Sarcasm-R1: Enhancing Sarcasm Detection through Focused Reasoning

Qi Yang[1], Liang Yang[1,2]*, Jingjie Zeng[1], Kai Ma[1], Hongfei Lin[1]
[1]School of Computer Science and Technology, Dalian University of Technology, China
[2]Key Laboratory of Social Computing and Cognitive Intelligence, Ministry of Education, China
qiyang@mail.dlut.edu.cn, liang@dlut.edu.cn

Abstract

Sarcasm detection is a crucial yet challenging task in natural language processing. Existing methods primarily rely on supervised learning or prompt engineering, which often struggle to capture the complex reasoning process required for effective sarcasm detection. This paper proposes a novel approach that decomposes sarcasm detection into three fundamental dimensions: language, context, and emotion, meticulously modeling the sarcasm reasoning process. To enhance the quality of reasoning, we employ reinforcement learning algorithms and design customized reward models for each dimension. We utilize five widely used sarcasm detection datasets and annotate the sarcasm reasoning process from these three dimensions to improve the performance of the reward models. Experiments demonstrate that our method outperforms state-of-the-art baseline methods in most cases. Additionally, we observe the central role of emotional contrast in sarcasm detection. Our research provides empirical insights into the mechanism of sarcasm, emphasizing that emotional contrast is at its core, supported by linguistic and contextual cues.[1]

1 Introduction

In recent years, large language models (LLMs) have demonstrated exceptional performance in natural language processing (NLP) tasks, particularly excelling in "System 1" fast-thinking tasks such as sentiment classification and topic analysis, which require rapid and intuitive processing capabilities (Pan et al., 2025). However, these models still face significant challenges in "System 2" slow-thinking tasks, which demand slow, deliberate, and multi-step reasoning, such as logical inference and commonsense reasoning (Wei et al., 2022). As a quintessential "System 2" task, sarcasm detection requires models to go beyond literal meanings and discern hidden sarcastic intent. From the perspective of linguistic cognition, understanding sarcasm is not merely a process of decoding literal language; it involves processing linguistic cues, integrating contextual information, and recognizing emotional contrasts (Matsui et al., 2016). This cognitive process is intricately linked to the mechanisms of the human brain in language processing, background knowledge integration, and emotional understanding (Davis et al., 2016). Therefore, the key problem can be framed as:

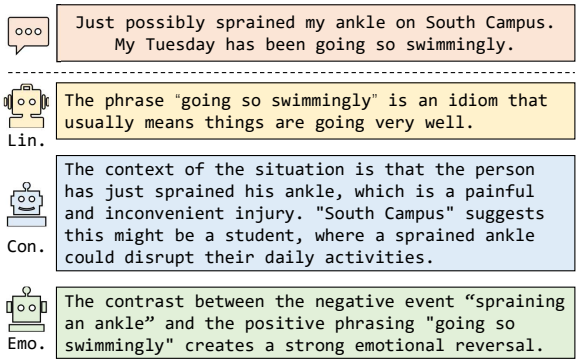

Figure 1: This is an example of sarcastic reasoning for a piece of sarcastic text. "Lin.", "Con.", and "Emo." represent reasoning in the linguistic dimension, contextual dimension, and emotional dimension, respectively.

How can we enhance a model's sarcasm reasoning ability by simulating human cognitive processes?

To tackle this question, we draw on insights from linguistic cognition (Fanari et al., 2023; Pexman, 2018; Pickering et al., 2018; Filik et al., 2016; Naing and Udomwong, 2024) and decompose sarcasm detection into three critical dimensions: language, context, and emotion, see figure 1. Starting from the definition of sarcasm in cognitive linguistics, we abstract the three dimensions of sarcasm. Experimental results show that dividing the analysis of sarcasm into these three dimensions can help the model make correct inferences in most cases.

*Corresponding author.
[1]https://github.com/yangqi1725/Sarcasm-R1

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10773–10785
November 4-9, 2025

This approach systematically guides the reasoning process of large models and is grounded in the cognitive mechanisms humans employ to comprehend sarcasm:

Linguistic Dimension: This focuses on rhetorical devices and expressive features (e.g., keywords, irony, hyperbole, punctuation), which serve as foundational cues for sarcastic expressions.

Contextual Dimension: This examines the text's topic, cultural background, and commonsense knowledge, enabling the model to grasp the context-dependent nature of sarcasm.

Emotional Dimension: This targets emotional expressions and their reversals (e.g., positive emotions masking negative intentions), uncovering the hidden emotional intent behind sarcasm.

