When Figures Speak with Irony: Investigating the Role of Rhetorical Figures in Irony Generation with LLMs

Pier Felice Balestrucci[1,†], Michael Oliverio[1,*,†], Soda Marem Lo[1], Luca Anselma[1], Valerio Basile[1], Cristina Bosco[1], Alessandro Mazzei[1] and Viviana Patti[1]

[1]*Computer Science Department, University of Turin, Italy*

Abstract

Irony poses a persistent challenge for computational models because it depends on context, implicit meaning, and pragmatic cues. This study investigates the ability of Large Language Models (LLMs) to generate ironic content by focusing on rhetorical figures—pragmatic devices that may shape and signal ironic intent. Using two datasets, TWITTIRÒ-UD and the Italian subset of MultiPICo, we fine-tune multilingual LLMs for rhetorical figure classification and evaluate their capacity to generate ironic Italian texts. Our work addresses two main questions: (1) how accurately LLMs can classify rhetorical figures in ironic Italian texts, and (2) whether such training supports the generation of irony that reflects human-like rhetorical usage. Human evaluation shows that LLMs achieve fair agreement with annotators in rhetorical figure classification, indicating a partial but promising alignment with human judgment. By leveraging rhetorical figures as a bridge between irony detection and generation, our results suggest that such training improves the stylistic control and interpretability of LLM-generated ironic language.

Keywords

Rhetorical Figures, Irony Generation, Large Language Models

1. Introduction

Irony is a complex linguistic phenomenon that involves expressing a meaning that contrasts with the literal interpretation of an utterance [1]. As a rhetorical figure, it is activated through multiple linguistic devices and pragmatic features to subvert literal meaning. Although irony is a pervasive and deeply rooted aspect of human communication, its computational modeling remains a complex and unresolved challenge.

Large Language Models (LLMs), especially when instruction-tuned, have shown remarkable progress in understanding pragmatic phenomena [2, 3]. However, their ability to leverage pragmatic features for the detection and generation of ironic content remains largely underexplored. One promising direction for addressing this challenge is to analyze the linguistic strategies through which irony is commonly expressed. Specifically, Karoui et al. [4] defined eight categories of irony, characterized by pragmatic features used to express meaning incongruence and grounded in rhetorical figures. Following their categorization of irony, this study investigates the capacity of LLMs to analyze and generate ironic texts in Italian when rhetorical figures are taken into account as cues for ironic intent. Thus, we focus on how they contribute to the expression of irony.

Indeed, irony can be also activated through the interaction with rhetorical figures, either amplifying their intended effects, as in the case of paradox, or subverting them entirely, as occurs with hyperbole. This interplay contributes to the richness and rhetorical complexity of ironic expressions [5].

In this work, we draw on two complementary datasets: TWITTIRÒ-UD, a corpus of ironic Italian tweets annotated using the rhetorical figure annotation scheme introduced by Karoui et al. [4], and MultiPICo, a multilingual collection of social media post–reply pairs annotated for irony by annotators with diverse sociodemographic characteristics, in which each reply is annotated with a binary label indicating whether it is ironic with respect to the corresponding post. By integrating fine-tuning and reasoning-enhanced prompting, we aim to evaluate both the classification and generative capabilities of LLMs in this domain for Italian.

Our study is structured around the following research questions (RQ):

- **RQ1:** To what extent can LLMs accurately classify rhetorical figures in ironic Italian texts?
- **RQ2:** Does fine-tuning LLMs on rhetorical fig-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author.

†These authors contributed equally.

pierfelice.balestrucci@unito.it (P. F. Balestrucci); michael.oliverio@unito.it (M. Oliverio); sodamarem.lo@unito.it (S. M. Lo); luca.anselma@unito.it (L. Anselma); valerio.basile@unito.it (V. Basile); cristina.bosco@unito.it (C. Bosco); alessandro.mazzei@unito.it (A. Mazzei); viviana.patti@unito.it (V. Patti)

0009-0001-2161-2263 (P. F. Balestrucci); 0009-0007-3448-2377 (M. Oliverio); 0000-0002-5810-0093 (S. M. Lo); 0000-0003-2292-6480 (L. Anselma); 0000-0001-8110-6832 (V. Basile); 0000-0002-8857-4484 (C. Bosco); 0000-0003-3072-0108 (A. Mazzei); 0000-0001-5991-370X (V. Patti)

ure classification lead to the generation of more human-like ironic replies, in terms of rhetorical devices?

