Towards a Perspectivist Understanding of Irony through Rhetorical Figures

Pier Felice Balestrucci[1]*, Michael Oliverio[1]*, Elisa Chierchiello[1], Eliana Di Palma[1], Luca Anselma[1], Valerio Basile[1], Cristina Bosco[1], Alessandro Mazzei[1], Viviana Patti[1],
[1]Computer Science Department, University of Turin, Italy,
Correspondence: pierfelice.balestrucci@unito.it

Abstract

Irony is a subjective and pragmatically complex phenomenon, often conveyed through rhetorical figures and interpreted differently across individuals. In this study, we adopt a perspectivist approach, accounting for the socio-demographic background of annotators, to investigate whether specific rhetorical strategies promote a shared perception of irony within demographic groups, and whether Large Language Models (LLMs) reflect specific perspectives. Focusing on the Italian subset of the perspectivist MultiPICo dataset, we manually annotate rhetorical figures in ironic replies using a linguistically grounded taxonomy. The annotation is carried out by expert annotators balanced by generation and gender, enabling us to analyze inter-group agreement and polarization. Our results show that some rhetorical figures lead to higher levels of agreement, suggesting that certain rhetorical strategies are more effective in promoting a shared perception of irony. We fine-tune multilingual LLMs for rhetorical figure classification, and evaluate whether their outputs align with different demographic perspectives. Results reveal that models show varying degrees of alignment with specific groups, reflecting potential perspectivist behavior in model predictions. These findings highlight the role of rhetorical figures in structuring irony perception and underscore the importance of socio-demographics in both annotation and model evaluation.

1 Introduction

Irony is a complex communicative phenomenon in which the intended meaning diverges from the literal interpretation of an utterance (Muecke, 1970). It often relies on pragmatic inference and contextual cues, making it a challenging target for computational modeling. Beyond its linguistic complexity, irony is also deeply subjective: its perception varies across individuals and is shaped by socio-demographic traits such as age, gender, or cultural background (Frenda et al., 2023a).

Linguistic studies distinguish several categories of irony, including hyperbole, exaggeration, and changes in register, conveyed through rhetorical figures (Karoui et al., 2017). These rhetorical figures can be seen as markers of different categories of irony, each relying on distinct communicative cues (Athanasiadou and Colston, 2020; Kühn and Mitrović, 2024). Recognizing such strategies may therefore aid in detecting irony and understanding how it is perceived across individuals.

At the same time, the subjectivity inherent in irony interpretation poses a challenge: what one person may find clearly ironic, another may interpret literally or fail to recognize altogether. This perspectivist dimension (Frenda et al., 2024) highlights the subjective variability in irony perception, posing challenges for both annotation and computational modeling.

In this paper, we study irony not as a uniform phenomenon, but as a set of rhetorical categories that shape its interpretation. Specifically, we investigate whether certain rhetorical figures promote a shared perception of irony categories among individuals who share socio-demographic traits—and whether such alignment can also be observed in the behavior of Large Language Models (LLMs).

Indeed, LLMs have emerged as powerful tools for natural language understanding and generation. Their ability to capture subtle patterns in language makes them promising candidates for modeling complex pragmatic phenomena such as irony (Balestrucci et al., 2024). Yet, LLMs are not neutral observers: their outputs reflect the data they were trained on, which may embed implicit cultural backgrounds, social perspectives, or biases (Kotek et al., 2023). When applied to subjective phenomena like irony, this raises the question of whether LLMs themselves adopt specific perspec-

*Equal contribution.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 27–36

tives in how they interpret rhetorical and ironic content (Basile et al., 2024).

To this end, in the first part of the paper, we focus on the Italian subset of the perspectivist MultiPICo dataset (Casola et al., 2024), which contains short social media conversations annotated for irony (*ironic* versus *not-ironic*) by a diverse pool of annotators. So, we augment the MultiPICo annotation by manually annotating the rhetorical figures, adopting the taxonomy proposed by Karoui et al. (2017), into the replies that were labeled as ironic by majority vote in the original campaign. This process is carried out by annotators grouped by generation and gender, allowing us to examine patterns of agreement both within and across demographic groups.

