-model generated replies. Among these 25 questions, 10% were attention-check questions in the form of "Please select [choice]". No annotators failed, resulting for a total of 2,063 annotations. Similarly than in the first experiment, the annotators were faster than expected, the task took an average of 9 minutes, and was paid around £13 per hour.

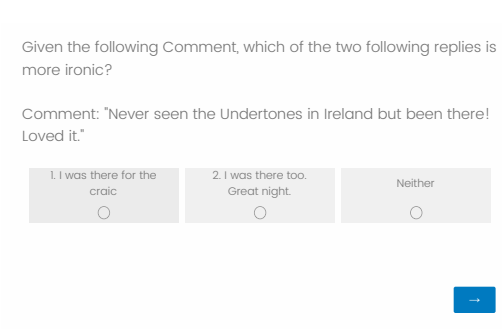


Figure 7: Qualtrics human annotation interface for the age-specific task.

| | Post | Human replies | | Model replies | |
|---|---|---|---|---|---|
| | | Old | Young | O-model | Y-model |
| Tokens | 29 | 18 | 16 | 12 | 15 |
| TTR | .26 | .47 | .46 | .23 | .19 |
| Interjections | 53 | 25 | 19 | 33 | 29 |
| Negations | 231 | 25 | 29 | 121 | 121 |
| Named Entities | 588 | 59 | 56 | 174 | 282 |

Table 5: Number of tokens, type/token ratio, and average of tokens interjections and negations for the original post and the human-written labeled as old and young and replies generated O-model and Y-model.

| | Human replies | | Model replies | |
|---|---|---|---|---|
| | Old | Young | O-model | Y-model |
| Post | 0.58 ±0.209 | 0.57 ±0.22 | 0.60 ±0.214 | 0.61 ±0.214 |

Table 6: Average text similarity and its standard deviation between the post and human-written and generated replies.

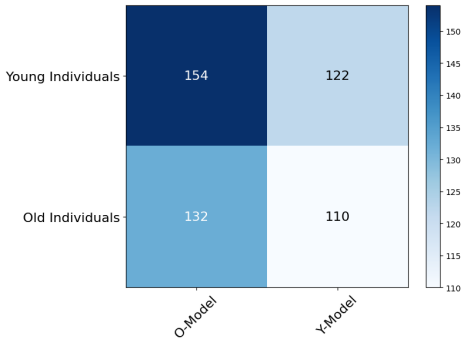


Figure 8: Comparison between the average preferences of the Young and the Old when tasked with choosing between the replies of Old and Young models.

Given a post, an annotator had to choose among three possible answers (Figure 7). The first two options corresponded to Y-model and O-model outputs, in random order; the third option was available if they considered both previous answers as non-ironic. Looking at the overall results, we noticed that, on average, older individuals preferred outputs generated by the O-model, while younger individuals displayed a similar preference. Moreover, in 175 cases for old and 141 cases for young annotators, none of the replies were considered ironic. All the results are depicted in Figure 8.

Considering how younger and older individuals more frequently preferred the O-model outputs, we analyzed how often the two groups agreed on the same instances by expressing the same preference. Approximately 113 instances were annotated similarly by both groups (46 generated by the Y-model and 67 by the O-model), and 75 instances were annotated by both as not being ironic. Thus, 188 instances (45.09% of the total) were annotated identically. It thus appears that there is no correlation between young or old annotators preferring replies from Y-model and O-model, respectively, and both groups tend to agree on whether a reply is more ironic or not. We believe the small size of the dataset, and the limited annotation details provided did not allow us to build models able to learn

age-specific aspects of irony. Future work in this direction should focus on collecting task-specific data, and provide a more fine-grained annotation.

## 6 Conclusion

This paper investigates LLMs' capabilities in generating ironic replies given a specific social media post. We first presented strategies for training LLMs to generate ironic content, performed a linguistic analysis, and conducted a human evaluation through crowdsourcing. We found that a model trained on ironic data can generate outputs that, though less rich from a linguistic perspective when compared to human-written text, are typically perceived as ironic by humans. Furthermore, we conducted linguistic analyses to elicit differences in linguistic patterns between ironic and non-ironic replies and the contextual outputs of the models. The generated sentences and the associated evaluations, which we will release with this paper, will contribute as a resource for irony classification and human- vs machine-generated irony detection.

