Rhetoric-Aware Counterfactual Reasoning for Fine-Grained Sarcasm Understanding

Anonymous ACL submission

Abstract

Sarcasm understanding requires modeling implicit meaning beyond surface sentiment. Existing methods frame sarcasm detection as coarse-grained classification, overlooking fine-grained rhetorical structures and compositional mechanisms. We present the first study of fine-grained rhetorical sarcasm classification. We propose **RACR**, a Rhetoric-Aware Counterfactual Reasoning framework that constructs multi-level counterfactuals and evaluates rhetorical hypotheses via semantic inconsistencies. RACR produces structured reasoning signals for downstream LLM inference. Experiments on fine-grained rhetorical sarcasm categories show that RACR outperforms supervised and prompting-based baselines. Further analysis reveals systematic LLM biases toward certain rhetorical patterns, which RACR mitigates via counterfactual reasoning, yielding more balanced and interpretable predictions.

1 Introduction

Sarcasm conveys meaning through a mismatch between surface form and intended semantics, requiring reasoning beyond literal sentiment. Most prior work models sarcasm as a coarse-grained classification problem (Khatri and P, 2020; He and Abisado, 2024; Babu et al., 2025), while overlooking diverse rhetorical structures such as irony, hyperbole, and rhetorical questions (Hong et al., 2025). These rhetorical mechanisms may co-occur within a single utterance (Table 1), increasing the complexity of sarcasm understanding.

Recent large language models rely on explicit surface cues (Zhang et al., 2025b; Bojić et al., 2025; Zhang et al., 2025a), leading to systematic biases across rhetorical categories and limited modeling of structurally implicit mechanisms. This suggests the need for reasoning beyond one-shot prediction.

Human interpretation typically follows a coarse-to-fine process: multiple plausible interpretations are first considered and then refined toward the most explanatory mechanism. This process aligns naturally with counterfactual reasoning (Song et al., 2023; Lemberger and Saillenfest, 2024; Chatzi et al., 2024), where alternative expressions are evaluated to reveal semantic inconsistencies.

Device	Definition
Irony	Surface meaning opposite to intent
Echoic Mention	Mocking repetition of prior utterance
Hyperbole	Extreme exaggeration
Rhetorical Question	Question implying assertion
Presupposition	Implicit shared assumption
Innuendo	Indirect negative implication
Reenactment	Exaggerated replay of events
Twist	Sudden narrative contrast
Self-deprecation	Self-mockery targeting others

Table 1: Fine-grained rhetorical devices in RedSD.

Hence, we propose **Rhetoric-Aware Counterfactual Reasoning (RACR)**, a two-stage framework for fine-grained sarcasm understanding. Stage 1 constructs multi-level counterfactual variants for coarse candidate selection, while Stage 2 performs causal verification and rule-guided refinement. RACR produces structured reasoning signals that improve interpretability and robustness.

To the best of our knowledge, this is the first work to investigate fine-grained rhetorical sarcasm classification via structured counterfactual reasoning. While (Hong et al., 2025) leverages counterfactual data augmentation for coarse-grained sarcasm detection, our approach differs in that it models fine-grained rhetorical categories through structured counterfactual reasoning with causal verification and rule-guided calibration, rather than using counterfactuals solely for data augmentation. Experiments on the RedSD dataset show that RACR improves classification performance while reducing rhetorical bias in LLM predictions.

Our main contributions are as follows:

- We introduce the task of fine-grained rhetorical sarcasm classification beyond coarse-grained sarcasm detection.
- We propose RACR, a two-stage framework integrating structured counterfactual reasoning, causal verification, and rule-guided calibration.
- We demonstrate that RACR improves performance and reduces rhetorical bias in LLM-based sarcasm understanding.

2 Related Work

Sarcasm Detection. Sarcasm detection is challenging due to its implicit and context-dependent nature. Early work formulates it as binary classification using feature-based or neural models (Khatri and P, 2020; Sharma et al., 2023; Kavitha

