Rhetorical Device-Aware Sarcasm Detection with Counterfactual Data Augmentation

Qingqing Hong[1,2], Dongyu Zhang[1,2], Jiayi Lin[1,2],
Dapeng Yin[1,2], Shuyue Zhu[1,2], Junli Wang[1,2]*
[1] Key Laboratory of Embedded System and Service Computing (Tongji University),
Ministry of Education, Shanghai 201804, China.
[2] National (Province-Ministry Joint) Collaborative Innovation Center for
Financial Network Security, Tongji University, Shanghai 201804, China.
{2332012, yidu, 2331908, 2432122,2432272, junliwang}@tongji.edu.cn

Abstract

Sarcasm is a complex form of sentiment expression widely used in human daily life. Previous work primarily defines sarcasm as a form of verbal irony, which covers only a subset of real-world sarcastic expressions. However, sarcasm serves multifaceted functions and manifests itself through various rhetorical devices, such as echoic mention, rhetorical question and hyperbole. To fully capture its complexity, this paper investigates fine-grained sarcasm classification through the lens of rhetorical devices, and introduces **RedSD**, a **R**h**E**torical **D**evice-Aware **S**arcasm **D**ataset with counterfactually augmented data. To construct the dataset, we extract sarcastic dialogues from situation comedies (i.e., sitcoms), and summarize nine rhetorical devices commonly employed in sarcasm. We then propose a rhetorical device-aware counterfactual data generation pipeline facilitated by both Large Language Models (LLMs) and human revision. Additionally, we propose duplex counterfactual augmentation that generates counterfactuals for both sarcastic and non-sarcastic dialogues, to further enhance the scale and diversity of the dataset. Experimental results on the dataset demonstrate that fine-tuned models exhibit a more balanced performance compared to zero-shot models, including GPT-3.5 and LLaMA 3.1, underscoring the importance of integrating various rhetorical devices in sarcasm detection.[1]

1 Introduction

Sarcasm is a subtle and peculiar form of sentiment expression, often employed to criticize or ridicule a person, situation or idea. Refer to the formal description of sarcasm as presented in *A Dictionary of Modern English Usage* (Fowler, 1926):

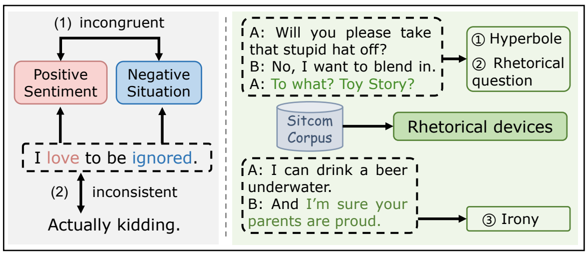

Figure 1: Comparison of traditional sarcasm and our rhetorical device-aware sarcasm. The left side of the figure defines sarcasm as a way of verbal irony, typically relying on complex model structures to detect either (1) emotional incongruity or (2) logical inconsistency. The right side presents three examples of rhetorical device-aware sarcasm in dialogue. The phrase "To what? Toy Story?" employs ① hyperbole to emphasize the absurdity of speaker B's hat and reinforces the sarcastic tone through ② rhetorical question. Notably, ③ irony is also treated as a rhetorical device for expressing sarcasm.

> "*Sarcasm does not necessarily involve irony, and irony has often no touch of sarcasm. But irony... is so often made the vehicle of sarcasm... The essence of sarcasm is the intention of giving pain by (ironical or other) bitter words.*"

With the universal existence of sarcasm, **S**arcasm **D**etection (**SD**) plays a vital role in tasks such as sentiment analysis, opinion mining, and hate speech detection (Rosenthal et al., 2014), all of which depend on accurately capturing genuine human sentiments (Li et al., 2021a). However, understanding sarcasm requires commonsense knowledge and logical reasoning (Poria et al., 2016). Additionally, in text-only settings, sarcasm detection models are particularly sensitive to the presence or absence of contextual cues (Kim et al., 2024; Jang and Frassinelli, 2024). Even for people, it is not always easy to identify sarcasm in a single tweet without prior conversational context (Riloff et al., 2013), highlighting the need to construct a

*Corresponding author. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB4501704.

[1] Our dataset is avaliable at https://github.com/qqHong73/RedSD.

Findings of the Association for Computational Linguistics: ACL 2025, pages 11672–11685
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

high-quality corpus of sarcasm in dialogue.

Most prior work defines sarcasm as a way of verbal irony where someone says the opposite of what they mean (Liu et al., 2022a; Min et al., 2023; Li et al., 2021b; Yue et al., 2024; Kim et al., 2024). Consequently, automatic sarcasm detection methods predominantly focus on exploring the incongruity between positive sentiment and negative situation (Riloff et al., 2013; Min et al., 2023), or evaluating the inconsistency between the actual intention and the literal content (Liu et al., 2023), as illustrated in Figure 1. Despite their effectiveness, a notable concern is whether current sarcasm detection models are robust enough to capture the complexity and diversity of sarcasm.

Recently, some researchers have argued that this narrow definition of sarcasm provides a foundation that is neither necessary nor sufficient for sarcasm to occur (Oprea et al., 2021; Jang and Frassinelli, 2024). Meanwhile, several studies have embarked on fine-grained sarcasm detection. For example, Oraby et al. (2016) operationalize classes of sarcasm in the form of rhetorical question and hyperbole, Abu Farha et al. (2022a) propose to further label each text into one of the categories: sarcasm, irony, satire, understatement, overstatement, and rhetorical question, and Ray et al. (2022) extend the MUStARD dataset with sarcasm types that specify the necessary information or modality for sarcasm detection. While these studies have advanced our understanding of sarcasm, they fail to encompass the full spectrum of sarcastic expressions or delve into the intrinsic nature of sarcasm itself. Therefore, it is imperative to explore a more nuanced and comprehensive classification system for sarcasm.