To address these questions, we fine-tune a set of multilingual open-weight LLMs on rhetorical figure classification and assess their performance. We then enrich the Italian subset of MultiPICo with automatic annotations and conduct a human evaluation to validate a small sample extracted from that corpus. Finally, we use the best-performing fine-tuned model to generate new replies to ironic posts in MultiPICo and carry out a linguistic analysis of the model-generated replies, comparing them with human-written ones.

This work contributes to (i) advancing the research into rhetorical figure classification using LLMs, by proving the effectiveness of Chain-of-Thought fine-tuning strategy; (ii) improving the interpretability of LLMs in pragmatic text generation, showing that rhetorical figure-aware models tend to create sentences stylistically more similar to human-written texts.[1]

2. Related Works

Rhetorical Figure Classification There are mainly two approaches to the automatic detection and classification of rhetorical figures in natural language: ontology-based methods and machine learning techniques [6, 7]. These approaches have shown effectiveness in supporting tasks such as sentiment analysis and intent classification [8, 9]. Several studies focus on their relationship with irony [10, 11], particularly in the context of irony detection. In this vein, Karoui et al. [4], drawing on well-established linguistic theories that explore the interplay between irony and rhetorical figures—such as oxymoron, paradox, false assertion, and analogy—propose an annotation schema for classifying these categories of irony in social media texts. Their work focuses on French, English, and Italian, highlighting the relevance of irony categories and markers for a linguistically informed approach to irony detection.

Irony Generation Irony generation remains a relatively underexplored area in Natural Language Generation. especially when compared to the growing literature on humor, puns, and sarcasm [12, 13]. Recent work has begun to model sarcasm through linguistic features such as valence reversal and contextual incongruity [14, 15], yet irony is still rarely addressed directly.

Among the more recent studies on irony generation, Balestrucci et al. [16] propose an approach that leverages LLMs to generate ironic text. The authors demonstrate that LLMs are capable of learning to produce ironic content, and explore the possibility of linking irony generation to the socio-demographic characteristics of user profiles—such as generational groups—with the goal of generating personalized ironic content tailored to different age groups.

3. Datasets

TWITTIRÒ-UD A collection of ironic Italian tweets annotated according to the Universal Dependencies framework. TWITTIRÒ-UD was created by enriching a resource originally developed for the fine-grained annotation of irony [17]. The original corpus consists of 1,424 tweets, with a total of 28,387 tokens [18]. Each tweet in the corpus has been annotated with the corresponding rhetorical figure used to convey irony, such as OXYMORON PARADOX, HYPERBOLE, or EUPHEMISM. The treebank includes both the fine-grained annotation for ironic tweets introduced in Karoui et al. [4] and the morphological and syntactic information encoded in the UD format.[2] Figure 1 shows the distribution of rhetorical figures in the corpus.

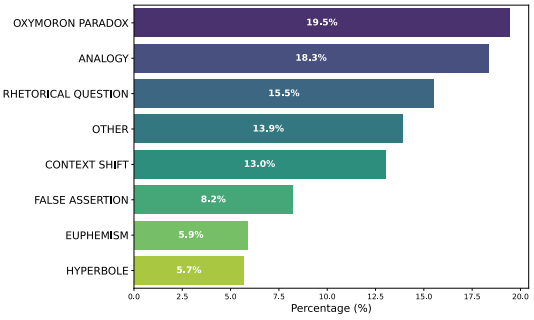

Figure 1: Distribution of rhetorical figures in the TWITTIRÒ corpus.

MultiPICo The dataset consists of disaggregated multilingual posts and replies from social media, each annotated to indicate whether the reply is ironic given the post. The corpus includes 18,778 post–reply pairs, collected from Reddit (8,956) and Twitter (9,822), and covers 9 different languages. A total of 506 annotators, with different sociodemographic information, carried out the annotations, producing 94,342 individual labels (an average of 5.02 per conversation). Each annotation is accompanied by sociodemographic metadata about the annotator, including gender, age, ethnicity, student status, and employment status. For the Italian subset of the

[1] All code and experimental results are publicly available at: https://github.com/MichaelOliverio/IronyDetection.

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

Table 1
Rhetorical figures used to convey irony. Reproduced from Karoui et al. [4].