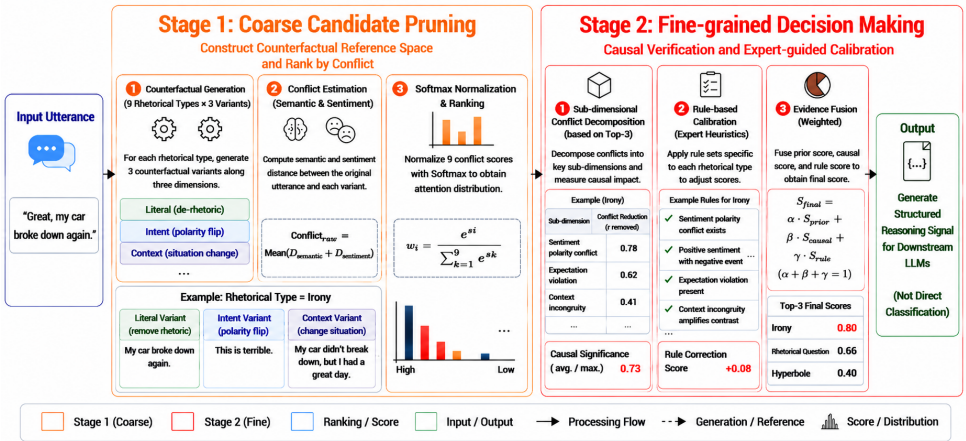

Figure 1: Overview of RACR. Stage 1 constructs counterfactual variants for coarse candidate selection, while Stage 2 performs causal verification and rule-guided refinement.

and Chittieni, 2022; He and Abisado, 2024; Šandor and Bagić Babac, 2024; Uma Maheswari and Dhenakaran, 2024; Verma et al., 2025), while recent approaches leverage contextual and emotional signals with Transformer-based architectures (Babu et al., 2025). However, most methods remain coarse-grained and overlook rhetorical compositionality (Wang et al., 2024; Yacoub et al., 2024). Recent work explores finer signals such as emotional incongruity and rhetorical patterns (Qiu et al., 2025; Zhang et al., 2025b), but lacks explicit modeling of diverse rhetorical mechanisms.

Counterfactual Reasoning. Counterfactual reasoning improves robustness and interpretability in NLP, with applications in data augmentation (Song et al., 2023), model explanation (Lemberger and Saillenfest, 2024), generation (Chatzi et al., 2024), and reasoning or emotion inference (Lewis and Mitchell, 2024; Wang et al., 2025; Li et al., 2021; Vu and Nguyen, 2025). Its use in sarcasm understanding remains limited, especially for fine-grained rhetorical modeling.

Datasets and Evaluation. Datasets such as RedSD (Hong et al., 2025) and benchmarks for figurative and implicit language understanding (Zheng et al., 2025) provide annotations of rhetorical devices. However, existing studies mainly focus on polarity-level or coarse-grained tasks, without addressing fine-grained rhetorical classification.

3 Method

3.1 Overview

We propose RACR, a two-stage framework for fine-grained sarcasm understanding via counterfactual reasoning. Given an utterance x, RACR constructs counterfactual variants and evaluates rhetorical hypotheses via semantic inconsistencies, following prior work (Song et al., 2023; Lemberger and Saillenfest, 2024; Chatzi et al., 2024). It produces structured reasoning signals for downstream inference. RACR consists of (1) counterfactual view construction with candidate selection, and (2) causal verification with rule-guided refinement.

3.2 Counterfactual View Construction

RACR constructs rhetoric-aware counterfactual views by prompting an LLM to generate three complementary variants for each rhetorical hypothesis $r \in \mathcal{R}$:

$$\mathcal{V}_r(x) = \{x_r^{lit}, x_r^{int}, x_r^{ctx}\} \tag{1}$$

where x_r^{lit} removes rhetorical effects specific to r, x_r^{int} neutralizes sentiment polarity under hypothesis r, and x_r^{ctx} modifies contextual assumptions associated with r (Song et al., 2023).

3.3 Stage I: Coarse Candidate Selection

RACR estimates rhetorical inconsistency via semantic and sentiment divergence under rhetoric-specific counterfactual interventions. For each rhetorical hypothesis $r \in \mathcal{R}$, a conflict score is computed as:

$$\text{Conflict}_r(x) = \sum_{v \in \mathcal{V}_r(x)} (\lambda_1 d_{\text{sem}}(x, v) + \lambda_2 d_{\text{sent}}(x, v)) \tag{2}$$

We set $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$ based on pilot experiments. Semantic divergence is defined as the cosine distance between Sentence-BERT (all-mpnet-base-v2) embeddings.

$$d_{\text{sem}}(x, v) = 1 - \cos(E(x), E(v)) \tag{3}$$

Sentiment divergence is computed as the absolute difference between LLM-predicted sentiment

scores $S(x) \in [-1, 1]$:

$$d_{\text{sent}}(x, v) = |S(x) - S(v)| \quad (4)$$

Conflict scores are normalized via a softmax to obtain rhetorical probabilities:

$$P(r|x) = \frac{\exp(-Conflict_r(x))}{\sum_{r'} \exp(-Conflict_{r'}(x))} \quad (5)$$

Finally, RACR selects the Top-3 rhetorical candidates for Stage II refinement.