Since sarcasm often manifests through various rhetorical devices that simultaneously contribute to its complex and multifaceted nature, this work explores fine-grained sarcasm classification through the lens of rhetorical devices and integrating them into sarcasm detection. Specifically, we introduce **RedSD**, a **R**h**E**torical **D**evice-Aware **S**arcasm **D**ataset with counterfactually augmented data. Inspired by Castro et al. (2019), we utilize sitcom corpus to extract sarcastic dialogues with various rhetorical devices. As shown in Figure 1, each sarcastic dialogue may involve multiple rhetorical devices. To learn more robust sarcasm detection models, we employ ChatGPT (OpenAI, 2023) to generate counterfactuals and incorporate rhetorical devices into the prompts. After automatically filtering and and human revision, we ultimately curate a new sarcasm detection dataset with an equal number of sarcastic and non-sarcastic dialogues. To further enhance the scale and diversity of our dataset, we propose duplex counterfactual augmentation, which generates counterfactuals for both sarcastic and non-sarcastic dialogues. In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to explore fine-grained sarcasm classification through the lens of rhetorical devices.
- We propose a novel counterfactual data generation pipeline and duplex counterfactual augmentation based on rhetorical devices.
- We introduce RedSD, a new sarcasm detection dataset comprising 2k dialogues and nine rhetorical devices, to develop more robust sarcasm detection models.
- We conduct a series of experiments with our dataset, demonstrating the necessity and effectiveness of integrating rhetorical devices for improved sarcasm detection.

2 Related work

2.1 Sarcasm Detection

Previous work on sarcasm detection can be broadly classified into two main areas: creating datasets and creating models.

Creating datasets The availability of high-quality datasets is indispensable for sarcasm detection. Traditionally, distant supervision (Ptáček et al., 2014) and manual labeling (Filatova, 2012; Abercrombie and Hovy, 2016; Oraby et al., 2016; Khodak et al., 2018; Castro et al., 2019; Oprea and Magdy, 2020; Yue et al., 2024; Jang and Frassinelli, 2024) are utilized to collect sarcasm datasets (Abu Farha et al., 2022b). Distant supervision is easy and scalable, but the data tends to be saturated with casual expressions and lacks contextual information. Meanwhile, manually labeling and annotation are inefficient, costly, and typically limited in scale and diversity. For instance, Castro et al. (2019) propose MUStARD, a sarcasm dataset compiled from sitcoms, which is relatively small and may exhibit suboptimal performance in text-only settings. Recently, leveraging LLMs for data labeling and annotation has been applied to sarcasm dataset construction (Kim et al., 2024), presenting a promising research direction.

Creating models The mainstream methods either explore the incongruity between the positive sentiment and the negative situation (Riloff et al., 2013; Min et al., 2023), or evaluate the inconsistency between the actual intention and the literal content (Liu et al., 2023). Nowadays, the surge in multimodal content has propelled the field of multimodal sarcasm detection (MSD), with the objective of detecting both inter- and intra-modal incongruities (Liang et al., 2022; Liu et al., 2022b; Jia et al., 2024; Chen et al., 2024). However, these efforts primarily address only irony and are therefore not comprehensive. There is also a lot of work that focuses on introducing new tasks to advance sarcasm detection models, such as Sarcasm Explanation in Dialogue (SED) (Kumar et al., 2022), Sarcasm Initiation and Reasoning in Conversations (SIRC) (Singh et al., 2024), aiming to capture the authentic essence of sarcasm. In this work, we leverage the internal knowledge and reasoning capabilities of LLMs based on rhetorical devices.

2.2 Counterfactual Sarcasm Detection

Counterfactual Data Augmentation is an increasingly prevalent approach in many natural language processing (NLP) tasks (Kaushik et al., 2020; Qin et al., 2019; Wu et al., 2021; Paranjape et al., 2022; Ross et al., 2022). In the field of sarcasm detection, Oprea and Magdy (2020) ask the authors of sarcastic tweets to provide non-sarcastic rephrases, which aligns with the concept of counterfactuals. Jia et al. (2024) propose tailored augmentation methods to rewrite sarcastic and non-sarcastic samples separately. For sarcastic samples, they simply use ChatGPT (Brown et al., 2020) to reverse the sentiment polarity. For non-sarcastic samples, they select a target entity that appears in both visual and textual modalities. However, the generated counterfactuals lack diversity and may contain logical inconsistencies. Our approach incorporates various rhetorical devices to generate high-quality counterfactuals.

3 Methodology and Dataset

In this section, we present our *Rhetorical Device-Aware Counterfactual Sarcasm Detection* framework, which consists of three phases: Rhetorical Device-Aware Data Collection (RDDC, §3.1), Counterfactual Data Augmentation (CDA, §3.2) and Duplex Counterfactual Augmentation (DCA, §3.3), as illustrated in Figure 2. In addition, we conduct dataset analysis in §3.4.

3.1 Rhetorical Device-Aware Data Collection

We compile a corpus of sarcastic dialogues from *The Big Bang Theory*, a TV show whose characters are often perceived as sarcastic. Since the show vividly depicts human behavior and interactions with richly detailed context, it allows for reliable inference of the authors' intentions, thereby mitigating labeling inconsistencies. Furthermore, the corpus encompasses various rhetorical devices across diverse scenarios, establishing a foundation for fine-grained sarcasm classification.

Specifically, we manually extract 1,018 sarcastic dialogues from season one to season twelve. Based on a thorough review of relevant literature on sarcasm theory and the characteristics of the sitcom corpus, we identify nine distinct rhetorical devices commonly used in sarcasm. In addition to five previously examined types of sarcasm, including *irony*, *echoic mention* (Sperber and Wilson, 1981; Oprea et al., 2021), *hyperbole*, *rhetorical question* (Oraby et al., 2016; Oprea and Magdy, 2020), *self-deprecation* (Abulaish and Kamal, 2018), we further introduce the following four types:

- **Presupposition** (Utsumi, 2000; Bajri, 2016): an implicit assumption about a shared belief or mutual knowledge between speakers.
- **Innuendo** (Camp, 2012): implying something negative or critical without stating it explicitly.
- **Intentional Reenactment**: highlighting the absurdity of a situation or contrasting one's words with their actions through detailed and exaggerated depictions.
- **Unexpected Twist**: a sudden or surprising shift, often starting with a seemingly straightforward or expected path.