Rhetorical Figure	Description
ANALOGY	Covers analogy, simile, and metaphor. Involves similarity between two things that have different ontological concepts or domains, on which a comparison may be based
HYPERBOLE	Make a strong impression or emphasize a point
EUPHEMISM	Reduce the facts of an expression or an idea considered unpleasant in order to soften the reality
RHETORICAL QUESTION	Ask a question in order to make a point rather than to elicit an answer
CONTEXT SHIFT	A sudden change of the topic/frame, use of exaggerated politeness in a situation where this is inappropriate, etc.
FALSE ASSERTION	A proposition, fact or an assertion fails to make sense against the reality
OXYMORON PARADOX	Equivalent to "False Assertion" except that the contradiction is explicit
OTHER	Humor or situational irony.

corpus, 24 annotators provided 4,790 annotations on 1,000 post–reply pairs [19].[3]

4. Methodology

To assess the ability of LLMs to analyze ironic Italian texts and classify rhetorical figures, we adopted the annotation scheme proposed by Karoui et al. [4], which defines a set of rhetorical figures commonly used to convey irony (summarized in Table 1).

We selected several open-weight multilingual LLMs trained on Italian data and fine-tuned them on the TWITTIRÒ dataset for the task of rhetorical figure classification. Models' performances were evaluated against two baselines: (i) a random classifier and (ii) a prompting-based approach. The best-performing model was then used to enrich the ironic Italian subset of the MultiPICo dataset—aggregated by majority vote—with rhetorical figure annotations. To validate the model's predictions, we conducted a human evaluation on a small subset of the annotated data.

Finally, to address the second research question, we focused on ironic post–reply pairs in Italian from MultiPICo, again selected via majority vote, and compared the distribution of rhetorical figures across three types of replies: (i) automatically generated by an LLM fine-tuned to recognize rhetorical figures, (ii) replies generated by the same model out-of-the-box, and (iii) written by humans. In addition to comparing the distributions, we conducted a linguistic analysis of these replies. A representative sample of the generated content was manually annotated to support this evaluation.

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

5. Rhetorical Figure Classification

In this section, we evaluate a set of LLMs for rhetorical figure classification. We fine-tune several open-weight, mid-sized LLMs using two different approaches on the original TWITTIRÒ-UD split (see Table 2). To highlight the impact of fine-tuning on rhetorical figure classification, we compare the performance of the fine-tuned models against two baselines: a random classifier and a zero-shot prompting approach. Our experiments involve five multilingual LLMs: Qwen2.5-7B-Instruct[4] (referred to as Qwen2.5-7B), Llama-3.1-8B-Instruct[5] (Llama-3.1-8B), Ministral-8B-Instruct-2410[6] (Ministral-8B), LLaMAntino-3-ANITA-8B-Inst-DPO-ITA[7] (LLaMAntino-3-8B), and Minerva-7B-instruct-v1.0 (Minerva-7B).[8]

Table 2
Data split statistics for the TWITTIRÒ-UD dataset.

	Train	Dev	Test
#Tweets	1,138	144	142
Avg. Tokens	20.77	20.80	20.96

Fine-tuning was performed using two different prompt strategies, described below, both relying on Low-Rank Adaptation (LoRA) [20].

Instruction Fine-Tuning In this approach, which we refer to as FT, we trained all the models (training details are available in Appendix A), using the following instruction:

> Given the ironic sentence (INPUT), identify and return the rhetorical figure

[4] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[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

it exemplifies in (OUTPUT).

Instruction CoT Fine-Tuning To explore an alternative, we apply a Chain-of-Thought fine-tuning strategy (referred to as CoT-FT), which guides the model to generate an explanation before predicting the rhetorical figure [3]. For example:

> **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.*
> **Input**: @user se continui sarò costretto a darti l'oscar (*@user if you keep going, I'll be forced to give you an Oscar.*)
> **Output**: The sentence draws a comparison between different domains to create irony through similarity. That's why it is an example of ANALOGY.

5.1. Model Evaluation

For the evaluation, we use the test split of TWITTIRÒ-UD. Each LLM is 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 distribution of the rhetorical figures in the dataset.

Table 3
Model performance: weighted averages of precision, recall, and F1-score across three runs per model. FT and CoT-FT indicate Fine-Tuning and Chain-of-Thought Fine-Tuning, respectively.