In the second part of the study, we first train LLMs to automatically classify rhetorical figures in ironic replies. In order to improve classification performance, we fine-tune the models on TWITTIRÒ-UD (Cignarella et al., 2017), a corpus of ironic Italian tweets annotated with rhetorical figures. We then examine whether the predictions made by the models reflect the annotation patterns of particular demographic groups—thus highlighting potential perspectivist biases in how LLMs handle complex pragmatic phenomena like irony.

Our study is guided by the following research questions (RQs):

- **RQ1:** Do rhetorical figures promote a shared perception of irony categories across different demographic groups?
- **RQ2:** Do LLMs exhibit perspectivist behavior when classifying rhetorical figures in ironic texts?

The remainder of the paper is structured as follows. Section 2 reviews the literature on irony, rhetorical figures, and perspectivist annotation. Section 3 introduces the MultiPICo dataset. Section 4 outlines our experimental design, followed by the manual annotation campaign and result analysis in Section 5. In Section 6, we present the automatic classification experiments with LLMs. We conclude with a summary of findings in Section 7 and a discussion of limitations in Section 8.[1]

[1]All code and the manually annotated corpus used in this study are available at: https://github.com/MichaelOliverio/perspectivist-understanding-rhetorical-figures.

2 Related Works

Recent work in NLP has increasingly emphasized the importance of taking annotators' perspectives into account when dealing with subjective linguistic phenomena such as irony or hate speech. Instead of treating disagreement as a flaw to be minimized, the *perspectivist approach* (Basile et al., 2021; Frenda et al., 2025) considers it meaningful variation that reflects different ways of interpreting language. To support this view, several studies have proposed modeling annotations at the level of individuals (Davani et al., 2022) or groups defined by shared beliefs or demographic traits (Frenda et al., 2023b; Akhtar et al., 2019).

This line of research relies on disaggregated datasets, where annotations are linked to metadata such as age, gender, ideology, or cultural background (Cabitza et al., 2023; Sachdeva et al., 2022). These datasets allow researchers to investigate how socio-demographic traits influence linguistic judgments, and to build models that better capture the diversity of interpretations (Sap et al., 2021; Wan et al., 2023). Incorporating this information has been shown to improve not only fairness, but also classification performance.

In the domain of irony detection, several studies have started to explore the relationship between perspectivism and the perception of irony (Frenda et al., 2023a,b), revealing, for instance, that irony can be more polarizing depending on the annotators' generation (Casola et al., 2024). In line with this direction, the present work aims to further investigate the perspectivist nature of irony by considering it as a phenomenon that can be classified into rhetorical categories (Karoui et al., 2017). Specifically, we propose a study that seeks to explain and analyze the role of annotators' perspectives in the perception and classification of irony through rhetorical figures.

3 MultiPICo

MultiPICo (Casola et al., 2024) is a multilingual dataset of short social media conversations, each consisting of a post and its reply, annotated to indicate whether the reply is ironic in response to the post. It contains a total of 18,778 post–reply pairs collected from Reddit (8,956) and Twitter (9,822), spanning nine languages. The annotations were obtained through crowdsourcing from 506 individuals with diverse demographic profiles, resulting in 94,342 labels—an average of 5.02 annotations per

post–reply pair. Each label is enriched with demographic metadata, including gender, age, ethnicity, student status, and employment.

In the Italian subset, 24 annotators provided 4,790 labels across 1,000 conversations.[2] Among them, 11 were female and 13 male. With respect to age groups, 11 annotators belonged to Gen Z (born between 1997 and 2012), 12 to Gen Y or Millennials (born between 1981 and 1996), and 1 to Gen X (born between 1965 and 1980).

4 Methodology

The first step of our methodology consists in the manual annotation of the Italian subset of MultiPICo by linguistically trained experts with specific knowledge of rhetorical figures. We adopt the taxonomy proposed by Karoui et al. (2017), which classifies irony into eight categories. Seven of these are grounded in rhetorical structures, while the eighth—OTHER —serves as an umbrella category encompassing situational irony and humor (Shelley, 2001; Niogret, 2004).