To further enhance reasoning quality, we employ a typical reinforcement learning (RL) algorithm, Group Relative Policy Optimization (GRPO) (Shao et al.), and design custom reward models for the three dimensions of language, context, and emotion. These reward models guide the model to focus on the core features of each of the three dimensions of language context and emotion during reasoning. Through interaction with the environment, RL optimizes the policy model, enabling it to gradually master complex sarcasm detection capabilities. Compared to traditional supervised learning, this approach more effectively simulates the human thought process for sarcasm, significantly improving detection performance.

To support this method, we refer to the work of Wang et al. (2024); Kanwal et al. (2025) and provide detailed annotations of the sarcasm reasoning process based on the above three dimensions on five mainstream sarcasm detection dataset. These annotations not only improve the reliability of the reward models, but also provide clear guidance for model reasoning, thereby improving the effectiveness of RL.

Experimental results show that our method outperforms the most advanced baseline methods on these five datasets, showing significant advantages. Ablation experiments further reveal the central role of emotional reversal in sarcasm detection. This observation helps us deeply understand the mechanism of sarcasm generation.

In summary, the contributions of this paper are as follows:

1. We divide the sarcastic reasoning process into three dimensions: language, context, and emotion, refine the modeling of the process, and annotate the process in the original dataset based on these three dimensions, and propose a new dataset named Sarcasm-Reason.

2. In order to learn the sarcastic reasoning process more accurately, we first introduce RL methods into the sarcasm detection task, design reward models for the three dimensions of language, context, and emotion, and effectively improve the model's reasoning quality and sarcastic detection ability.

3. Through extensive experiments and analysis, we find that emotional reversal plays a central role in sarcasm detection, and this discovery help us gain a deeper understanding of the mechanism of sarcasm generation.

2 Related Work

With the rise of deep learning, pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have been widely applied to sarcasm detection tasks. These models capture rich linguistic representations through pre-training on large corpora and achieve significant performance improvements when fine-tuned on sarcasm detection datasets. For instance, Wang et al. (2022) propose an unsupervised method that uses masking and generation techniques to uncover sarcastic cues in text; Liu et al. (2021) employ attention mechanisms to analyze sarcasm from multiple perspectives; and Wen et al. (2023) design a dual inconsistency-aware network focusing on linguistic and emotional inconsistencies. These approaches have improved the accuracy of sarcasm detection by enhancing the model's understanding of linguistic features and context.

In recent years, multimodal methods have gained attention in social media contexts (Pang et al., 2024), as sarcasm is often expressed through multiple modalities such as text, images, or videos. For example, Maity et al. (2022) introduce a multi-task framework that combines textual and visual information to simultaneously detect sentiment, emotion, and sarcasm in memes for cyberbullying identification. Similarly, Jia et al. (2024) use contrastive learning to reduce bias in multimodal sarcasm detection. While these methods excel at handling multimodal data, their primary focus is on eliminating differences between modalities, leaving the exploration of the reasoning process behind sarcasm insufficient.

The advent LLMs, such as Llama (Touvron et al., 2023; Chen et al., 2024) and Qwen (Yang et al., 2024b), marks a new phase in sarcasm detection research. Studies have shown that carefully designed prompt strategies enable LLMs to better understand the implicit intent of sarcasm. For instance, Yao et al. (2025) investigates whether sarcasm detection requires step-by-step reasoning and propose four prompt strategies that significantly outperform standard prompting methods. Nimase and Hong (2024) develops a framework that integrates multiple contextual cues, demonstrating that sequentially adding context can substantially improve performance. Yu et al. (2023) enhances the construction of emotion dependency graphs by incorporating commonsense knowledge, while Qiu et al. (2025) uses commonsense reasoning to capture emotional inconsistencies. These works indicate that LLMs have potential in handling the semantic complexity and emotional contrasts of sarcasm, but systematically guiding their reasoning process requires further exploration.

Recent research has shown that RL can significantly enhance the performance of LLMs in complex reasoning tasks. For example, DeepSeek-AI et al. (2025) demonstrates naturally emerging reasoning capabilities under unsupervised fine-tuning through RL training. The GRPO algorithm introduced in DeepSeekMath (Shao et al.) improves model performance in mathematical reasoning tasks, providing an efficient RL method for our sarcasm detection task. Additionally, research on "Slow-Thinking" Gan et al. (2025) reveals a mechanism for reducing error accumulation by extending reasoning time and refining steps, supporting our strategy of decomposing sarcasm detection into language, context, and emotion dimensions.