	Model	Precision	Recall	F1
FT	Qwen2.5-7B	0.346	0.359	0.350
	Llama-3.1-8B	0.370	0.394	0.378
	LLaMAntino-3-8B	0.373	0.399	0.379
	Ministral-8B	0.371	0.371	0.366
	Minerva-7B	0.382	0.399	0.388
CoT-FT	Qwen2.5-7B	0.350	0.352	0.349
	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
Baseline	Random	0.138	0.122	0.125
	Zero-Shot	0.213	0.218	0.185

Table 3 reports the evaluation results. The baselines used are: (i) a random classifier (Random), which assigns one of the eight possible labels uniformly at random to each input, and (ii) a zero-shot prompting approach. For the latter, we selected the best-performing model overall (Ministral-8B CoT-FT) in its non–fine-tuned version, and included the full list of rhetorical figures as candidate outputs in the prompt.

The random baseline serves as a reference point to assess the task's intrinsic difficulty: with eight possible classes, achieving high performance by chance is highly unlikely. The zero-shot results, instead, lead to two relevant observations: (i) LLMs exhibit some prior knowledge of rhetorical figures and their usage, as evidenced by their better performance compared to random guessing; and (ii) fine-tuning on the TWITTIRÒ dataset yields a considerable improvement in classification performance.

Among the fine-tuned models (FT), Italian-developed models generally outperform multilingual ones, with Minerva-7B achieving the best results in this setting, followed by LLaMAntino-3-8B.

When reasoning capabilities are introduced through Chain-of-Thought fine-tuning, performance improves consistently for most models—with the notable exception of Minerva-7B. This might be due to the fact that Minerva-7B is trained on nearly 2.5 trillion tokens—1.14 trillion of which are in Italian, which could make it less effective at generalizing reasoning when prompted in English. This behavior is evident in the outputs, where it often mixes Italian and English, producing labels such as EUFEMISMO instead of EUPHEMISM.

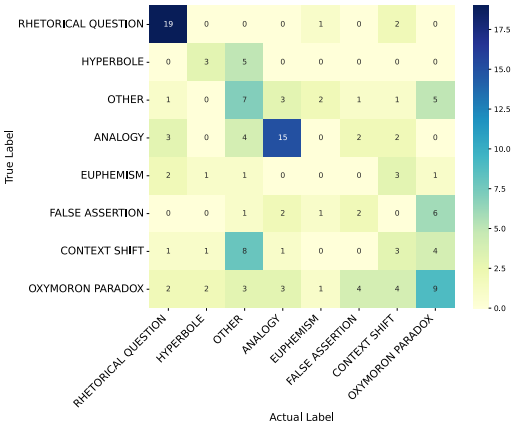

Figure 2: Confusion matrix from the third generation run of Ministral-8B with CoT-FT.

Figure 2 shows the confusion matrix for the third run of the best-performing model, Ministral-8B with CoT-FT. We observe that some rhetorical figures are easier for the model to recognize than others. In particular, the model performs well on RHETORICAL QUESTION (19 out of 22 correctly predicted) and ANALOGY (15 out of 26), which are among the most represented figures in the TWITTIRÒ dataset.

In contrast, the model struggles with several other cat-

egories—especially EUPHEMISM, for which it made no correct predictions (0 out of 8). These results highlight a substantial margin for improvement in this task and suggest the need for further investigation into the model's behavior and the characteristics of under-represented or more challenging rhetorical categories.

6. MultiPICo Enrichment

This section focuses on enriching the Italian MultiPICo with annotations of rhetorical figures. To this end, we employ the best-performing rhetorical figure classification model (see Table 3), Ministral-8B with CoT-FT, to classify rhetorical figures in the Italian post-reply pairs. As mentioned in Section 3, MultiPICo consists of both ironic and non-ironic post-reply pairs. Therefore, we extract only the ironic pairs from the dataset, using a majority vote approach to determine whether a post-reply pair is ironic, given the disaggregated nature of MultiPICo, resulting in a subset of 278 ironic post-reply pairs.

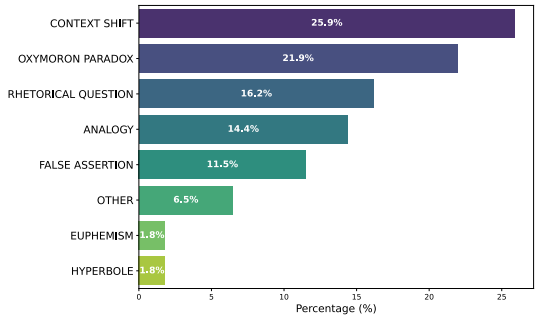

Figure 3: Distribution of rhetorical figures extracted from the Italian MultiPICo corpus.