3.4 Stage II: Causal Verification and Rule-guided Refinement

Given the Top-3 candidates from Stage I, we refine them via causal verification and rule-guided calibration (Wang et al., 2025).

Causal verification. We decompose the utterance x into key rhetorical spans using spaCy dependency parsing and type-specific heuristics (e.g., presupposition triggers, extreme quantifiers). For each component k, we construct $x \setminus k$ by masking the span and using an LLM to generate a fluent version. The contribution of each component is:

$$\Delta_k^{(r)} = \text{Conflict}_r(x) - \text{Conflict}_r(x \setminus k), \quad (6)$$

and the causal score is $S_{\text{causal}}(r) = \sum_k \Delta_k^{(r)}$. Positive $\Delta_k^{(r)}$ indicates higher importance of component k for hypothesis r.

Rule-guided calibration. We incorporate linguistic priors as soft constraints. Specifically, we check polarity reversal for Irony, interrogative markers for Rhetorical Question, extreme quantifiers for Hyperbole, presupposition triggers for Presupposition, and contrast markers for Twist. For other categories, we use analogous feature-matching rules. We compute a rule-matching score $S_{\text{rule}}(r) \in [0, 1]$ for each candidate.

The final score is a weighted combination:

$$S_{\text{final}}(r) = \alpha S_{\text{causal}}(r) + (1 - \alpha) S_{\text{rule}}(r) \quad (7)$$

We set $\alpha = 0.6$ based on pilot experiments.

3.5 Evidence Fusion and Structured Reasoning Signal

RACR fuses the above results into a structured reasoning signal z, which includes the Top-3 ranked candidates with their S_{final} scores, key semantic conflicts, causal attribution scores $\Delta_k^{(r)}$, and activated rules. This signal z is provided as auxiliary context to downstream LLMs, enabling more informed and interpretable sarcasm understanding (Wei et al., 2022; Xu et al., 2025).

4 Experiments

4.1 Experimental Setup

We evaluate RACR on the sarcastic subset of RedSD, a rhetorical device-aware benchmark containing 1,018 manually annotated sarcastic dialogues across nine rhetorical categories, where each instance may contain up to two labels to capture compositional sarcasm. RACR generates structured counterfactual reasoning signals and injects them into four representative downstream LLMs: LLaMA-3-8B-Instruct, GPT-4o, DeepSeek-V3, and Qwen-2.5.

As fine-grained rhetorical sarcasm classification remains underexplored, no established supervised SOTA systems currently exist. We therefore compare RACR against both supervised and prompting-based baselines, including BERT-large, RoBERTa-large, standard zero-shot prompting, few-shot prompting, and counterfactual CoT prompting. Final results are reported as the average performance across five independent inference runs.

We evaluate performance using three protocols: Full Match (exact label set match), Top-1 (Any Match) (top prediction matches any ground-truth label), and Top-2 (Any Match) (any of the top-2 predictions matches the ground truth).

4.2 Main Results

Table 2 presents the overall results. RACR consistently outperforms all baselines across evaluation protocols, with the largest improvements under Full Match, indicating stronger modeling of complete rhetorical structures. Under Top-2 (Any Match), RACR achieves the best F1 score (0.52), demonstrating improved coverage of multiple plausible rhetorical interpretations under compositional sarcasm reasoning. Notably, supervised pretrained classifiers such as BERT-large and RoBERTa-large perform worse than zero-shot LLM prompting, suggesting that fine-grained rhetorical sarcasm understanding relies more heavily on implicit reasoning and pragmatic knowledge than conventional discriminative classification.

4.3 Category-wise Analysis

Table 3 presents a fine-grained comparison between CoT prompting and RACR under the Top-2 (Any-match) protocol.

We select CoT as the primary baseline because it achieves the strongest overall performance among all baselines,while explicitly incorporating reasoning before prediction. Unlike conventional CoT, which relies on free-form natural language reason-