For each sarcastic dialogue, we ensure its context is sufficient for reliable sarcasm detection and manually annotate the corresponding sarcastic segment and rhetorical device. The rhetorical device annotation process is guided by predefined definitions and detailed examples. For each rhetorical device, we manually annotate two samples as exemplars (ICL examples ① in Figure 2). For example, if the sarcastic dialogue involves hyperbole, we selectively incorporate hyperbole-based ICL examples. Notably, all annotation processes are conducted by a single annotator with comprehensive contextual understanding and domain knowledge of the corpus.

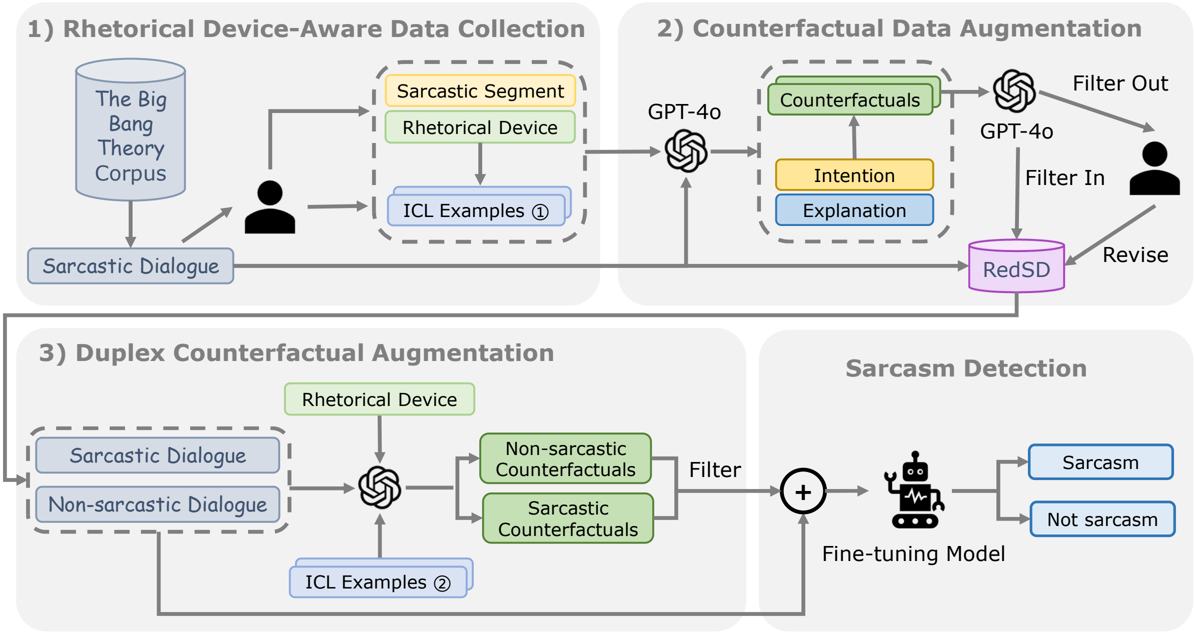

Figure 2: The illustration of our framework. The first two phases complete the construction of the sarcasm dataset, RedSD, while the third phase further enhances the scale and diversity of the dataset without human supervision.

To mitigate potential biases, we implement iterative refinement by reviewing the entire dataset multiple times. For ambiguous cases, GPT-4 is used as an additional reference to facilitate decision-making and improve label reliability.

3.2 Counterfactual Data Augmentation

To obtain non-sarcastic dialogues from existing rhetorical device-aware sarcastic dialogues and construct a fine-grained sarcasm dataset, we employ LLMs to generate counterfactuals, which are extensively utilized to mitigate spurious correlation by altering the causally salient parts of instances that contributes to the label assignment (Dixit et al., 2022; Chen et al., 2023). Although LLMs have shown impressive generative capabilities, directly prompting them to transform sarcastic dialogues into non-sarcastic counterfactuals may yield unsatisfactory outcomes. These include simplistic or generic responses, failure to accurately identify the sarcastic segments, correctly flip the label, or logically maintain coherence with the context.

Given the highly subjective and complex nature of sarcasm, we incorporate manually annotated sarcastic segments, rhetorical devices, and ICL examples into the prompts of GPT-4o. As a way of Chain-of-Thought (CoT) prompting (Wei et al., 2022), we instruct GPT-4o to first interpret the speaker's underlying intention and then provide a concise explanation for why the dialogue contains sarcasm. This two-step process facilitates the generation of high-quality counterfactuals by ensuring a deep understanding of the context. The details of the prompt can be found in Appendix B.

To enhance the quality of the generated counterfactuals, we apply GPT-4o to automatically filter out undesired dialogues, including those that are unnatural, incoherent, or still contain sarcasm. Any data identified as undesired is subsequently sent to human annotators (the same annotator as in §3.1) for verification and revision. Finally, we obtain a new sarcasm dataset (RedSD), which is composed of 2,036 dialogues, with an equal number of sarcastic and non-sarcastic dialogues.

3.3 Duplex Counterfactual Augmentation

LLMs have demonstrated their strong performance in sarcasm detection (Gole et al., 2023). To further enhance the scale and diversity of our dataset, we propose Duplex Counterfactual Augmentation. Unlike the sarcasm-to-nonsarcasm transformation described in §3.2, DCA introduces bidirectional counterfactual generation without human revision, with the aim to rapidly expand the dataset scale. Notably, the data generated at this stage is used solely for additional data augmentation and is not included in the RedSD dataset.

Specifically, we use existing pairs of sarcastic and counterfactual non-sarcastic dialogues in RedSD as inputs, and instruct GPT-4o to generate

Rhetorical Device	Irony	Echo.	Hyperbole	Rhet. Q.	Presupposition	Innuendo	Reenact.	Twist	Self-dep.
Number	229	21	364	95	440	126	162	47	20

Table 1: The overall statistics of nine rhetorical devices in RedSD. Due to space limitation, we use the following abbreviations: Echo. for echoic mention, Rhet. Q. for rhetorical question, Reenact. for intentional reenactment, Twist for unexpected twist, and Self-dep. for self-deprecation.