We then use our model to classify the rhetorical figures in this subset. As shown in Figure 3, the most frequently extracted rhetorical figures in the post–reply pairs are CONTEXT SHIFT (25.9%) and OXYMORON PARADOX (21.9%), while the least frequent are EUPHEMISM and HYPERBOLE (1.8% each). This distribution closely resembles that of TWITTIRÒ, and the high frequency of CONTEXT SHIFT may be attributed to the nature of post–reply interactions, where replies often reframe or shift the meaning of the corresponding posts. Given the difficulty in classifying some rhetorical figures, as highlighted in Table 2, we carry out a human evaluation in Section 6.1 to assess the quality of the model predictions.

6.1. Human Evaluation

Following the annotation guidelines in Karoui et al. [4], two authors of this paper—both expert in computational linguistics—manually annotated a subset of 20 out of the 278 ironic post-reply pairs. The annotators were tasked to specify the rhetorical figures used to express irony in the reply given the corresponding post, selecting one or more labels from those reported in Table 1.

The annotators achieved an average Cohen's κ score [21] of 0.63 on a subset of 20 post–reply pairs, a value comparable to that reported by Karoui et al. [4] for the same task (0.60), indicating substantial agreement. Krippendorff's α [22] was also computed, yielding a score of 0.60, which confirms a similarly substantial level of inter-annotator reliability.

We then compared the human annotations with the predictions produced by our automatic model. The resulting Krippendorff's α was 0.21, corresponding to a *fair* level of agreement.

To better understand this result, we examined the 14 out of 20 pairs where both annotators assigned the same label. In 3 of these cases, the model's prediction matched the human annotation exactly.

For example, for the post: *Due si candidano in quanto "ci vuole una donna" nel #Pd: #Schlein e #DeMicheli. Una sola domanda: perché?"* (Two women are running for office in the Democratic Party because 'we need a woman': Schlein and DeMicheli. One question: why?") the reply: *@USER Perché per un canguro è ancora presto."* (Because for a kangaroo it's still too early.") was labeled as CONTEXT SHIFT by both annotators and the model. The label was assigned due to the sudden change in topic, introducing an unexpected element (the kangaroo) that breaks coherence and signals irony.

In the remaining 11 cases where the model's prediction did not match humans' annotations, the model frequently labeled replies as OXYMORON PARADOX when annotators had chosen OTHER—this occurred in 6 out of the 11 pairs.

Consider the following example: *"Salvini ripropone il ponte sullo stretto di Messina, opera imprescindibile per lo sviluppo economico. Condivido e rilancio: contestualmente realizzerei anche il tunnel sottomarino Civitavecchia - Cagliari. Dai non facciamo come al solito la figura dei barboni, pensiamo in grande" ("Salvini reintroduces the Strait of Messina bridge proposal, a crucial infrastructure for economic development. I agree and raise: let's also build the Civitavecchia–Cagliari submarine tunnel. Let's not be our usual broke selves—let's think big!")* with the reply: *"Si può proporre il ponte Palermo–Cagliari già che ci siamo… una spesa unica… compri uno, paghi tre… no com'è la storia?" ("We might as well propose a Palermo–Cagliari bridge while we're at it… one payment for three projects… or how does it go again?")*

Here, the model likely interpreted the absurdity of the reply as a rhetorical figure of type OXYMORON PARADOX, whereas human annotators labeled it as a case of sarcasm, and thus as OTHER.

An illustrative example of the remaining cases is the

following: *"Lo scrivo per tanti idioti che rispondono ai Twitter come le pecore. Sono un Sovranista, non sono vaccinato, non pagherò la multa e la mia Libertà non è in svendita." ("I write this for all the idiots who respond to tweets like sheep. I'm a sovereignist, I'm unvaccinated, I won't pay the fine, and my freedom is not for sale.")* with the reply: *"Lo scrivo per te … non bere più" ("I write this for you… stop drinking.")*

In this case, the model assigned the label ANALOGY, possibly misled by the introductory phrase in the post, failing to capture the sarcastic tone of the reply. This example suggests that prompt design could be improved to better guide the model's focus toward the reply and its pragmatic intent.