The seven rhetorical categories are as follows:

- ANALOGY (Ritchie, 2005; Burgers, 2010): involves similarity between two things that have different ontological concepts or domains, on which a comparison may be based.
- HYPERBOLE (Berntsen and Kennedy, 1996; Mercier-Leca, 2003; Didio, 2007): makes a strong impression or emphasizes a point.
- EUPHEMISM (Muecke, 1978; Seto, 1998): reduces the facts of an expression or an idea considered unpleasant in order to soften the reality.
- RHETORICAL QUESTION (Barbe, 1995; Berntsen and Kennedy, 1996): asks a question in order to make a point rather than to elicit an answer.
- CONTEXT SHIFT (Haiman, 1998; Leech, 2016): a sudden change of topic or frame; use of exaggerated politeness in a situation where it is inappropriate, etc.
- FALSE ASSERTION (Didio, 2007): a proposition, fact, or assertion that fails to make sense against reality.
- OXYMORON/PARADOX (Gibbs, 1994; Barbe, 1995; Tayot, 1984): equivalent to "False assertion" except that the contradiction is explicit.

All annotators belong to the same demographic groups considered in the original MultiPICo annotation campaign. For this study, we focus on two dimensions: gender and generation. A subset of 200 ironic Italian post–reply pairs was annotated by six individuals—three male and three female—balanced across generations: two from Gen X, two from Gen Y, and two from Gen Z.

We then analyze whether these groups show consistent patterns in the identification of rhetorical figures for ironic texts, both within and across demographic groups, in order to address our first research question.

In the second phase of the study, we fine-tune various LLMs on rhetorical figure classification. We then evaluate their capability to classify rhetorical figures in ironic post–reply pairs from MultiPICo. Finally, we investigate whether these LLMs exhibit specific perspectives in their classification outputs, analyzing potential alignment with human demographic groups.

5 MultiPICo Annotation

In this section, we describe the annotation of the Italian subset of MultiPICo using the taxonomy proposed by Karoui et al. (2017), which was specifically developed for the analysis of ironic texts. We focus exclusively on post–reply pairs annotated as ironic in MultiPICo, selected through a majority vote strategy. This yields a total of 278 ironic post–reply pairs.

The annotation was performed by six volunteer native Italian speakers, all with a strong academic background in linguistics, on 200 out of the 278 ironic post–reply pairs.

The annotation process follows these steps:

- We adopt the annotation guidelines released by Karoui et al. (2017) to ensure consistency with their framework.[3]
- We label the reply, using the post as contextual information to support the classification of rhetorical figures;
- We assign one or more labels to each reply, depending on the rhetorical figures identified.

[2] https://huggingface.co/datasets/Multilingual-Perspectivist-NLU/MultiPICo

[3] Guidelines available at: https://github.com/Jihen-Karoui/Scheme

Annotator Agreement across Rhetorical Figures Once the annotation phase was completed, we analyzed the level of agreement among annotators to understand whether certain rhetorical figures promote a more shared perception of irony.

Our hypothesis is that, if some rhetorical figures are more easily or intuitively recognized as markers of irony, they should yield higher agreement scores across annotators. To test this, we computed inter-annotator agreement for each figure using both Fleiss' κ (Fleiss, 1971) and Krippendorff's α (Krippendorff, 2011), as shown in Table 1.

The results reveal notable differences across labels: RHETORICAL QUESTION achieves the highest agreement ($\kappa = 0.426$, $\alpha = 0.426$), followed by HYPERBOLE and ANALOGY. This may be due to the fact that these figures often exhibit salient syntactic or lexical markers in Italian—such as the use of a question mark in rhetorical questions, or comparative structures introduced by *come* ("like/as") in analogies—making them more easily recognizable and less open to interpretive ambiguity. Other figures—such as EUPHEMISM, OXYMORON, and CONTEXT SHIFT—show much lower agreement scores.

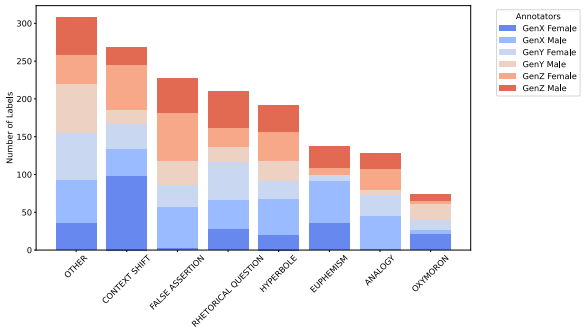

Figure 1: Distribution of Rhetorical Figures Annotated per Annotator

An analysis of label distribution (Figure 1) shows that the most frequent categories are OTHER and CONTEXT SHIFT, further confirming that annotator agreement is driven more by the presence of recognizable linguistic cues than by the predominance of any single category within the annotated sample.