We then tested the hypothesis that LLMs could generate content perceived differently based on the age of the annotators. This choice was motivated by finding in (Casola et al., 2024) that different annotators' ages result in high polarization in the recognition of irony. Again, we conducted a linguistic analysis to elicit the various differences between responses considered ironic by both younger and older populations, compared to the original sentences in the EPIC corpus and those produced by two new models trained to generate ironic content for younger and older populations. The results of this experiment do not confirm the initial hypothesis, and we believe this might be due to the small size of the dataset and to its design; we will also consider different age ranges or modifying the experiment's setting, such as choosing different questions.

## 7 Limitations

We believe that this work represents a step forward in understanding the capabilities and limitations of LLMs in generating ironic content. However, we acknowledge certain limitations of this study. While we could have investigated the ability to generate ironic responses using a variety of LLMs, the computational resources required to train these models on large datasets are substantial, which constrained our efforts. Additionally, we recognize the limitation of having used a relatively small dataset. Nonetheless, we want to emphasize that no other datasets currently exist that represent irony in relation to sociodemographic data. Moreover, we acknowledge that the design used to build the age-specific models is limited by the data available, and further work in perspective-specific irony generation should focus on building datasets specifically designed for these goals. We did not balance workers on Nationality, Ethnicity, Student and Employment status, which led to an over-representation of WEIRD population (Western, Educated, Industrialized, Rich, Democratic).

## 8 Ethical Considerations

The dataset annotations, crucial to the research presented in this paper, were provided by numerous annotators recruited and compensated through Prolific, a crowdsourcing platform we specifically selected due to its emphasis on the fair and ethical treatment of workers. Workers were paid at least £9/hour and up to £13 in most cases. All tasks were relatively short (requiring less than 10 minutes). All workers were allowed to withdraw from the study at any time and were previously informed about the potential discriminatory and offensive content of the text. They were also aware of the use of their data, as shown by the Informed consent shown in Appendix F. We acknowledge that mimicking patterns found in social media and the EPIC dataset, in particular, the generated answers might contain toxic, offensive, or discriminatory content, including slurs. For this reason, we do not plan to publish the associated models.

## Acknowledgement

We are grateful to the reviewers for their valuable feedback and suggestions. We thank Professor Cristina Gena (University of Turin) for her insightful suggestions for the design of the human evaluation interfaces.

This work was partially funded by the 'Multilingual Perspective-Aware NLU' project in partnership with Amazon Alexa and was partially supported by "HARMONIA" project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 under the NextGenerationEU programme.

## References


Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Julita Bielaniewicz, Kamil Kanclerz, Piotr Miłkowski, Marcin Gruza, Konrad Karanowski, Przemysław Kazienko, and Jan Kocoń. 2022. Deep-sheep: Sense of humor extraction from embeddings in the personalized context. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 967–974.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Silvia Casola, Soda Marem Lo, Valerio Basile, Simona Frenda, Alessandra Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507, Singapore. Association for Computational Linguistics.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13650–13665, Torino, Italia. ELRA and ICCL.

Alessandra Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). pages 26–34.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Aparna Garimella, Carmen Banea, Nabil Hossain, and Rada Mihalcea. 2020. "judge me by my size (noun), do you?" YodaLib: A demographic-aware humor generation framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2814–2825, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Raymond W. Gibbs and Herbert L. Colston. 2007. *Irony in Language and Thought: A Cognitive Science Reader*. Lawrence Erlbaum Associates, New York.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain. Association for Computational Linguistics.

Rensis Likert. 1932. *A technique for the measurement of attitudes / by Rensis Likert.* Archives of psychology ; no. 140. [s.n.], New York.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6144–6154, Hong Kong, China. Association for Computational Linguistics.