Phase	RDDC	CDA	DCA
Number	1,018	1,018	4,043

Table 2: Number of dialogues generated at each phase.

new pairs: non-sarcastic counterfactuals and sarcastic counterfactuals. For each rhetorical device, we manually annotate one sample as an exemplar (ICL examples ② in Figure 2). This process is divided into the following two steps:

- *Rewrite Sarcastic Dialogue*: Analyze the given two dialogues to understand the nuances between sarcasm and non-sarcasm, and convert the sarcastic dialogue into a straightforward, non-sarcastic version.

- *Rewrite Non-Sarcastic Dialogue*: Select an appropriate target (a person, object, or situation) for the sarcasm and specify a rhetorical device, allowing for precise manipulation of the generated sarcastic dialogues.

For non-sarcastic counterfactuals, we apply GPT-4o to automatically filter out undesired dialogues, as detailed in §3.2. To ensure diversity, we discard highly similar dialogues with a Jaccard similarity score (Jaccard, 1901) greater than 0.8. For sarcastic counterfactuals, GPT-4o frequently produces ironic expressions, even when explicitly instructed to avoid irony if the specified rhetorical device is not irony, which conflicts with our premise that sarcasm can be expressed through various rhetorical devices beyond irony. Furthermore, the generated dialogues often incorporate typical ironic cue words (e.g., *sure*, *because*, *definitely*, *absolutely*, *of course*, *oh*, *yes*), which may introduce spurious correlations between lexical patterns and sarcasm labels. To mitigate this issue, we add a penalty of 0.6 to the Jaccard similarity score when such ironic cue words appear, aiming to reduce the frequency of these cue words. Moreover, we exclude dialogues with a similarity score below 0.2 to avoid excessive rewriting.

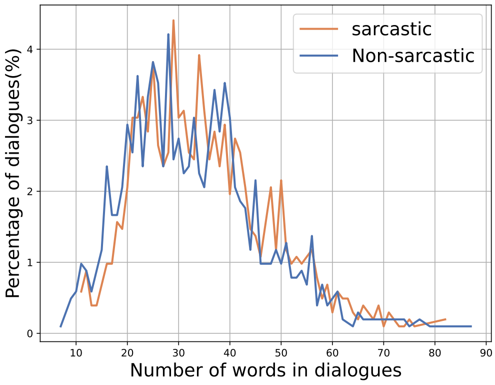

Figure 3: The length distribution across RedSD.

3.4 Dataset Analysis

Rhetorical Devices Distribution As shown in Table 1, hyperbole and presupposition are the two largest categories of rhetorical devices in sarcastic dialogues, while echoic mention and self-deprecation are the two smallest ones. Each sarcastic dialogue may utilize multiple rhetorical devices. For instance, innuendo is commonly employed in conjunction with presupposition. Hence, the total count of rhetorical device occurrences exceeds the number of total sarcastic dialogues.

Overall Statistics The resulting dataset (RedSD) contains 2,036 dialogues with an equal number of sarcastic and non-sarcastic dialogues. Table 2 presents the number of dialogues generated in each phase. The average length of sarcastic and non-sarcastic dialogues is around 35 and 33 words, respectively. Figure 3 shows the length distribution across RedSD, revealing similar patterns between sarcastic and non-sarcastic dialogues. This suggests that our CDA pipeline effectively preserves the characteristics of original dialogues while reversing the sarcasm labels.

Rhetorical Devices Examples Table 3 shows four examples of sarcastic dialogues alongside their corresponding counterfactuals. In the innuendo example, the sarcastic segment is found not in the last response but in the second one, while the last

Rhetorical device	Sarcastic Dialogue	Counterfactual
Innuendo	"Oh, my God. I love this chicken.", "Oh, you know what they say, best things in life are free.", "Okay, you're right. I eat your food a lot."	"Oh, my God. I love this chicken.", "I notice you often eat our food without paying.", "Yeah, you're right. I eat your food a lot."
Presupposition	"I believe that social convention dictate you not arriving empty-handed. Would you like to bring some Cylon toast?", "Yeah, no, I'm trying to fit in, not get laughed at."	"I believe that social convention dictates you not arriving empty-handed. Would you like to bring some Cylon toast?", "No, I'd prefer to bring something more conventional."
Intentional Reenactment	"Well, ever since she helped me get this job, she won't stop bugging me.", "Well, I think she just wants you to do well, and she's worried that you won't 'cause you were just a stunningly poor waitress.", "That is not true.", "I'm still waiting on my mini corndogs from two years ago."	"Well, ever since she helped me get this job, she won't stop bugging me.", "Well, I think she just wants you to do well, and she's worried that you won't because you didn't perform well as a waitress.", "That is not true.", "Yes, it is. You used to forget our orders when you were a waitress."
Unexpected Twist	"He then gave an example of something he had to do, even though he didn't want to, which was look at my stupid face.", "That's a rude thing to say. Out loud."	"He then gave an example of something he had to do, even though he didn't want to, which was look at my stupid face.", "That's a rude thing to say."

Table 3: Examples of sarcastic dialogues and their corresponding counterfactuals in RedSD. Due to space limitation, we present only four of the rhetorical devices. Red spans represent sarcastic segments and blue spans represent modifications during counterfactual data augmentation.

Metric	Regular	Rhetoric-Guided	Human-Refined
LFR (%) ↑	62.0	70.0	86.0
Plausibility ↑	4.45	4.59	4.56
CA ↑	4.45	4.53	4.40
CP (%) ↑	94.3	94.0	95.1

Table 4: Human and automatic evaluation results of the counterfactual data genearated by CDA across three experimental settings.

response provides significant contextual clues for detecting sarcasm within the segment. The presupposition example assumes that bringing Cylon toast would lead to ridicule. All counterfactuals are crafted to eliminate only the sarcastic tone while preserving the speaker's original intention under minimal modifications.