To illustrate how certain rhetorical figures may be more easily and consistently identified, we report two representative examples from our dataset:

- **Post:** "@USER Not exactly good morning." *("@USER Non troppo buongiorno.")*
- **Reply:** "@USER Grandpa! Already awake???" *("@USER Nonnino! Già sveglio???")*

Five annotators labeled the reply as a *rhetorical question*. The ironic tone emerges from the contrast between the reply's exaggerated cheerfulness and the original negative tone. The question is not meant to be answered, but rather functions as a rhetorical device to underscore the mismatch in mood, making irony both recognizable and effective.

- **Post:** "If you find university easier than high school, I would seriously question your degree program. After all, that's how it should work—you grow, you mature, and gradually you deal with more difficult topics. But the truth is, many universities are just daycare 2.0 for people in their twenties." *("Se trovate più facile l'università che il liceo mi farei serie domande sulla vostra facoltà. D'altronde dovrebbe essere l'ordine naturale delle cose, si cresce, si matura e pian piano si affrontano argomenti più difficili. La verità è però che tante università non sono altro che un asilo 2.0 per ventenni.")*
- **Reply:** "Of course, everyone knows that in every RPG, the final boss is always the hardest one—especially if it's the biggest in the game." *("Del resto lo sanno tutti che in ogni GDR il boss più difficile in assoluto è quello finale, soprattutto se è il più grosso del gioco.")*

Also in this case, five out of six annotators labeled the reply as an *analogy*. The ironic intent is conveyed through a comparison between university education and video game dynamics, suggesting that an academic path should progressively become more challenging—just like in a role-playing game. The analogy is built around a clearly structured evaluative comparison, making the rhetorical figure relatively unambiguous and contributing to the high level of agreement among annotators.

While these examples show that some rhetorical figures can be consistently identified by different annotators, the overall picture remains more nuanced. The average agreement across all figures is modest ($\kappa = 0.198$, $\alpha = 0.199$), suggesting that only some rhetorical strategies promote a shared perception of irony categories.

Crucially, all annotators involved are trained linguists with expertise in rhetorical analysis, and

were provided with detailed annotation guidelines. One might therefore expect a high level of objectivity and consistency. However, the observed variation indicates that the classification of rhetorical figures in ironic texts is not a straightforward or universally shared process, but rather a task that involves subjective interpretation—even among experts.

Label	Fleiss' κ	Krippendorff's α
ANALOGY	0.238	0.238
CONTEXT SHIFT	0.112	0.112
EUPHEMISM	0.089	0.090
FALSE ASSERTION	0.194	0.194
HYPERBOLE	0.304	0.304
OTHER	0.142	0.143
OXYMORON	0.084	0.085
RHETORICAL QUESTION	0.426	0.426
Average	0.198	0.199

Table 1: Inter-annotator agreement scores (Fleiss' κ and Krippendorff's α) for each rhetorical figure.

Annotators' Polarization Following the analysis proposed by Casola et al. (2024), we used the Polarization Index (P-index) introduced by Akhtar et al. (2019). This measure evaluates, for each instance—in our case, each post–reply pair—the polarization in annotations provided by annotators grouped according to specific sociodemographic characteristics. An example of such grouping, shown in Table 2, is by gender (male/female) or by generation (Gen X/Y/Z).

The P-index ranges between 0 and 1, where 0 indicates complete agreement across different groups (no polarization), and 1 indicates maximum internal agreement within each group but total disagreement between groups (maximum polarization).

Formally, the P-index for an instance i is defined as:

$$P(i) = \frac{1}{k} \sum_{w=1}^{k} a(G_w) \cdot (1 - a(G)) \quad (1)$$

where k is the number of groups (for example, 3 in the case of grouping by generation), $a(G_w)$ is the internal agreement level within group G_w for instance i, and $a(G)$ is the overall agreement level of all annotators on instance i. Following the original proposal, the agreement (a) is calculated using a normalized χ^2 statistic:

$$a(G) = \frac{\chi^2(G)}{|M|} \quad (2)$$

where $\chi^2(G)$ denotes the chi-square statistic for group G, and $|M|$ is the number of annotations for the corresponding instance.