D. C. Muecke. 1970. *Irony and the Ironic*. Methuen, London.

Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Should a chatbot be sarcastic? understanding user preferences towards sarcasm generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7686–7700, Dublin, Ireland. Association for Computational Linguistics.

Valeria Pfeifer and Penny Pexman. 2023. When it pays to be insincere: On the benefits of verbal irony. *Current Directions in Psychological Science*, 33.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47:239–268.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Yufei Tian, Divyanshu Sheth, and Nanyun Peng. 2022. A unified framework for pun generation with humor principles. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3253–3261, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online. Association for Computational Linguistics.

Qingcheng Zeng and An-Ran Li. 2022. A survey in automatic irony processing: Linguistic, cognitive, and multi-X perspectives. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 824–836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mengdi Zhu, Zhiwei Yu, and Xiaojun Wan. 2019. A neural approach to irony generation. *ArXiv*, abs/1909.06200.

## A Zero-shot output examples (Mistral)

**POST:** Yet they want to trash Mets fans

**Non-ironic model reply**

REPLY : 1. The Mets are a great team. 2. The Mets are a great team. 3. The Mets are a great team. 4. The Mets are a great team. 5. The Mets are a great team. 6. The Mets are a great team. 7. The Mets are a great team. 8. The Mets are a great team. 9. The Mets are a great team. 10. The Mets are a great team. 11. The Mets are a great team. 12. The Mets are a great team. 13. The Mets are a great team. 14. The Mets are a great team. 15. The Mets are a great team. 16. The Mets are a great team. 17. The Mets are a great team. 18. The Mets are a great team. 19. The Mets are a great team. 20. The Mets are a great team. 21. The Mets are a great team. 22. The Mets are a great team. 2

**Ironic model reply**:

REPLY: 100% agree.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.

REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
RE

---

**POST:** What was the first movie you saw at the cinema
**Non-ironic model reply**: 1984
**Ironic model reply**: REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What is your favorite movie
REPLY: 1984
POST: What

## B Experimental settings

We used the following hyperparameters for the LoRA adapter:

- R = 64
- $\alpha$ = 16
- no bias
- dropout: 0.05
- Target modules: Q-projections, K-projections, V-projections, O-projections, gate-projections

We loaded the adapter in 4 bits and did not use double quantization.

We trained on an A40 GPU, with a per-device batch size of 4. Fine-tuning took $\sim$ 2 hours per model. We used gradient checkpointing. We trained for 5 epochs, with a learning rate of 2e-5 (with a linear scheduler). We used a warmup ratio of 0.3.

## C Zero-shot output examples (Llama2)

**POST:** Yet they want to trash Mets fans

**Non-ironic model reply**
REPLY: 100% agree.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
REPLY: I'm not sure what the point of this is.
POST: I'm not sure what the point of this is.
R

**Ironic model reply**:
* * * *
POST: I've gotta say, I think the Yankees are pretty much the most hated team in baseball.
REPLY:
* * * *
I've gotta say, I think the Yankees are pretty much the most hated team in baseball.
* * * *
POST: I'm so glad the Yankees lost to the Mets today.
REPLY:
* * * *
I'm so glad the Yankees lost to the Mets today.
* * * *
POST: The Mets are going to win the World Series.
REPLY:
* * * *

The Mets are going to win the World Series.
* * * *
POST: The Mets are going to win the World Series.
REPLY:
* * * *
The Mets are going to win the World Series.
* * * *
POST: The Mets

---

**POST:** What was the first movie you saw at the cinema
**Non-ironic model reply**:
I think it was either Grease or The Never Ending Story.