This evaluation highlights the LLM's ability to produce overall reasonable outputs. Although its performance is not particularly high, it can still serve as a useful tool for silver annotation, thanks to the reasoning and explanations it provides.

7. Irony Generation

Inspired by previous work on irony generation [16], we investigated whether a model trained to classify rhetorical figures would also be capable of employing them during generation—producing ironic outputs comparable to those written by humans in terms of rhetorical figures. To explore this hypothesis, we considered the 278 post–reply pairs selected in Section 6, using the posts as input to the best-performing model for rhetorical figure classification. The model was prompted to generate an ironic reply for each post, which was then compared to the original human-written reply. As a baseline, we used the same model in its non–fine-tuned version, applying the same prompting strategy. To illustrate this process, we provide the following example:

> **Instruction**: Ti viene fornito in input (INPUT) un post estratto da conversazioni sui social media. Fornisci in output (OUTPUT) una risposta ironica in italiano. (*You are given as input (INPUT) a post extracted from social media conversations. Provide as output (OUTPUT) an ironic reply in Italian.*)
>
> **Input**: Consigli su workout in casa in questo periodo di palestre chiuse? (*Any tips for home workouts during this period of gym closures?*)
>
> **Output**: Ah, certo! Prova a fare squat con il divano, push-up con il tavolino e flessioni con il gatto! (*Oh, sure! Try doing squats with the couch, push-ups using the coffee table, and some presses with the cat!)*

Table 4
Distribution of rhetorical figures in human and model-generated ironic replies (rep.) from MultiPICo. CoT-FT refers to the fine-tuned model; Baseline to the non-fine-tuned version.

	Human rep.	**Model rep.** CoT-FT	Baseline
ANALOGY	40	58	41
HYPERBOLE	5	3	2
EUPHEMISM	5	9	6
RHETORICAL QUESTION	45	34	64
OXYMORON PARADOX	61	67	51
CONTEXT SHIFT	72	62	52
FALSE ASSERTION	32	35	34
OTHER	18	10	28

Table 4 presents the distribution of rhetorical figures in the ironic replies generated by humans, the fine-tuned model, and the baseline model, all classified by Ministral-8B with CoT-FT. Overall, the differences across distributions are not substantial, but some trends are worth noting.

The fine-tuned model produces slightly more ANALOGY and EUPHEMISM compared to humans, which may reflect the influence of the TWITTIRÒ training data, where these categories are relatively well represented. Conversely, CONTEXT SHIFT appears underrepresented in the model outputs compared to human replies, which could be due to either the complexity of capturing discourse-level phenomena.

Interestingly, the baseline model shows a notable increase in the use of RHETORICAL QUESTION and OTHER, suggesting a more generic or less targeted use of rhetorical strategies when the model is not fine-tuned. This may indicate that zero-shot generation leads to a reliance on broadly applicable or ambiguous rhetorical patterns, as already seen in Balestrucci et al. [16].

To better understand these patterns and assess the reliability of the automatic classification, we conducted a human evaluation on a subset of 20 model-generated replies from both systems.

Specifically, the same two annotators from Section 6.1 independently labeled the rhetorical figures predicted by the models. Inter-annotator agreement was substantial, with a Cohen's κ of 0.68 and a Krippendorff's α of 0.65. In contrast, the Krippendorff's α between the annotators and the classifier was 0.26, confirming all the previous results.

7.1. Linguistic Analysis

Following the approach proposed by Balestrucci et al. [16], we also conducted a linguistic analysis focusing on specific stylistic markers—namely, average token length, type-token ratio (TTR), and the use of interjections and negations—across human-written replies and model-generated outputs.

Table 5
Linguistic analysis for human-written posts, human-written replies, fine-tuned model generations (CoT-FT), and baseline generations (Baseline): average number of tokens (Tokens), type/token ratio (TTR), and average occurrences of interjections (Interjections) and negations (Negations).

	Human		**Model Replies**	
	Post	Reply	CoT-FT	Baseline
Tokens	30.586	12.471	20.173	22.399
TTR	0.924	0.956	0.938	0.935
Interjections	0.594	0.273	0.381	0.507
Negations	0.050	0.072	0.410	0.982

Table 5 reports a linguistic analysis of human-written replies compared to those generated by the fine-tuned and baseline models. The comparison includes the average number of tokens, type-token ratio (TTR), and the average occurrences of interjections and negations.