We employed the P-index for groups defined by gender and generation. Due to the multi-label nature of our annotation scheme, where annotators can assign multiple rhetorical figures to a single instance, we compute the P-index independently for each rhetorical figure and report the average across all figures. An example of the P-index on an instance can be seen in Table 3.

To establish a baseline, we calculated the P-index for each rhetorical figure over all possible random combinations of annotators—pairs for gender grouping and triplets for generation grouping—and averaged the results accordingly.

Additionally, we also calculated the percentage difference (%Δ) between the real P-index and the random P-index, to highlight the degree of polarization actually observed compared to a random baseline.

Gender			**Generation**		
real	random	%Δ	real	random	%Δ
0.124	0.132	−6.10	0.191	0.146	31.13

Table 2: Polarization index values calculated for annotator groups based on gender and generation. The table shows the real P-index, the random P-index obtained by averaging over random permutations of annotators, and the relative percentage difference (%Δ) between the real and random values.

The results in Table 2 show that for the gender dimension, the real P-index value (0.124) is lower than the one expected by chance (0.132), with a negative percentage difference of −6.10%. This suggests that annotators do not tend to polarize based on gender; in fact, their annotations appear to be slightly less variable within gender groups than would be expected randomly. In contrast, for the generation dimension, the real P-index value (0.191) is higher than the random baseline (0.146), with a positive difference of 31.13%. This indicates that generation is a polarizing trait in the annotation of rhetorical figures. In other words, annotators within the same age group tend to agree more with each other, while differing more from those in other generational groups.

Post	Reply	Ann. Gen.	An	Cs	Eu	Fa	Hy	Ot	Ox	Rq	P-index
@USER It will be the first strong team they face.... (*@USER Sarà la prima squadra forte che affrontano....*	@USER Which one of the two? ? (*@USER Quale delle due? ?*)	X	0	1	0	0	0	0	0	1	0.100
		X	0	1	0	0	0	0	0	0	
		Y	0	0	0	0	0	0	0	1	
		Y	0	0	0	0	0	0	0	1	
		Z	0	0	0	0	0	0	0	1	
		Z	0	0	0	0	0	0	0	1	

Table 3: Example of polarization in the annotations. While the reply *"Which one of the two?"* may appear as a rhetorical question, the table reveals disagreement among annotators from different generations. All Gen Z and Gen Y annotators labeled it as a *Rhetorical question* (Rq), whereas only one Gen X annotator agreed, with another opting for *Context shift* (Cs). Abbreviations: Ann. Gen. = Annotator Generation, An = Analogy, Cs = Context shift, Eu = Euphemism, Fa = False assertion, Hy = Hyperbole, Ot = Other, Ox = Oxymoron, Rq = Rhetorical question.

6 Rhetorical Figure Classification and Perspective Alignment in LLMs

In this section, we explore whether LLMs reflect specific perspectives when classifying rhetorical figures in ironic texts. As a first step, we fine-tuned a set of multilingual LLMs on the TWITTIRÒ-UD dataset, aiming to enhance their performance in the classification of rhetorical figures within ironic language. Indeed, while the TWITTIRÒ-UD dataset serves to fine-tune and evaluate the LLMs' classification abilities, the MultiPICo data instead allow us to assess whether model predictions align more closely with specific demographic perspectives.

TWITTIRÒ-UD TWITTIRÒ-UD is a corpus of ironic Italian tweets annotated with rhetorical figures and linguistic information following the Universal Dependencies (UD) framework.[4] It contains 1,424 tweets and over 28,000 tokens, originally collected for the fine-grained annotation of irony. Each tweet is labeled with the rhetorical figure used to convey irony, based on the taxonomy proposed by Karoui et al. (2017).

Model Setup and Fine-Tuning We fine-tuned four LLMs on TWITTIRÒ-UD using a reasoning instruction format, in which the model is prompted to first generate a short explanation before producing the final label, following the Chain-of-Thought prompting strategy (Wei et al., 2022). The models we used are:

- Llama-3.1-8B-Instruct[5],
- Ministral-8B-Instruct-2410[6],
- LLaMAntino-3-ANITA-8B-Inst-DPO-ITA[7],
- Minerva-7B-instruct-v1.0[8].

Model fine-tuning Fine-tuning was performed using the Low-Rank Adaptation (LoRA) method (Hu et al., 2021). All models were prompted in English and trained to output both the explanation and the final rhetorical figures using the labels from the original annotation schema. The training was conducted using the `transformers` and `peft` libraries. Table 4 summarizes the main parameters used in the `TrainingArguments` class and in the LoRA configuration.