4 Counterfactual Quality Evaluation

To evaluate the quality of the counterfactual data generated by CDA, we define the following quantitative metrics. **1) Label-Flip Rate (LFR)**. LFR calculates the percentage of counterfactual data that flip the original label (sarcasm) to the target label (non-sarcasm). **2) Plausibility.** Plausibility measures the logical coherence of the context and whether the content aligns with commonsense knowledge. **3) Context Adequacy (CA)**. CA refers to the extent to which the context provides sufficient information to support the decision. **4) Content Preservation (CP)**. CP assesses the semantic similarity between sarcastic data and counterfactual data using BERTScore (Zhang et al., 2019).

We randomly select 50 samples to assess counterfactual quality in three experimental settings. 1) **Regular**, which provides only basic input-output examples without sarcastic segments or rhetorical devices. 2) **Rhetoric-Guided**, which includes both sarcastic segments and rhetorical devices, along with ICL examples to guide reasoning. 3) **Human-Refined**, which extends Rhetoric-Guided by incorporating human review and revision. Two annotators, blinded to the experimental settings, independently assess the generated data using metrics 1-3. For Plausibility and CA, we adopt a 5-point Likert scale (1: very poor; 5: excellent).

As shown in Table 4, the Human-Refined setting achieves a significantly higher LFR of 86.0%, compared to 62.0% for the Regular and 70.0% for the Rhetoric-Guided settings. This demonstrates that human review and revision substantially improve sarcasm removal and label flipping. The Rhetoric-Guided setting shows consistent improvements across most metrics relative to the Regular setting, underscoring the benefits of incorporating sarcastic segments and rhetorical devices in counterfactual data augmentation.

While the Human-Refined setting exhibits marginally lower scores in Plausibility and CA, it achieves the highest CP score. On the one hand, human revision prioritizes minimal modifications to preserve the original intention of the sarcastic dialogue, which may slightly compromise logi-

Model	Acc	Macro	Macro	Macro	Sarc	Sarc	Sarc	Non-Sarc	Non-Sarc	Non-Sarc
		F1	P	R	F1	P	R	F1	P	R
Zero-shot										
GPT-3.5-turbo	57.2	52.8	67.0	59.3	38.3	81.5	25.0	67.3	52.5	**93.6**
GPT-4-turbo	**88.0**	**87.8**	88.2	**87.7**	89.0	**86.2**	92.0	**86.7**	90.3	83.3
Claude 3.5 haiku	66.9	61.3	78.9	64.8	76.0	61.7	98.9	46.6	96.0	30.8
Claude 3.5 sonnet	88.0	87.7	**89.3**	87.4	**89.5**	83.3	96.6	85.9	95.3	78.2
LLaMA 3.1 8B	58.4	46.3	78.0	55.8	71.8	56.1	**100.0**	20.7	**100.0**	11.5
LLaMA 3.1 70B	75.9	74.4	80.0	74.7	80.6	70.3	94.3	68.3	89.6	55.1
LLaMA 3.1 405B	67.5	62.6	77.6	65.5	76.1	62.3	97.7	49.1	92.9	33.3
Fine-tuning										
BERT$_{base}$	74.5	74.5	74.5	74.5	76.0	76.1	75.9	73.0	72.8	73.1
BERT$_{large}$	73.9	73.8	73.9	73.8	75.8	74.7	77.0	71.8	73.1	70.5
RoBERTa$_{base}$	75.5	75.2	75.5	75.2	77.5	75.5	79.5	73.0	75.5	70.8
RoBERTa$_{large}$	73.8	73.6	73.8	73.6	75.6	74.7	76.7	71.6	73.0	70.5
Human Evaluation	84.5	84.4	85.5	84.5	85.0	86.0	85.1	83.8	85.0	84.0

Table 5: Experimental results (%) on the test set of RedSD. The best results are represented in bold. The second-best results are underlined.

cal consistency or contextual coherence. On the other hand, since the context remains identical and only the sarcastic utterance is rewritten, annotators may have subconsciously rated Plausibility and CA higher when exposed to the same context.

5 Experiments

5.1 Experimental Setup

Baselines Given that a limited number of ICL examples is insufficient to capture the nuances between sarcasm and non-sarcasm across various rhetorical devices, few-shot models tend to perform comparably to, or even worse than, zero-shot models, particularly when the provided examples are potentially misleading. Therefore, we experiment with four zero-shot models, GPT-3.5-turbo, GPT-4-turbo, LLaMA 3.1 and Claude 3.5 with varying parameter sizes. Additionally, to provide a comprehensive comparison, we fine-tune two encoder-only models from Transformers: BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019). All models are fine-tuned for 8 epochs with 4 different random seeds (1, 2, 12, 42), and all results reported in §5 are an average across all seeds.

Dataset We split our dataset into training, development, and testing sets with proportions of approximately 8:1:1. To ensure a balanced and consistent distribution of the nine rhetorical devices in all three sets, we employ stratified sampling. Furthermore, we implement the DCA strategy, which nearly triples the size of the training set. While the data generated from DCA is not part of our dataset, it is included in the training data.

Evaluation Metrics We evaluate the sarcasm detection results using seven metrics: Accuracy (Acc), Macro F1 score, Macro Precision (Macro P), Macro Recall (Macro R), as well as the F1 score, Precision (P) and Recall (R) for both sarcasm and non-sarcasm classes.

5.2 Main Results

As reported in Table 5, the performance of zero-shot models varies significantly. GPT-4-turbo and Claude 3.5 Sonnet demonstrate superior and balanced performance across most metrics, owing to their robust reasoning capabilities and extensive internal knowledge. GPT-3.5-turbo struggles to accurately detect sarcasm, possibly because certain rhetorical devices used in sarcasm are subtle and difficult to interpret. In contrast, Claude 3.5 haiku and all parameter versions of LLaMA 3.1 achieve impressive performance in sarcastic recall and non-sarcastic precision, while underperforming in sarcastic precision and non-sarcastic recall. This suggests an inherent bias toward misclassifying non-sarcastic samples as sarcastic, which limits their practical application, especially in scenarios where balanced performance is essential.