Human replies tend to be shorter (12.47 tokens on average) than those generated by both the fine-tuned model (20.17) and the baseline (22.40), suggesting that human-written irony is often more concise. The type-token ratio remains high across all outputs, indicating a generally rich lexical variety. Notably, the TTR of the fine-tuned model (0.938) is slightly higher than that of the baseline (0.935), and closer to the human replies (0.956), suggesting that fine-tuning may help preserve or recover some degree of lexical diversity.

Regarding stylistic markers, human replies make limited use of interjections (0.273 per reply), while both models tend to use them more frequently—especially the baseline (0.507), possibly as a compensatory strategy to signal irony more explicitly. A similar trend is observed for negations: while human replies contain very few (0.072), model generations show a noticeable increase—particularly in the baseline output (0.982). This may indicate a tendency of the baseline model to overuse negative constructions, possibly due to a lack of fine control over tone and pragmatics in ironic generation.

Overall, these findings suggest that while model outputs differ in length and surface features from human replies, the fine-tuning on rhetorical figure classification task helps reduce some of the stylistic drift, bringing the generations closer to human-like patterns in terms of lexical variation and use of pragmatic markers.

8. Conclusions

Our study explored the extent to which rhetorical figures can serve as a bridge between the detection and generation of ironic content in Italian. We showed that fine-tuning LLMs on rhetorical figure classification enables models to identify key linguistic devices involved in irony with reasonable accuracy. The best results were obtained using a CoT strategy, which guided models to provide explanations before predicting the rhetorical category. While the models performed well on frequently represented figures such as ANALOGY and RHETORICAL QUESTION, they struggled with more subtle or under-represented categories like EUPHEMISM, suggesting that further refinement and data augmentation may be needed.

For the irony generation task, we observed that models fine-tuned on rhetorical figure classification produced ironic replies that more closely resembled human outputs in terms of rhetorical devices and stylistic markers. Although the overall distribution of rhetorical figures remained similar across models, the fine-tuned version demonstrated a more balanced use of devices, reducing the over-reliance on rhetorical questions and interjections observed in the baseline. This suggests that rhetorical figure awareness acquired through classification can positively influence generation, even in the absence of explicit training on ironic text generation.

Manual evaluation confirmed the model's ability to generate plausible annotations and replies, albeit with fair agreement compared to human annotators. Nonetheless, the consistency and interpretability of its outputs highlight its potential as a tool for silver annotation—particularly valuable in low-resource settings. Finally, our linguistic analysis showed that the fine-tuned model better preserved lexical diversity and pragmatic subtlety than its non-fine-tuned counterpart, indicating that rhetorical figure classification fine-tuning may also serve as a form of stylistic control. Taken together, these findings point to the value of leveraging rhetorical figures to enhance both the interpretability and expressiveness of LLMs in pragmatic language generation.

As future work, we plan to extend this study to other languages, such as French and English, with the goal of comparing the capacity of LLMs to classify rhetorical figures and generate ironic content across different linguistic contexts.

Moreover, a key research direction we intend to pursue concerns the perspectivist nature of the MultiPICo dataset. In particular, we aim to explore whether rhetorical figures function as shared cues in the perception of irony across different sociodemographic groups, thereby pointing to the existence of rhetorical devices that act as universal markers of ironic intent.

9. Limitations

Despite the promising results, this work presents several limitations that call for further investigation.

First, the rhetorical figure classification task was trained and evaluated on a relatively small dataset (TWITTIRÒ-UD), which may hinder the generalizability of the models—particularly for under-represented categories such as EUPHEMISM and HYPERBOLE. While fine-tuning contributes to improved performance, the models still struggle with these categories, likely due to data sparsity and the intrinsic ambiguity of certain rhetorical devices.

Second, the human evaluation was conducted on a relatively limited subset, which reduces the statistical robustness of the agreement scores. Although the results align with previous studies and provide qualitative insights into model behavior, a larger annotation effort would be needed to draw more conclusive findings—especially when distinguishing between closely related rhetorical categories. However, large-scale human annotation remains time-consuming and costly.

Finally, this study did not include a direct comparison with models explicitly fine-tuned for irony generation. Such a comparison would be necessary to better assess the specific contribution of rhetorical figure classification to the generation of ironic content, and to determine whether the observed improvements are attributable to rhetorical awareness or other factors.