Method	Full Match				Top-1 (Any Match)				Top-2 (Any Match)			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
BERT-large	0.05	0.24	0.30	0.26±0.02	0.37	0.35	0.27	0.28±0.02	0.53	0.36	0.42	0.34±0.02
RoBERTa-large	0.06	0.26	0.32	0.28±0.02	0.39	0.37	0.28	0.29±0.02	0.55	0.38	0.44	0.36±0.02
Zero-shot LLaMA-3-8B-Instruct	0.07	0.28	0.36	0.31±0.02	0.43	0.41	0.31	0.32±0.02	0.58	0.40	0.47	0.38±0.02
Zero-shot GPT-4o	0.08	0.30	0.39	0.32±0.02	0.46	0.42	0.31	0.34±0.01	0.60	0.41	0.46	0.41±0.02
Zero-shot DeepSeek-V3	0.07	0.29	0.35	0.31±0.02	0.40	0.41	0.33	0.32±0.01	0.60	0.42	0.48	0.39±0.03
Zero-shot Qwen-2.5	0.06	0.27	0.34	0.29±0.02	0.42	0.40	0.30	0.33±0.02	0.59	0.39	0.48	0.38±0.03
Few-shot LLaMA-3-8B-Instruct	0.08	0.29	0.36	0.31±0.02	0.42	0.40	0.32	0.31±0.02	0.58	0.41	0.48	0.39±0.03
Few-shot GPT-4o	0.10	0.31	0.39	0.33±0.01	0.47	0.42	0.32	0.34±0.02	0.61	0.42	0.48	0.43±0.02
Few-shot DeepSeek-V3	0.08	0.29	0.36	0.31±0.02	0.42	0.42	0.34	0.33±0.02	0.60	0.42	0.48	0.39±0.03
Few-shot Qwen-2.5	0.08	0.29	0.35	0.30±0.02	0.43	0.41	0.31	0.33±0.02	0.60	0.41	0.49	0.39±0.03
CoT LLaMA-3-8B-Instruct	0.11	0.33	0.39	0.32±0.02	0.43	0.42	0.33	0.33±0.02	0.59	0.41	0.48	0.39±0.03
CoT GPT-4o	0.12	0.32	0.40	0.33±0.01	0.47	0.46	0.32	0.35±0.02	0.62	0.44	0.49	0.44±0.02
CoT DeepSeek-V3	0.12	0.30	0.41	0.32±0.03	0.42	0.42	0.35	0.33±0.02	0.60	0.43	0.48	0.40±0.03
CoT Qwen-2.5	0.11	0.31	0.35	0.31±0.02	0.45	0.43	0.32	0.35±0.03	0.60	0.42	0.50	0.40±0.03
RACR (Ours)	**0.19**	**0.41**	**0.48**	**0.37±0.02**	**0.52**	**0.50**	**0.38**	**0.40±0.01**	**0.65**	**0.49**	**0.53**	**0.52±0.02**

Table 2: Overall performance comparison under different evaluation settings. Precision, Recall, and F1 are macro-averaged across the nine rhetorical categories. Statistical significance is tested using paired t-tests ($p < 0.03$).

ing, RACR performs structured counterfactual reasoning through rhetorical conflict modeling and causal verification.

We adopt Top-2 evaluation because sarcasm in RedSD is inherently compositional and often involves overlapping rhetorical mechanisms. Prior linguistic studies suggest that sarcastic utterances are frequently multi-interpretable rather than associated with a single deterministic meaning (Camp, 2012; Utsumi, 2000). Therefore, Top-2 evaluation better measures whether models can preserve multiple plausible rhetorical interpretations during compositional reasoning.

Rhetorical Type	COT-average			RACR			Δ F1
	P	R	F1	P	R	F1	
Irony	0.42	0.48	0.43	0.50	0.56	0.52	+0.09
Echoic Mention	0.38	0.44	0.40	0.47	0.52	0.48	+0.08
Hyperbole	0.41	0.46	0.42	0.49	0.54	0.50	+0.08
Rhetorical Question	0.36	0.42	0.38	0.46	0.51	0.47	+0.09
Presupposition	0.35	0.40	0.37	0.44	0.49	0.45	+0.08
Innuendo	0.33	0.39	0.35	0.43	0.48	0.44	+0.09
Reenactment	0.34	0.41	0.36	0.45	0.50	0.46	+0.10
Twist	0.37	0.43	0.39	0.48	0.53	0.49	+0.10
Self-deprecation	0.39	0.45	0.41	0.47	0.52	0.48	+0.07

Table 3: Comparison across rhetorical categories.

RACR improves performance across all categories, with larger gains on structurally implicit types such as *Reenactment* and *Twist*. LLM baselines tend to over-predict frequent categories while underperforming on implicit rhetorical mechanisms, suggesting an over-reliance on surface cues. RACR mitigates these biases through counterfactual verification and conflict-based reasoning.

4.4 Ablation Study

Table 4 reports ablation results under the same Top-2 evaluation setting.

Method	Acc	P	R	F1
RACR	**0.65**	**0.49**	**0.53**	**0.52±0.02**
w/o Counterfactual Views	0.58	0.42	0.48	0.41±0.02
w/o Causal Reasoning	0.61	0.46	0.50	0.46±0.01
w/o Rule Calibration	0.60	0.44	0.49	0.45±0.02

Table 4: Ablation study under Top-2 (Any-match).