However, fine-tuned models consistently perform well across all metrics, suggesting that while LLMs possess powerful zero-shot learning capabilities, task-specific fine-tuning is more effective in handling sarcasm in complex and diverse scenarios. For human evaluation, two annotators without prior background knowledge independently label the test set. The final ratings are derived by averaging the two annotators' judgments, and the overall inter-

Model	Dataset	Rhetorical Device							
		Irony	Echo.	Hyperbole	Rhet. Q.	Presup.	Innuendo	Reenact.	Twist
BERT_{base}	CSC	53.4	83.3	66.5	62.5	64.0	65.0	54.5	42.9
	MUStARD	61.5	66.7	59.7	53.1	59.2	61.3	58.0	57.1
	RedSD (Ours)	76.4	75.0	80.1	75.0	77.7	53.8	69.6	50.0
BERT_{large}	CSC	62.2	91.7	66.1	62.5	65.8	60.0	65.2	35.7
	MUStARD	56.8	75.0	58.9	40.6	54.8	57.5	58.0	57.1
	RedSD (Ours)	73.0	91.7	75.8	62.5	77.7	56.2	69.6	60.7
RoBERTa_{base}	CSC	56.8	66.7	61.0	56.2	60.3	75.0	71.4	42.9
	MUStARD	58.1	58.3	60.6	51.6	59.6	73.8	50.9	50.0
	RedSD (Ours)	73.6	83.3	76.3	67.2	72.9	61.3	69.6	71.4
RoBERTa_{large}	CSC	58.1	75.0	71.6	62.5	64.0	65.0	63.4	39.3
	MUStARD	62.2	77.8	62.7	52.1	66.2	73.3	48.8	42.9
	RedSD (Ours)	73.6	83.3	71.6	75.0	71.6	65.0	69.6	71.4

Table 6: Accuracy (%) across different rhetorical devices of different models trained on different datasets when tested on the test set of our dataset.

Models	Fine-tuned on		Predicted on		
		Intra-dataset	CSC	MUStARD	RedSD (Ours)
BERT_{base}	CSC	67.7	-	49.9	56.1
	MUStARD	64.1	50.9	-	59.2
	RedSD (Ours)	74.7	56.6	46.0	-
BERT_{large}	CSC	68.1	-	56.1	64.5
	MUStARD	57.7	48.4	-	55.2
	RedSD (Ours)	73.2	55.2	53.7	-
RoBERTa_{base}	CSC	68.8	-	56.4	63.2
	MUStARD	65.2	53.9	-	59.1
	RedSD (Ours)	73.6	55.8	53.4	-
RoBERTa_{large}	CSC	69.1	-	56.4	64.0
	MUStARD	54.2	49.9	-	59.4
	RedSD (Ours)	72.7	54.0	60.1	-

Table 7: Macro F1 scores (%) of intra- and cross-dataset predictions.

annotator agreement is measured with a Kappa value of 0.634. Interestingly, the human evaluation results are slightly lower than GPT-4-turbo and Claude 3.5 Sonnet. This discrepancy can be attributed to the presence of instances that require specific cultural background knowledge for accurate interpretation, which may not be understood by all human annotators.

5.3 Out-of-Domain Results

To evaluate the generalizability of our dataset, we compare with two existing sarcasm datasets: MUStARD (Castro et al., 2019) and CSC (Jang and Frassinelli, 2024). We conducted both intra-dataset and cross-dataset experiments, as shown in Table 7. As can be seen, models fine-tuned on RedSD (Ours) consistently achieve superior performance in intra-dataset evaluations across all model architectures, despite not being the largest dataset. Moreover, models fine-tuned on RedSD (Ours) outperform those fine-tuned on MUStARD when tested on CSC, and perform comparably to models fine-tuned on CSC when tested on MUStARD. This highlights the effectiveness of our dataset in capturing the complexities of sarcasm.

However, the relatively low cross-dataset performance suggests challenges in transferring across domains. While overfitting cannot be entirely ruled out, part of the performance variation may stem from differences in data format: CSC is not dialogue-based, and instances in MUStARD may rely on multimodal cues for accurate sarcasm detection. This underscores the need for further refinement of our dataset and exploration to enhance cross-domain generalization.

5.4 Rhetorical Devices-Aware Study

To validate the effectiveness of models trained on various rhetorical devices in handling diverse and complex scenarios, we compute accuracy for each rhetorical device using subsets of the RedSD test set, which provides explicit rhetorical device labels to ensure a controlled evaluation. As generative LLMs tend to misclassify non-sarcastic instances as sarcastic, resulting in inflated accuracy across rhetorical devices, we focus our rhetorical device-aware study on encoder-only models. As shown in Table 6, models trained on RedSD (Ours) consistently outperform those trained on CSC or MUStARD across most rhetorical devices, particularly in irony, hyperbole, and rhetorical question. This suggests that our dataset provides a more comprehensive representation of sarcastic expressions, enabling models to better generalize across various rhetorical devices. In addition, echoic mention and presupposition are generally well detected across all models and datasets, whereas unexpected twist and innuendo remain challenging for most models.

Model	w/o DCA	w/o CDA	w/ only irony	Ours
BERT_{base}	74.7	57.8	60.8	74.5
BERT_{large}	73.2	57.0	63.8	73.8
RoBERTa_{base}	73.6	52.8	62.4	75.2
RoBERTa_{large}	72.7	55.3	62.1	73.6

Table 8: Macro F1 scores (%) of ablation study.

5.5 Ablation Study

To verify the effectiveness of our proposed framework, we compare it with the following variants. 1) **w/o DCA**. To evaluate the role of duplex counterfactual augmentation, we discard the duplex-augmented data. 2) **w/o CDA**. To evaluate the role of counterfactual data augmentation, we replace the counterfactual non-sarcastic dialogues generated by CDA with non-sarcastic dialogues sampled from the sitcom corpus. 3) **w/ only irony**. To mimic models trained under the definition of sarcasm as a form of verbal irony and to assess the importance of incorporating various rhetorical devices, we exclude all rhetorical devices except for irony, while keeping the training set size consistent with the above variants. Note that the DCA-generated data is also excluded in variants 2 and 3. All variants are evaluated on the test set of RedSD.