Acknowledgments Michael Oliverio was partially funded by the 'Multilingual Perspective-Aware NLU' project in partnership with Amazon Alexa.

References

[1] D. C. Muecke, Irony and the Ironic, Methuen, London, 1970.

[2] S. Sravanthi, M. Doshi, P. Tankala, R. Murthy, R. Dabre, P. Bhattacharyya, Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities, in: Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 12075–12097.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[4] J. Karoui, F. Benamara, V. Moriceau, V. Patti, C. Bosco, N. Aussenac-Gilles, Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 262–272. URL: https://aclanthology.org/E17-1025/.

[5] A. Athanasiadou, H. L. Colston, The Diversity of Irony, volume 65, Walter de Gruyter GmbH & Co KG, 2020.

[6] M. Mladenovic, Ontology-based recognition of rhetorical figures, Infotheca, Journal for Digital Humanities 16 (2016) 24–47.

[7] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, M. Wroczynski, Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection, Information Processing & Management 58 (2021) 102600.

[8] C. W. Strommer, Using rhetorical figures and shallow attributes as a metric of intent in text (2011).

[9] M. Dubremetz, J. Nivre, Rhetorical figure detection: Chiasmus, epanaphora, epiphora, Frontiers in Digital Humanities Volume 5 - 2018 (2018). URL: https://www.frontiersin.org/journals/digital-humanities/articles/10.3389/fdigh.2018.00010. doi:10.3389/fdigh.2018.00010.

[10] L. Neuhaus, On the relation of irony, understatement, and litotes, Pragmatics & Cognition 23 (2016) 117–149.

[11] C. Burgers, M. van Mulken, P. J. Schellens, Type of evaluation and marking of irony: The role of perceived complexity and comprehension, Journal of Pragmatics 44 (2012) 231–242.

[12] M. Zhu, Z. Yu, X. Wan, A neural approach to irony generation, ArXiv abs/1909.06200 (2019). URL: https://api.semanticscholar.org/CorpusID:202572954.

[13] Y. Tian, D. Sheth, N. Peng, A unified framework for pun generation with humor principles, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 3253–3261. URL: https://aclanthology.org/2022.findings-emnlp.237. doi:10.18653/v1/2022.findings-emnlp.237.

[14] Q. Zeng, A.-R. Li, A survey in automatic irony processing: Linguistic, cognitive, and multi-X perspectives, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea,

2022, pp. 824–836. URL: https://aclanthology.org/2022.coling-1.69.

[15] A. Mishra, T. Tater, K. Sankaranarayanan, A modular architecture for unsupervised sarcasm generation, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6144–6154. URL: https://aclanthology.org/D19-1636. doi:10.18653/v1/D19-1636.

[16] P. F. Balestrucci, S. Casola, S. M. Lo, V. Basile, A. Mazzei, I'm sure you're a real scholar yourself: Exploring ironic content generation by large language models, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 14480–14494. URL: https://aclanthology.org/2024.findings-emnlp.847/. doi:10.18653/v1/2024.findings-emnlp.847.

[17] A. T. Cignarella, C. Bosco, V. Patti, et al., Twittiro: a social media corpus with a multi-layered annotation for irony, in: CEUR Workshop Proceedings, volume 2006, CEUR, 2017, pp. 1–6.

[18] A. Cignarella, C. Bosco, V. Patti, TWITTIRÒ: a Social Media Corpus with a Multi-layered Annotation for Irony, 2017, pp. 101–106. doi:10.4000/books.aaccademia.2382.

[19] S. Casola, S. Frenda, S. M. Lo, E. Sezerer, A. Uva, V. Basile, C. Bosco, A. Pedrani, C. Rubagotti, V. Patti, D. Bernardi, MultiPICo: Multilingual perspectivist irony corpus, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 16008–16021. URL: https://aclanthology.org/2024.acl-long.849/. doi:10.18653/v1/2024.acl-long.849.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[21] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37 – 46. URL: https://api.semanticscholar.org/CorpusID:15926286.

[22] K. Krippendorff, Computing krippendorff's alpha-reliability, 2011.

A. Experimental Setup

This appendix reports the hyperparameter configuration used during model fine-tuning. All experiments were performed using LoRA. Training was conducted using the `transformers` and `peft` libraries. The table below summarizes the main parameters used in the `TrainingArguments` class and in the LoRA configuration.

Table 6
Configuration of hyperparameters used in the LoRA-based fine-tuning process.

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