Removing counterfactual views causes the largest drop, highlighting their importance. Causal reasoning and rule calibration provide complementary gains, indicating that both fine-grained reasoning and linguistic priors are necessary.

5 Conclusion

We propose RACR, a two-stage Rhetoric-Aware Counterfactual Reasoning framework for fine-grained sarcasm understanding. To the best of our knowledge, this is the first work to investigate fine-grained classification of sarcasm. RACR models sarcasm as progressive reasoning by constructing multi-level counterfactuals and refining predictions via causal verification and rule-guided calibration.

RACR consistently outperforms strong baselines across evaluation protocols and rhetorical categories, with notable gains on complex types such as presupposition and reenactment. It also mitigates systematic LLM biases, including category collapse and over-prediction of frequent classes.

RACR follows a coarse-to-fine reasoning paradigm that generates alternative interpretations via counterfactuals and validates them through consistency and conflict analysis. This *generate-then-verify* process aligns with human-like interpretation and shifts sarcasm understanding from surface cue matching to causal pragmatic reasoning.

Limitations

RACR is currently evaluated in a lightweight inference setting using representative open- and closed-source LLMs, without exploring larger-scale or frontier-level reasoning models. While this setting better reflects practical and cost-efficient deployment scenarios, future work may investigate whether stronger foundation models further enhance structured counterfactual reasoning.

In addition, our experiments are limited to the English RedSD benchmark. Since rhetorical sarcasm is closely tied to linguistic conventions and cultural context, the effectiveness of RACR across multilingual or cross-cultural settings remains underexplored.

Finally, although RACR introduces structured counterfactual reasoning, the framework still depends on LLM-generated intermediate reasoning signals, which may occasionally propagate inconsistencies or unstable interpretations from the underlying models.

References

Suryavamshi Sandeep Babu, S. V. Suryanarayana, M. Sruthi, P. BhagyaLakshmi, T. Sravanthi, and M. Spandana. 2025. Enhancing sentiment analysis with emotion and sarcasm detection: A Transformer-based approach. *Metallurgical and Materials Engineering*, 31(5):794–803.

Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaitė, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific Reports*, 15:11477.

Elisabeth Camp. 2012. Sarcasm, pretense, and the semantics/pragmatics distinction. *Noûs*, 46(4):587–634.

Ivi Chatzi, Nina Corvelo Benz, Eleni Straitouri, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. 2024. Counterfactual token generation in large language models. *Preprint*, arXiv:2409.17027.

Aixiang He and Mideth Abisado. 2024. Text sentiment analysis of douban film short comments based on BERT-CNN-BiLSTM-Att model. *IEEE Access*, 12:45229–45237.

Qingqing Hong, Dongyu Zhang, Jiayi Lin, Dapeng Yin, Shuyue Zhu, and Junli Wang. 2025. Rhetorical device-aware sarcasm detection with counterfactual data augmentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11672–11685.

K. Kavitha and Suneetha Chittieni. 2022. An intelligent metaheuristic optimization with deep convolutional recurrent neural network enabled sarcasm detection and classification model. *International Journal of Advanced Computer Science and Applications*, 13(2):304–314.

Akshay Khatri and Pranav P. 2020. Sarcasm detection in tweets with bert and glove embeddings. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 56–60.

Pirmin Lemberger and Antoine Saillenfest. 2024. Explaining text classifiers with counterfactual representations. *Preprint*, arXiv:2402.00711.

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *Preprint*, arXiv:2402.08955.

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021. Emotion inference in multi-turn conversations with addressee-aware module and ensemble strategy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3935–3941.

Ziqi Qiu, Jianxing Yu, Yufeng Zhang, Hanjiang Lai, Yanghui Rao, Qinliang Su, and Jian Yin. 2025. Detecting emotional incongruity of sarcasm by commonsense reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9062–9073, Abu Dhabi, UAE. Association for Computational Linguistics.

Daniel Šandor and Marina Bagić Babac. 2024. Sarcasm detection in online comments using machine learning. *Information Discovery and Delivery*, 52(2):213–226.

Dilip Kumar Sharma, Bhuvanesh Singh, Saurabh Agarwal, Nikhil Pachauri, Amel Ali Alhussan, and Hanaa A. Abdallah. 2023. Sarcasm detection over social media platforms using hybrid ensemble model with fuzzy logic. *Electronics*, 12(4):937.

Rui Song, Fausto Giunchiglia, Yingji Li, and Hao Xu. 2023. Automatic counterfactual augmentation for robust text classification based on word-group search. *Preprint*, arXiv:2307.01214.