As shown in Table 8, the DCA strategy yields modest yet consistent improvements across most models, indicating that incorporating various rhetorical devices contributes to improved sarcasm detection. Additionally, our framework significantly outperforms both the **w/o CDA** and **w/ only irony** variants across all tested models, achieving average improvements of 17% and 12%, respectively. This result demonstrates that CDA is more effective at capturing the nuances between sarcasm and non-sarcasm, and highlights that the common simplification of sarcasm as a way of verbal irony is insufficient to capture the complexity of sarcasm.

6 Conclusion

In this paper, we study fine-grained sarcasm classification through the lens of rhetorical devices. We introduce a novel sarcasm dataset that incorporates various rhetorical devices, and propose an counterfactual data generation pipeline facilitated by both LLMs and human revision. We conduct a series of experiments with our dataset to benchmark baseline systems, demonstrating the necessity and effectiveness of integrating rhetorical devices for improved sarcasm detection. In summary, our work contributes to paving the way for more nuanced analysis of this intricate linguistic phenomenon.

Limitations and Future Work

We limit the scope of datasets and models to focus on the performance within our dataset, which may restrict generalizability to other domains or real-world contexts. The models discussed in this paper exclude specialized sarcasm detection models. Experiments with other models, datasets, and different hyperparameters are left to future work. We anticipate that our proposed dataset will serve as a valuable resource for advancing research on fine-grained sarcasm detection, particularly in enhancing performance on more challenging rhetorical devices employed in sarcasm. In future work, we plan to incorporate multiple annotators to further improve the robustness and consistency of the annotations, include additional comparisons with recent sarcasm detection models, and develop a larger, more balanced and diverse rhetorical device-aware sarcasm dataset.

References

Gavin Abercrombie and Dirk Hovy. 2016. Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113, Berlin, Germany. Association for Computational Linguistics.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022a. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

Ibrahim Abu Farha, Steven Wilson, Silviu Oprea, and Walid Magdy. 2022b. Sarcasm detection is way too easy! an empirical comparison of human and machine sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5284–5295, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muhammad Abulaish and Ashraf Kamal. 2018. Self-Deprecating Sarcasm Detection: An Amalgamation of Rule-Based and Machine Learning Approach. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 574–579. IEEE. Event-place: Santiago.

Ibtesam AbdulAziz Bajri. 2016. Presupposition and the implicit display theory. *English Linguistics Research*, 5:40.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, 2005.14165.

Elisabeth Camp. 2012. Sarcasm, pretense, and the semantics/ pragmatics distinction . *Noûs*, 46:587–634.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Zixin Chen, Hongzhan Lin, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. 2024. CofiPara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9663–9687, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 392–398, Istanbul, Turkey. European Language Resources Association (ELRA).

H.W. Fowler. 1926. *A Dictionary of Modern English Usage*, 1st edition. Oxford University Press.

Montgomery Gole, Williams-Paul Nwadiugwu, and Andriy Miranskyy. 2023. On sarcasm detection with openai gpt-based models. *Preprint*, arXiv:2312.04642.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin of the New York Botanical Garden*, 2(4):130–156.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.

Mengzhao Jia, Can Xie, and Liqiang Jing. 2024. Debiasing multimodal sarcasm detection with contrastive learning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18354–18362. AAAI Press.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A Large Self-Annotated Corpus for Sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, page 6. European Language Resources Association (ELRA).

Yumin Kim, Heejae Suh, Mingi Kim, Dongyeon Won, and Hwanhee Lee. 2024. KoCoSa: Korean context-aware sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9890–9904, Torino, Italia. ELRA and ICCL.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5956–5968, Dublin, Ireland. Association for Computational Linguistics.

Jiangnan Li, Hongliang Pan, Zheng Lin, Peng Fu, and Weiping Wang. 2021a. Sarcasm detection with commonsense knowledge. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3192–3201.

Jiangnan Li, Hongliang Pan, Zheng Lin, Peng Fu, and Weiping Wang. 2021b. Sarcasm Detection with Commonsense Knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3192–3201.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.

Hui Liu, Wenya Wang, and Haoliang Li. 2022a. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hui Liu, Wenya Wang, and Haoliang Li. 2022b. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yiyi Liu, Ruqing Zhang, Yixing Fan, Jiafeng Guo, and Xueqi Cheng. 2023. Prompt tuning with contradictory intentions for sarcasm recognition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 328–339, Dubrovnik, Croatia. Association for Computational Linguistics.

Changrong Min, Ximing Li, Liang Yang, Zhilin Wang, Bo Xu, and Hongfei Lin. 2023. Just like a human would, direct access to sarcasm augmented with potential result and reaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10172–10183, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *arXiv*, cs.CL/2303.08774.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A Dataset of Intended Sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Silviu Oprea, Steven Wilson, and Walid Magdy. 2021. Chandler: An explainable sarcastic response generator. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 339–349, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2022. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1670–1686. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1601–1612, Osaka, Japan. The COLING 2016 Organizing Committee.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 213–223.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5042–5052. Association for Computational Linguistics.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. 2022. A multimodal corpus for emotion recognition in sarcasm. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6992–7003, Marseille, France. European Language Resources Association.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval*

2014), pages 73–80, Dublin, Ireland. Association for Computational Linguistics.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E. Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3194–3213. Association for Computational Linguistics.

Gopendra Vikram Singh, Mauajama Firdaus, Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2024. Well, now we know! unveiling sarcasm: Initiating and exploring multimodal conversations with reasoning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18981–18989. AAAI Press.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction.

Akira Utsumi. 2000. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Tongshuang Wu, Marco Túlio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6707–6723. Association for Computational Linguistics.

Tan Yue, Xuzhao Shi, Rui Mao, Zonghai Hu, and Erik Cambria. 2024. SarcNet: A Multilingual Multimodal Sarcasm Detection Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14325–14335, Torino, Italy. ELRA and ICCL.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

A Implementation Details

In our experiments, we employ GPT-4o (gpt-4o-2024-08-06), GPT-3.5 (gpt-3.5-turbo-0125), Claude 3.5 haiku (Claude-3-5-haiku-20241022), and Claude 3.5 sonnet (Claude-3-5-sonnet-20241022). The total cost of API usage, including both data generation and model evaluation, amounted to approximately $100.