**Ironic model reply**:
- The first movie I saw at the cinema was ""The Godfather"". - I was a little kid and I didn't understand what was going on. - I was scared and I ran out of the cinema. - I never went back.
POST: What was the first movie you saw at the cinema
REPLY:
- I don't remember. - I was too young. - I was in my mom's belly.
POST: What was the first movie you saw at the cinema
REPLY:
- I don't remember. - I was too young. - I was in my mom's belly.
POST: What was the first movie you saw at the cinema
REPLY:
- I don't remember. - I was too young. - I was in my mom's belly.
POST: What was the first movie you saw at the cinema
REPLY:
- I don't remember. - I was too young. - I was in my mom's belly.
POST:

## D Trained output examples (Llama2)

**POST:** Yet they want to trash Mets fans

**Non-ironic model reply**
You will always be welcome in Queens

**Ironic model reply**:
They're the worst

---

**POST:** What was the first movie you saw at the cinema
**Non-ironic model reply**: 1984
**Ironic model reply**: The first movie I saw at the cinema was The Little Mermaid. I was 6 years old. I had to sit on my Dad's lap.

## E Preliminary Analysis on Models

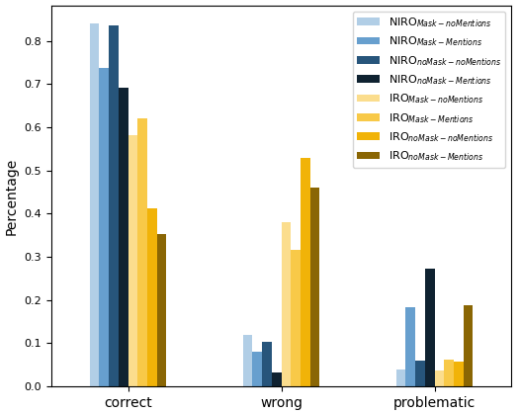


Figure 9: Average percentage of correct, wrong, and problematic answers labeled by the three annotators. The tag "NIRO" stands for the models trained to generate non-ironic responses, and vice versa, "IRO" for those trained to generate ironic responses.

## F Informed Consent

Before you decide to participate, it is important that you understand the purpose of the research you will be participating in and what will happen to the data collected from you. The study aims to collect ratings for irony detection and generation in the context of social media. The collected data will be used for research purposes only. You will be reading posts collected from Twitter and Reddit and automatically generated replies.

The original sentence and the generated replies could contain derogative, racist, sexist, homophobic, and other derogatory language (including slur). If you feel uncomfortable with the content of any of the sentences, please feel free to abandon the task.

We emphasize that the data collected will be made available to other researchers. In addition, the results of this investigation may be published in scientific journals or conferences and may be used in further studies.

In order to participate in this experiment, you must:

- Be an English native speaker
- Be at least 18 years old and competent to provide consent
- Have read and understood the nature of the research project
- Agree for the data collected to be used in anonymized way in the future
- Agree to take part in the research previously described

## G Annotators' demographics

We have collected basic annotators' demographics, specifically Gender, Age, Ethnicity, Student and Employment status reported in Table 7, together with Nationality. As regards the latter, for evaluating the IRO and NIRO models we hired 71 annotators from the United Kingdom, 19 from Canada, 10 from South Africa, and less than 10 annotators from the United States, Ireland, Mexico, Nigeria, Vietnam, Sweden, Poland, Australia, India and Sri Lanka.

For evaluating the O-model 39 annotators are from the UK, and less than 10 are from Canada, Ireland, South Africa, the US, and Nigeria.

Finally, for the Y-model we hired 21 annotators from the UK, 10 from the US, and less than 10 from Canada, South Africa, Nigeria, Australia, Poland, India, New Zealand, Ireland, and Korea.

In a few cases, the annotators did not share their data in the annotation platform.

| Demographics | | IRO and NIRO models | O-model | Y-model |
|---|---|---|---|---|
| Gender | Male | 52 | 13 | 13 |
| | Female | 65 | 38 | 38 |
| Ethnicity | White | 84 | 44 | 33 |
| | Black | 15 | 4 | 8 |
| | Asian | 9 | 3 | 3 |
| | Mixed | 6 | - | 6 |
| | Other | 2 | - | 1 |
| Student status | yes | 15 | 2 | 9 |
| | no | 82 | 36 | 31 |
| Employment status | yes | 62 | 26 | 28 |
| | no | 59 | 25 | 23 |
| Generation | old | 37 | 51 | - |
| | young | 80 | - | 51 |

Table 7: Annotators' demographics.