Parameter	Value
LoRA configuration	
LoRA rank (r)	64
LoRA alpha	16
Dropout probability	0.1
TrainingArguments	
Number of training epochs	5
Enable fp16 training	`False`
Enable bf16 training	`True`
Batch size per GPU for training	1
Batch size per GPU for evaluation	1
Gradient accumulation steps	1
Maximum gradient norm	0.3
Initial learning rate	2e−4
Weight decay	0.001
Optimizer	`adamw_torch`
Learning rate schedule	`cosine`
Warmup ratio	0.03

Table 4: Configuration of hyperparameters used in the LoRA-based fine-tuning process.

The input prompt for the fine-tuning followed this format:

[4] https://github.com/UniversalDependencies/UD_Italian-TWITTIRO

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[6] https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

[7] https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA

[8] https://huggingface.co/sapienzanlp/Minerva-7B-instruct-v1.0

Instruction: Given the ironic sentence (INPUT), identify and return the rhetorical figure it exemplifies in (OUTPUT). Explain your reasoning first, and then answer with the rhetorical figure.

Baselines To contextualize the performance of the fine-tuned models, we defined two baselines:

- **Random**: a naive classifier that assigns one of the eight possible rhetorical categories uniformly at random. This provides a sense of the task's inherent difficulty.
- **Zero-Shot prompting**: we prompted the best-performing model in its non–fine-tuned version using the same instruction and listing all rhetorical categories as candidate outputs. This baseline allows us to estimate how much LLMs know about rhetorical devices without fine-tuning.

Model	Precision	Recall	F1-score
Llama-3.1-8B	0.378	0.406	0.384
LLaMAntino-3-8B	0.382	0.397	0.385
Ministral-8B	**0.393**	**0.408**	**0.396**
Minerva-7B	0.367	0.385	0.372
Random	0.138	0.122	0.125
Zero-Shot	0.213	0.218	0.185

Table 5: Performance of fine-tuned models on the TWITTIRÒ-UD test set. Scores are reported as weighted averages of precision, recall, and F1-score across three runs.

Results on TWITTIRÒ-UD Table 5 reports the classification results on the TWITTIRÒ-UD test split. Each LLM was run three times per input using a temperature of 0.1. We report the results as the weighted average of Precision, Recall, and F1-score, in order to account for the different distribution of the rhetorical figures in the dataset.

The random baseline acts as a benchmark to evaluate the inherent difficulty of the task: given the presence of eight possible classes, it is very unlikely to achieve strong results through chance alone. Within this challenging setup, Ministral-8B achieves the highest performance, narrowly surpassing other fine-tuned models. Moreover, the zero-shot results obtained by prompting Ministral-8B reveal that LLMs possess some prior understanding of rhetorical figures and their use, as evidenced by their performance exceeding random chance. Finally, fine-tuning on the TWITTIRÒ-UD dataset leads to a substantial improvement in their classification performance.

Do LLMs Exhibit a Specific Perspective? To explore whether LLMs adopt a specific perspective when classifying rhetorical figures, we assessed their performance against gold references derived from different demographic groups. Specifically, for each group in the Italian subset of MultiPICo (Female, Male, Gen X, Gen Y, Gen Z), we computed the most frequently assigned rhetorical figure label across all instances, based on the annotations provided by human annotators belonging to that group in Section 5. These labels were then used as gold references to calculate precision, recall, and F1-scores for each model. We also computed an additional "Global" reference, using the most frequent label aggregated across all annotators, regardless of group.

Table 6 reports model performance under these different evaluation perspectives. The results show consistent variation depending on which group's labels are used as gold. For instance, Llama-3.1-8B performs notably better when evaluated against the Gen X labels (F1 = 0.215), suggesting a closer alignment with the rhetorical preferences of Gen X annotators. Minerva-7B shows a similar trend, also achieving its highest F1-score (0.260) with Gen X. In contrast, LLaMAntino-3-8B performs best when evaluated against the labels assigned by the Gen Z group (F1 = 0.241), while Ministral-8B performs best with the Female group (F1 = 0.261)

These findings suggest that LLMs may align more closely with certain annotation patterns, reflecting differences in how rhetorical figures are interpreted across demographic groups.