S. Uma Maheswari and S. S. Dhenakaran. 2024. Improved ensemble based deep learning approach for sarcastic opinion classification. *Multimedia Tools and Applications*, 83:38267–38289.

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

Deval Verma, Pranay Kumar, Manish Agarwal, Gaurav Verma, Sujanya Sombatteera, and Sujin Butdisuwan. 2025. Comparative analysis of different models for sarcastic and non-sarcastic text detection using nlp. In *Lecture Notes in Networks and Systems*, volume 1164, pages 671–682.

Hoang-Long Vu and Cam-Van Thi Nguyen. 2025. Integrating cause-awareness for improving emotional support in conversational ai. In *2025 17th International Conference on Knowledge and System Engineering (KSE)*.

Xinzhi Wang, Jiayan Qian, Yudong Chang, Hang Yu, and Hui Zhang. 2025. Emotion inference of text based on counterfactual behavior knowledge. *Information Fusion*, 120:103060.

Yongjie Wang, Xiaoqi Qiu, Yue Yu, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A survey on natural language counterfactual generation. *Preprint*, arXiv:2407.03993.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Xingle Xu, Shi Feng, Daling Wang, Yifei Zhang, and Xiaocui Yang. 2025. Enhancing zero-shot emotion perception in conversation through the internal-to-external chain-of-thought. In *DAAP 2024*.

Ahmed Derbala Yacoub, Salwa O. Slim, and Amal Elsayed Aboutabl. 2024. A survey of sentiment analysis and sarcasm detection: Challenges, techniques, and trends. *International Journal of Electrical and Computer Engineering Systems*, 15(1):69–78.

Xinyu Zhang, Long Zhang, Qiusheng Zheng, and Yanli Liang. 2025a. Identification of chinese internet neologism and implicit sentiment analysis based on cot. In *2025 5th International Conference on Communication Technology and Information Technology (ICCTIT)*. IEEE.

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2025b. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *IEEE Transactions on Affective Computing*, 16(4):2560–2576.

Yilun Zheng, Sha Li, Fangkun Wu, Ziyi Yang, Hongchao Lin, Zhichao Hu, Xinjun Cai, Ziming Wang, Jinxuan Chen, Sitao Luan, Jiahao Xu, and Lihui Chen. 2025. Fanchuan: A multilingual and graph-structured benchmark for parody detection and analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*.

A Case Study and Analysis

We compare RACR with Chain-of-Thought (CoT) prompting, which also introduces intermediate reasoning steps and can implicitly reflect counterfactual considerations. While CoT improves reasoning transparency, it does not explicitly model hypothesis-conditioned counterfactual interventions.

A.1 Presupposition vs. Irony

Dialogue:
A: "Is this the high-IQ sperm bank?"
B: "If you have to ask, maybe you shouldn't be here."
Prediction:
CoT LLM: *Irony* (incorrect)
RACR: *Presupposition* (correct)

CoT reasoning identifies surface-level pragmatic tension and often attributes the response to ironic reversal. Although CoT may implicitly consider alternative interpretations, it does not systematically test whether removing the presupposition resolves the inconsistency.

RACR instead explicitly constructs counterfactual variants that neutralize the implicit assumption that "intelligent users should not ask such a question." Under these interventions, the pragmatic conflict disappears, indicating that the rhetorical force originates from presupposition violation rather than polarity inversion.

A.2 Hyperbole vs. Echoic Mention

Dialogue:
A: "So Penny's a little messy."
B: "A little messy? The Mandelbrot set of complex numbers is a little messy. This is chaos."
Prediction:
CoT LLM: *Echoic Mention* (incorrect)
RACR: *Hyperbole* (correct)

CoT reasoning tends to focus on lexical repetition and analogy interpretation, which leads to over-attribution of echoic structure. While CoT may decompose reasoning steps, it lacks explicit perturbation of the cross-domain exaggeration mechanism.

RACR performs counterfactual replacement that maps extreme comparisons into neutral intensity expressions, revealing that the rhetorical effect is driven by semantic amplification across domains rather than quotation-like reuse.

A.3 Presupposition vs. Twist

Dialogue:
A: "Leonard was telling me about dark matter, but I didn't really understand it."
B: "Don't feel bad. Neither does he."
Prediction:
CoT LLM: *Twist* (incorrect)
RACR: *Presupposition* (correct)

CoT reasoning captures the apparent narrative reversal and concludes a twist-based interpretation. However, even when reasoning is explicitly articulated, CoT does not explicitly verify whether underlying assumptions hold under counterfactual removal.

RACR instead tests the implicit assumption that an explainer should understand the topic. When this assumption is removed via counterfactual intervention, the apparent twist structure collapses, indicating that the humor arises from presupposition failure.