In summary, although the field of sarcasm detection has made progress with pre-trained models, multimodal methods, and LLMs, existing studies still lack systematic guidance for the reasoning process and have limit exploration of the working mechanism of sarcasm. Inspired by the aforementioned research, this study combines RL and slow-thinking methods to optimize the model's sarcasm reasoning ability and explore the working mechanism of sarcasm.

3 Methodology

This section introduces the training method of our sarcasm detection model, Sarcasm-R1. The discussion encompasses three key aspects. Initially, it provides a concise overview of the GRPO algorithm that we have employed in our approach. Following this, it elaborates on the training methodology for our reward model, where we have crafted distinct reward models tailored to the dimensions of language, context, and emotion—collectively termed SarGRM for simplicity, as their training processes are uniform. The names of these three reward models are LinGRM, ConGRM and EmoGRM. Finally, the section delves into our data synthesis strategy, which is a three-step process of synthesizing data using LLMs and supplemented by human intervention. And offering detailed statistical insights into the Sarcasm-Reason dataset we have proposed. The framework diagram of this process is shown in figure 2.

3.1 Group Relative Policy Optimization

GRPO is a variant of Proximal Policy Optimization (PPO) (Schulman et al., 2017) that can significantly enhance the reasoning capabilities of model. The core idea of GRPO is to optimize the policy model π_θ by evaluating the relative quality of a group of candidate answers. For a given input question q, GRPO first employs the current policy π_θ to generate G distinct answers $o = \{o_1, o_2, \ldots, o_G\}$. Then, it computes the reward for each answer $r = \{r_1, r_2, \ldots, r_G\}$ using a predefined reward functions. To assess the relative quality of each response, GRPO normalizes the rewards by calculating their mean and standard deviation:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}, \tag{1}$$

where $\text{mean}(\mathbf{r})$ denotes the mean of the rewards of each answer, $\text{std}(\mathbf{r})$ represents the standard deviation of the rewards, and $\hat{A}_{i,t}$ is the normalized advantage score that indicates the relative quality of the $i-\text{th}$ answer relative to all G answers.

To prevent the policy model π_θ from forgetting its original knowledge during training, which could lead to issues like catastrophic forgetting (Li et al., 2024), GRPO incorporates the model parameters from the post-instruction fine-tuning stage as a reference model π_{ref}. It then uses KL divergence to constrain the distribution of prediction scores between π_θ and π_{ref}. In terms of algorithm implementation, GRPO refers to the kl divergence approximator proposed by A Schulman:

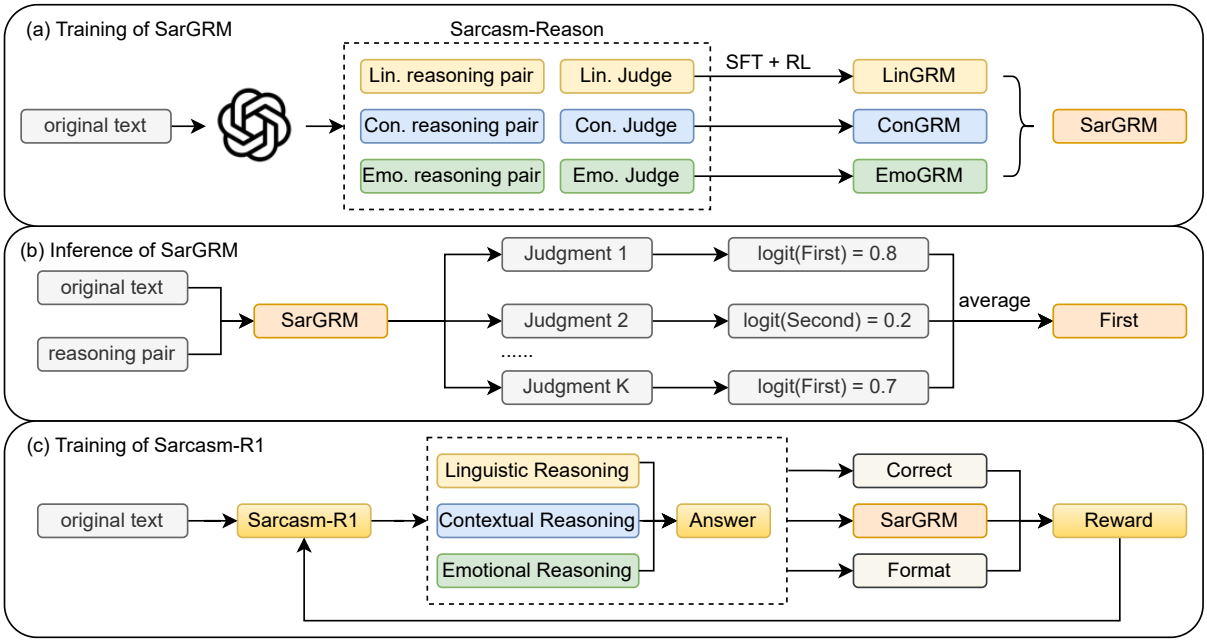

Figure 2: This is the framework diagram of our approach. During SarGRM training, we use pairwise method to train three reward models: LinGRM, ConGRM, and EmoGRM for different dimensions. During SarGRM reasoning, we first sample K judgments on the input through SarGRM, and then confirm the winning reasoning by taking the average of the predicted logits of 'First' and 'Second' token by SarGRM. For Sarcasm-R1 training, in addition to these three reward models, we also use the correctness and format reward function.

$$\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta \| \pi_{\mathrm{ref}}\right] = \frac{\pi_{\mathrm{ref}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} - \log \frac{\pi_{\mathrm{ref}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} - 1, \tag{2}$$

where $o_{i,t}$ represents the $t-$th token of the $i-$th answer o_i. $o_{i,<t}$ represents the first $t-1$ tokens of the $i-$th answer o_i, $\pi_{\mathrm{ref}}(o_{i,t} \mid q, o_{i,<t})$ represents the logit score of the π_{ref} for $o_{i,t}$, $\pi_\theta(o_{i,t} \mid q, o_{i,<t})$ represents the logit score of the π_θ for $o_{i,t}$.

Combining the relative advantage $\hat{A}_{i,t}$ and the KL penalty term $\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta \| \pi_{\mathrm{ref}}\right]$ is the final optimization goal of GRPO:

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{[\pi_\theta(o_{i,t} \mid q, o_{i,<t})]_{\text{no grad}}}\hat{A}_{i,t} - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta \| \pi_{\mathrm{ref}}\right]\right], \tag{3}$$

where β represents the coefficient of the kl penalty term, the default value is 0.04, $|o_i|$ represents the length of o_i.

3.2 Training of SarGRM

The training of SarGRM is divided into two stages, chain-of-thought(CoT) finetune and rule-based RL.

3.2.1 CoT Finetune

An autoregressive language model generates an output sequence $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ given a input context $\mathbf{x}$ by predicting tokens one at a time, based on the previously generated tokens. Assuming that the language model is parameterized by θ, the conditional probability distribution of generating a sequence $\mathbf{y}$ given context $\mathbf{x}$ is $p_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{T} p_\theta(y_t \mid \mathbf{x}, y_{<t})$, with the convention $y_{<1} = \emptyset$ and $\mathbf{y}_{<t} = (y_1, y_2, \ldots, y_{t-1})$. For ease of notation, we define $p_\theta(y_t \mid \mathbf{x}) := p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{x})$. For a vocabulary size M, the probability of predicting the $t-$th token $y_t, p_\theta(y_t \mid \mathbf{x})$ is determined using a softmax with temperature γ on logit scores z of all the tokens: $p_\theta(y_t \mid \mathbf{x}) = \frac{\exp(z_t/\gamma)}{\sum_{i=1}^{M}\exp(z_i/\gamma)}$, where $z_t = \mathrm{logit}_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t})$. Higher values of temperature γ introduce more randomness, while setting $\tau = 0$ corresponds to greedy decoding.

Next-token prediction is the typical approach for pre-training and fine-tuning LLMs. In particular, supervised fine-tuning(SFT) minimizes the cross-entropy loss between the model's predicted next token and the actual target token in a given sequence. Given a dataset $D = (x, y)$ of input context x and target response y, the SFT loss is given by:

Datasets	#Train	#Test	#Avg.Len	#Lin.Avg	#Con.Avg	#Emo.Avg	%Sarcasm
Ghosh	41373	2000	16.8	67.6	78.5	69.2	45.1%
SARC	13667	3406	28.4	66.6	86.9	70.4	50.0%
IAC-V2	7497	1877	48.6	74.8	85.0	73.9	50.0%
iSarcasm	3465	1400	18.1	64.1	77.4	66.4	21.9%
SemEval2018	3832	784	13.9	68.47	79.68	66.42	48.1%

Table 1: #Train and #Test represent the number of data items, #Avg.Len represents the average length of the text to be predicted, #Lin.Avg, #Con.Avg, and #Emo.Avg represent the average lengths of linguistic, contextual, and emotion dimension reasoning, respectively, and %Sarcasm represents the proportion of positive samples.