Error Analysis To better understand the classifications produced by the models, we conducted an analysis of the most frequent errors.

One of the most common issues involves the distinction between the post and the reply. In many cases, the models tend to assign the label to the post rather than the reply, which is actually the correct target for classification. For example, in the following pair:

- **Post**: "Do you think a MORTADELLA SANDWICH could be considered HOMEOPATHIC?" (*"Secondo voi il PANINO CON LA MORTADELLA si può considerare OMEOPATICO?"*)

Model	Group	Precision	Recall	F1-Score
Llama-3.1-8B	Female	0.236	0.214	0.204
	Male	0.220	0.199	0.177
	Gen X	**0.305**	0.199	**0.215**
	Gen Y	0.187	0.194	0.160
	Gen Z	0.161	0.159	0.138
	Global	0.217	**0.219**	0.195
LLaMAntino-3-8B	**Female**	**0.333**	0.174	0.187
	Male	0.271	0.189	0.202
	Gen X	0.251	0.179	0.193
	Gen Y	0.267	0.199	0.204
	Gen Z	0.275	**0.249**	**0.241**
	Global	0.311	0.204	0.224
Ministral-8B	**Female**	0.327	**0.244**	**0.261**
	Male	0.254	0.199	0.200
	Gen X	0.258	0.184	0.202
	Gen Y	0.275	0.224	0.218
	Gen Z	0.193	0.189	0.182
	Global	**0.346**	0.239	0.250
Minerva-7B	Female	0.305	0.214	0.220
	Male	0.327	0.184	0.183
	Gen X	**0.367**	**0.234**	**0.260**
	Gen Y	0.296	0.184	0.166
	Gen Z	0.181	0.184	0.167
	Global	0.314	0.209	0.202

Table 6: Performance of each model on the Italian subset of MultiPICo, reported as weighted averages of precision, recall, and F1-score. Gold labels correspond to the most frequent label assigned by human annotators for each demographic group (Female, Male, Gen X, Gen Y, Gen Z) and overall (Global).

- **Reply**: "@USER Yes" (*"@USER Si"*)

LLaMAntino-3-8B assigns the label RHETORICAL QUESTION, which is more appropriate for the post than for the reply. In this case, most human annotators labeled the reply as FALSE ASSERTION, a rhetorical figure that better reflects the content of the response.

Another critical issue is the presence of hallucinations in the models' responses. For instance:

- **Post**: "@USER No no, it's right, it has to be there, you feed it, cuddle it, keep it warm, it has to be there" (*"@USER No no è giusto, ce deve sta, la nutri la coccoli la tieni calda, ce deve sta"*)
- **Reply**: "@USER Actually, the other one handles it. I'm just a disruptive element." (*"@USER Veramente ce pensa quell'altro. Io sono un mero elemento di disturbo."*)

In this case, Llama-3.1-8B labels the reply as SITUATIONAL IRONY, which is not part of the label set used during fine-tuning. The appropriate label would be OTHER, which was in fact the most frequently assigned category by annotators in similar situations.

This analysis highlights the need for improvements in the fine-tuning phase of the models, particularly to ensure clarity that the classification should refer exclusively to the reply, with the post serving only as contextual information. Additionally, it is important to reinforce the alignment between the available labels and those used by the model, in order to avoid generating labels not included in the adopted taxonomy.

7 Conclusions

In this paper we investigated irony as a multifaceted phenomenon, structured by different rhetorical figures that guide its interpretation. By focusing on the Italian subset of the perspectivist MultiPICo dataset, we conducted a manual annotation campaign in which expert annotators labeled rhetorical figures in ironic replies. The annotators were balanced across gender and generation, allowing us to explore patterns of agreement both within and across demographic groups.

Our findings show that only some rhetorical figures—such as RHETORICAL QUESTION, HYPERBOLE, and ANALOGY—promote a more shared perception of irony categories. Others yielded lower agreement, highlighting the subjective nature of this task. Despite the linguistic expertise of the annotators and the use of detailed guidelines, the overall agreement remained modest, supporting the perspectivist view that irony interpretation is influenced by socio-demographic background.

We then trained and evaluated LLMs on rhetorical figure classification. While fine-tuned models outperformed baselines, their predictions showed variation depending on which group's annotations were used as gold labels. In particular, different models aligned more closely with different demographic perspectives—suggesting that LLMs may replicate specific patterns observed in human annotation.