A.4 Irony vs. Twist

Dialogue:
A: "I'm not angry at all that my boyfriend called you instead of me. I love that!"
Prediction:
CoT LLM: *Twist* (incorrect)
RACR: *Irony* (correct)

CoT reasoning improves interpretability but remains sensitive to abrupt sentiment transitions, often over-predicting twist-like structures. Even with step-by-step reasoning, it does not explicitly isolate polarity conflict as a causal factor.

RACR explicitly evaluates sentiment consistency under counterfactual perturbations, identifying a stable polarity contradiction between contextual expectation and literal expression, which is characteristic of irony.

Summary These cases show that while Chain-of-Thought improves intermediate reasoning, it remains grounded in natural language explanation rather than explicit counterfactual intervention. In contrast, RACR performs structured hypothesis-conditioned perturbations, enabling finer-grained identification of rhetorical mechanisms.

B Counterfactual Examples

This section presents the counterfactual generation procedure used in RACR. We first define nine categories of rhetorical phenomena and their corresponding transformation strategies. Then, we provide a single input example and show how each category produces distinct counterfactual variants.

B.1 Rhetorical Phenomena and Transformation Strategies

1. Irony **Core characteristic:** Surface expression contradicts intended meaning.

- **Literal inversion:** Replace evaluative words with antonyms to align surface meaning with literal intent.
- **Intent alignment:** Rewrite into a direct non-ironic statement expressing the underlying meaning.
- **Context strengthening:** Add explicit negation cues (e.g., "actually not") to remove ironic interpretation.

2. Echoic Mention **Core characteristic:** Repetition of prior utterances with implicit evaluation.

- **Remove repetition:** Eliminate or rewrite repeated segments to form an original statement.
- **Explicit intent:** Convert implied sarcasm into explicit agreement or disagreement.
- **Context neutralization:** Remove quotation markers or add affirming context.

3. Hyperbole **Core characteristic:** Extreme exaggeration for emphasis.

- **Literal reduction:** Replace exaggerated expressions with mild descriptions.
- **Intent normalization:** Reformulate into literal statements without amplification.
- **Context balancing:** Add quantifiers such as "somewhat" or "approximately".

4. Rhetorical Question **Core characteristic:** Questions used to express assertions.

- **Declarativization:** Convert interrogative forms into declarative statements.
- **Explicit intent:** Directly state the implied evaluative meaning.
- **Response grounding:** Add explicit answer structure to remove rhetorical reading.

5. Presupposition **Core characteristic:** Implicit assumptions embedded in utterances.

- **Explicitization:** Make implicit assumptions explicit.
- **Assumption removal:** Rewrite to eliminate dependency on shared assumptions.
- **Contradiction injection:** Introduce alternative or negated presuppositions.

6. Innuendo **Core characteristic:** Indirect implication of negative meaning.

- **Direct expression:** Replace implicit hints with explicit statements.
- **Transparency shift:** Remove hidden meaning layers.
- **Context clarification:** Add neutralizing context to eliminate implication.

7. Reenactment **Core characteristic:** Dramatic reconstruction of past events.

- **Compression:** Reduce reenactment into factual summary.
- **Neutral narration:** Rewrite as objective description.
- **Abstraction:** Remove dialogue mimicry and temporal dramatization.

8. Twist **Core characteristic:** Sudden narrative reversal.

- **Linearity enforcement:** Remove contrast markers such as "but".
- **Early revelation:** Anticipate outcome in advance.
- **Smooth transition:** Remove abrupt shifts in meaning.

9. Self-deprecation **Core characteristic:** Self-directed criticism masking external critique.

- **Self-neutralization:** Replace self-deprecation with neutral/self-positive expressions.
- **Target reorientation:** Redirect evaluation toward external entities.
- **Self-removal:** Eliminate self-reference to expose true target.

B.2 Unified Example

Original input x:

"Is this the high-IQ sperm bank? If you have to ask, maybe you shouldn't be here."

B.3 Counterfactual Variants Across Nine Categories

Irony

- "This is a place for highly intelligent individuals."
- "This place is not suitable for you."
- "This is a place with specific entry requirements, and asking questions is normal."

Echoic Mention

- "This is a high-IQ sperm bank."
- "The statement reflects exclusionary judgment."
- "This is a facility with defined eligibility criteria."

Hyperbole

- "This is an extremely exclusive place."
- "This place has strict entry requirements."
- "This place has certain admission standards."

Rhetorical Question

- "This does not seem appropriate."
- "This situation is inappropriate."
- "Is this an appropriate place?" (literal question)

Presupposition

- "This place is for highly intelligent people."
- "There is no confirmed intelligence requirement."
- "This place has unclear admission criteria."