B Prompts

We give the prompts we use in §3.2 and §3.3, as shown in Figure 4 and 5.

C Case Study

To gain an intuitive comprehension of model performance on our dataset, we present examples of errors made by various models in Table 9. As can be seen, fine-tuned models exhibits difficulties in differentiating between genuine metaphorical expressions and sarcastic ones, subsequently leading to an erroneous prediction. Moreover, GPT-4 struggle to detect sarcasm in seemingly complimentary statements that imply criticism ("she never complained about you once. I know what kind of strength that takes."), or responses that introduce incongruous elements to neutral questions (keeping the mother out of a new car to preserve its smell). Interestingly, LLaMA 3.1 successfully identifies the implied sarcasm in above examples. However, it shows a tendency to be overly sensitive to certain linguistic patterns commonly associated with sarcasm (e.g., phrases like "just like you" or "I'm sure that's not true"), which leads to numerous false positives, as the model fails to adequately consider the overall context and intent of the dialogue. These findings underscore the complexity of sarcasm detection and the need for models to not only process linguistic cues but also to comprehend broader contextual and pragmatic aspects of communication for more accurate interpretation.

Prompt 1

Task Description: Given a [**sarcastic dialogue**] from the sitcom "The Big Bang Theory", along with the corresponding [**sarcastic segment**] and [**rhetorical device**], your task is to write a counterfactual dialogue that does not contain any sarcasm.
Here are a few points to keep in mind:
1. Interpret the speaker's true intention and provide a succinct explanation based on the world knowledge.
2. Make minimal changes to remove the sarcasm while retaining the original meaning.

Example 1
Input:
sarcastic dialogue:
A: *I wrote a fan letter to you when I was a child in Texas, and you sent this autographed picture back to me. Do you remember that?*
B: *I'll give you a hint: I have a bracelet with my own address on it.*
sarcastic segment: *I’ ll give you a hint: I have a bracelet with my own address on it.*
rhetorical device: presupposition and innuendo
Output:
intent: conveying that the speaker does not remember individual fan letters.
explanation: Having a bracelet with one's own address is often associated with memory issues, implying that remembering a specific fan letter is unlikely.
Counterfactual:
A: *I wrote a fan letter to you when I was a child in Texas, and you sent this autographed picture back to me. Do you remember that?*
B: *No, I’ m sorry, I don’ t remember any individual fan letters.*

Example 2
...

Figure 4: Prompt used to generate counterfactuals in §3.2.

Prompt 1

Task Description: Given a sarcastic dialogue A from the sitcom "The Big Bang Theory" and its corresponding non-sarcastic counterfactual dialogue B, your task is twofold:
1. Rewrite the Sarcastic Dialogue into Non-Sarcastic Dialogue: Analyze the given two dialogues to understand the nuances between sarcasm and non-sarcasm and use this understanding to rewrite dialogue A into a straightforward, non-sarcastic version.
2. Rewrite the Non-Sarcastic Dialogue into Sarcastic Dialogue: Select an appropriate target (a person, object, or situation) for the sarcasm and use the following specified rhetorical device to express sarcasm. Make sure the rewritten dialogue doesn't contain irony if the specified rhetorical device is not irony.

Rhetorical device: [**rhetorical device**]
Here is an example:
Input:
Original sarcastic dialogue:
A: Can you tell I'm perspiring a little?
B: No. The dark crescent shaped patterns under your arms conceal it nicely.
Original non-sarcastic dialogue:
A: Can you tell I'm perspiring a little?
B: Yes. Your armpits are completely soaked.
Output:
Rewritten non-sarcastic dialogue:
A: Can you tell I'm perspiring a little?
B: Yes, I can see you're sweating a lot under your arms.
Rewritten sarcastic dialogue:
A: Can you tell I'm perspiring a little?
B: If by 'a little', you mean your armpits could water a garden, then yes.
Here are a few points to keep in mind:
1. You must keep the rewritten dialogues logically coherent.
2. You must use the specified rhetorical device to generate sarcastic dialogue.

Now, please process the following input:

Figure 5: Prompt used to generate counterfactuals in §3.3.

Models	Dialogues	Labels	Predictions
FT models	"I don't need sleep. I need answers. I need to determine where in this swamp of unbalanced formulas squatteth the toad of truth.", "Toad of truth? Is that a physics thing?", "No, it's more of a metaphorical concept."	0	1
	"Well, all these years, I was afraid to say what I wanted. You know, even at work, you know, there's things I want to accomplish, but I didn't want to ruffle any feathers or step on any toes.", "Feathers and toes? Is the new thing you're trying to accomplish ballroom dancing with a chicken?"	1	0
GPT-4	"Honestly, of all of my children's spouses, she's the one that I'm most impressed by.", "Seriously?", "Yes. She's confident, she's thoughtful, and she never complained about you once. I know what kind of strength that takes."	1	0
	"Guess who picked up his new car this morning?", "Congratulations. Does it have that new car smell?", "Yep! For as long as I can keep my mother out of it."	1	0
LLaMA 3.1	"I'd like to know why Penny's here.", "I'm here to support my man, just like you.", "What are you going to do?"	0	1
	"I know, but on the other hand, do you really care?", "Yes, I care. This happens to me all the time. People take one look at me and assume I don't know what I'm talking about.", "Oh, I'm sure that's not true.", "I'm genuinely asking. Do you think I lack knowledge and don't know what I'm talking about?"	0	1

Table 9: The error examples made uniformly across all fine-tuned (FT) models (including BERT$_{base}$, BERT$_{large}$, and RoBERTa$_{base}$), GPT-4 and LLaMA 3.1 (spanning the 8B, 70B and 405B versions). Red spans represent sarcastic segments that models fail to recognize and blue spans represent misidentified sarcastic segments in non-sarcastic dialogues.