These results emphasize the importance of incorporating socio-demographic information when modeling complex pragmatic phenomena such as irony, both to improve classification performance and to better account for variation in human interpretation.

8 Limitations

This study presents a first attempt to investigate the perspectivist nature of irony through the lens of rhetorical figures. However, it presents some limitations that open directions for future work.

First, our analysis is limited to the Italian subset of the MultiPICo dataset. While this choice enabled a controlled and linguistically grounded study, future work will extend the approach to other languages and cultural contexts, to assess whether similar perspectivist patterns emerge cross-linguistically.

Second, the annotation was carried out by a small group of six annotators. This limited sample size may restrict the generalizability of our findings. Nonetheless, we opted for a small but expert group of annotators—all with a background in linguistics—to ensure a high-quality annotation of complex rhetorical phenomena. Relying on larger but less specialized crowdsourcing platforms could have introduced noise and inconsistencies, particularly in the classification of fine-grained rhetorical strategies.

Third, to improve model performance in the automatic classification task, we fine-tuned the LLMs on the TWITTIRÒ-UD dataset. While this resource provides valuable rhetorical annotations for ironic content, its use may introduce a potential source of bias, as the labels reflect the interpretative choices of a different group of annotators.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI* IA 2019–Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18*, pages 588–603. Springer.

Angeliki Athanasiadou and Herbert L Colston. 2020. *The Diversity of Irony*, volume 65. Walter de Gruyter GmbH & Co KG.

Pier Felice Balestrucci, Silvia Casola, SODA Lo, Valerio Basile, Alessandro Mazzei, and 1 others. 2024. I'm sure you're a real scholar yourself: Exploring ironic content generation by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14480–14494. Association for Computational Linguistics.

Katharina Barbe. 1995. Irony in context.

Valerio Basile, Silvia Casola, Simona Frenda, and Soda Marem Lo. 2024. PERSEID - perspectivist irony detection: A CALAMITA challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1074–1081, Pisa, Italy. CEUR Workshop Proceedings.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, and 1 others. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Dorthe Berntsen and John M Kennedy. 1996. Unresolved contradictions specifying attitudes—in metaphor, irony, understatement and tautology. *Poetics*, 24(1):13–29.

Christian Frederik Burgers. 2010. *Verbal irony: Use and effects in written discourse*. sl: sn.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

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.

Alessandra Cignarella, Cristina Bosco, and Viviana Patti. 2017. *TWITTIRÒ: a Social Media Corpus with a Multi-layered Annotation for Irony*, pages 101–106.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Lucie Didio. 2007. *Une approche sémantico-sémiotique de l'ironie*. Ph.D. thesis, Limoges.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Lang. Resour. Evaluation*, 59(2):1719–1746.

Simona Frenda, SODA Lo, Silvia Casola, Bianca Scarlini, Cristina Marco, Valerio Basile, Davide Bernardi, and 1 others. 2023a. Does anyone see the irony here? analysis of perspective-aware model predictions in irony detection. In *CEUR WORKSHOP PROCEEDINGS*, volume 3494, pages 1–11. CEUR-WS.

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. 2023b. 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.

Raymond W Gibbs. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.

John Haiman. 1998. *Talk is cheap: Sarcasm, alienation, and the evolution of language*. Oxford University Press.

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.

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.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Ramona Kühn and Jelena Mitrović. 2024. The elephant in the room: Ten challenges of computational detection of rhetorical figures. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 45–52, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Geoffrey N Leech. 2016. *Principles of pragmatics*. Routledge.

Florence Mercier-Leca. 2003. *L'ironie*. Hachette Éducation.

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

Douglas C Muecke. 1978. Irony markers. *Poetics*, 7(4):363–375.

Philippe Niogret. 2004. Les figures de l'ironie dans" a la recherche du temps perdu".

David Ritchie. 2005. Frame-shifting in humor and irony. *Metaphor and Symbol*, 20(4):275–294.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Ken-ichi Seto. 1998. On non-echoic irony. *PRAGMATICS AND BEYOND NEW SERIES*, pages 239–256.

Cameron Shelley. 2001. The bicoherence theory of situational irony. *Cognitive Science*, 25(5):775–818.

Claudine Tayot. 1984. *L'ironie*. Ph.D. thesis.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.