Innuendo

- "You do not meet the expectations of this place."
- "This interaction contains implicit judgment."
- "This is a structured institutional environment."

Reenactment

- "The conversation involved questioning access suitability."
- "The interaction was confrontational."
- "A conversation occurred regarding entry conditions."

Twist

- "This appears to be a standard facility."
- "The situation is more complex than expected."
- "This is a facility with defined rules."

Self-deprecation

- "I misunderstood the nature of this place."
- "The situation reflects exclusionary behavior."
- "There was confusion about the nature of this place."

B.4 Summary

This framework demonstrates how a single input x can be systematically transformed into nine types of counterfactual variants. Each category isolates a specific rhetorical mechanism, enabling fine-grained causal analysis of pragmatic and rhetorical phenomena in RACR.

C Prompt Templates

This section describes prompt templates used in few-shot, Chain-of-Thought (CoT), and LLM-based evaluation settings.

C.1 Few-shot Prompt

We construct few-shot examples from representative dataset instances, ensuring coverage of diverse rhetorical mechanisms and preserving dialogue-level context.

> You are a rhetorical classification system. Given a dialogue, identify the dominant rhetorical device(s) from the predefined taxonomy.
>
> **Rhetorical Labels and Definitions:**
>
> - **Irony**: Surface meaning opposite to intent.
> - **Echoic Mention**: Mocking repetition of a prior utterance.

- **Hyperbole**: Extreme exaggeration.
- **Rhetorical Question**: A question implying an assertion.
- **Presupposition**: Implicit shared assumption.
- **Innuendo**: Indirect negative implication.
- **Reenactment**: Exaggerated replay of events.
- **Twist**: Sudden narrative contrast or reversal.
- **Self-deprecation**: Self-mockery targeting oneself.

Example 1
Context: "What are you, a little kid? Is she gonna cut your dinner into little pieces, too?"
Label: rhetorical question, hyperbole

Example 2
Context: "I'm gonna miss her." (after Raj breaks Emily's drawer)
Label: presupposition, innuendo

Example 3
Context: "You could barely survive a tiny turtle bite."
Label: intentional reenactment

Now classify: Input: "INPUT" Output:

C.2 Chain-of-Thought Prompting Strategy

We design a structured prompting strategy that integrates few-shot demonstrations, chain-of-thought reasoning, and counterfactual augmentation for rhetorical classification. The model is instructed to reason over both the original dialogue and its counterfactual variant to determine the underlying rhetorical mechanism.

Instruction:
You are a rhetorical analysis system. You are given a dialogue and its counterfactual reformulation.

Please reason step by step:

1. Interpret the literal meaning of the original utterance.
2. Consider contextual assumptions and pragmatic expectations.
3. Compare the original and counterfactual utterances to determine whether rhetorical meaning is weakened, preserved, or removed.
4. Assign one or more rhetorical labels from the predefined taxonomy.

Rhetorical Labels and Definitions:

- **Irony**: Surface meaning opposite to intent.
- **Echoic Mention**: Mocking repetition of a prior utterance.
- **Hyperbole**: Extreme exaggeration.
- **Rhetorical Question**: A question implying an assertion.
- **Presupposition**: Implicit shared assumption.
- **Innuendo**: Indirect negative implication.
- **Reenactment**: Exaggerated replay of events.
- **Twist**: Sudden narrative contrast or reversal.
- **Self-deprecation**: Self-mockery targeting oneself.

Input Format:
Dialogue: {DIALOGUE}
Counterfactual: {COUNTERFACTUAL}[1]
Output Format:
final label(s).
Few-shot Demonstrations:
Example 1: Hyperbole
Dialogue: A: "So Penny's a little messy."
B: "A little messy? The Mandelbrot set of complex numbers is a little messy. This is chaos."
Counterfactual: A: "So Penny's a little messy."
B: "This is extremely messy."
Label: Hyperbole
Example 2: Irony
Dialogue: A: "I'm not angry at all that my boyfriend called you instead of me. I love that!"
Counterfactual: "I am actually upset that my boyfriend called you instead of me."
Label: Irony
Example 3: Rhetorical Question, Innuendo
Dialogue: A: "What's happening?"
B: "Raj was snooping through Emily's drawers and broke one."

[1]The counterfactual texts used in the COT input are taken directly from the RedSD dataset rather than being generated by RACR.

C: "I’m gonna miss her."
Counterfactual: A: "What’s happening?"
B: "Raj was snooping through Emily’s drawers and broke one."
C: "I think Emily may break up with Raj because of this."
Label: Rhetorical Question, Innuendo