$$\mathcal{L}_{\mathrm{SFT}}(\theta, \mathcal{D}) = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\sum_{t=1}^{|\mathbf{y}|} \log p_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t})\right] \quad (4)$$

We refer to the method of Zhang et al. (2025) and propose SarGRM, which is trained by the standard next token prediction equation 4 to predict which sarcastic reasoning is better in a pair of sarcastic reasoning. SarGRM uses the probability distribution of LLM over labels to characterize the correctness of the solution, rather than predicting independent numerical scores. For example, input of SarGRM is the original text $\mathbf{x}$ and the sarcastic reasoning pair $(\mathbf{y}_1, \mathbf{y}_2)$, if the first reason $\mathbf{y}_1$ is better among the reason pairs, then maximize $\log p_\theta(\mathbf{'First'}|(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{I}))$, where $\mathbf{I}$ = *"Between the two sarcastic reasons above, which is better(First / Second)?"*. We divide sarcastic reasoning into three dimensions, so there are also three types of sarcastic reasoning pairs $(\mathbf{y}_1, \mathbf{y}_2)$, corresponding to LinGRM, ConGRM and EmoGRM. Considering the complexity of judging the quality of sarcastic reasoning, inspired by Wei et al. (2022), we improve the judgment accuracy by generating CoTs. The training data D_{CoT} is constructed as follows:

$$\begin{aligned}\mathcal{D}_{\mathrm{CoT}} = &\{(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-, \mathbf{I_{CoT}}), (\mathbf{v_{CoT}}, \mathbf{I}, \mathbf{'First'})\} \\ &\bigcup\{(\mathbf{x}, \mathbf{y}^-, \mathbf{y}^+, \mathbf{I_{CoT}}), (\mathbf{v_{CoT}}, \mathbf{I}, \mathbf{'Second'})\},\end{aligned} \quad (5)$$

where $\mathbf{I_{CoT}}$ = *"Let's think about which of these two sarcastic reasons is better step by step."* $\mathbf{v_{CoT}}$ is the process of judgment synthesized by model. $\mathcal{D}_{\mathrm{CoT}}$ is also composed of three parts: language, context, and emotion. In the training phase, $\mathbf{v_{CoT}}$ is synthesized by the external LLMs, and in the inference phase, $\mathbf{v_{CoT}}$ is synthesized by SarGRM.

To sum up the above, at this stage, we use equation 4 to fine-tune SarGRM. The input of SarGRM is $(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-, \mathbf{I_{CoT}})$ or $(\mathbf{x}, \mathbf{y}^-, \mathbf{y}^+, \mathbf{I_{CoT}})$ and the predicted label of SarGRM is $(\mathbf{v_{CoT}}, \mathbf{I}, \mathbf{'First'})$ or $(\mathbf{v_{CoT}}, \mathbf{I}, \mathbf{'Second'})$ respectively.

3.2.2 Rule-Based Reinforcement Learning

SarGRM is further fine-tuned using rule-based online RL. Specifically, we leverage the original GRPO setup, integrating rule-based format rewards and outcome rewards. During the rollout phase, SarGRM generates judgments and predicted scores based on the input text and the corresponding sarcastic reasoning pairs, and then we use a predefined regular expression to match the output format and extract the predicted score and compare it with the correct answer.

3.2.3 SarGRM Inference

In the inference phase, we first use SarGRM to generate correctness judgments CoT $\mathbf{v_{CoT}}$, and then assign correctness scores to candidate solutions based on the probabilities of "first" and "second" tokens. When sampling CoT, SarGRM may generate different reasoning paths for the same text and reasoning pair, resulting in changes in correctness probabilities. As such, we would like to marginalize out these reasoning paths to select the most consistent correctness answer (Wang et al., 2023). To do so, we use majority voting: for the generated K judgments chain reasons, we compute the average scores of these reasons as follows:

$$r = \frac{1}{K}\sum_{i=1}^{K} p_\theta\left(\mathbf{'First'}|\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-, \mathbf{I_{CoT}}, \mathbf{v}_{\mathbf{CoT}}^{(i)}, \mathbf{I}\right), \quad (6)$$

where $\mathbf{v}_{\mathbf{CoT}}^{(i)} \sim p_\theta(\cdot|\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-, \mathbf{I_{CoT}})$, p_θ is SarGRM. Since individual CoT judgment may contain errors, majority voting effectively mitigates the impact of these errors by averaging the correctness scores across multiple judgments. More importantly, this demonstrates that Sarcasm-GRM can improve accuracy by increasing computational resources during inference (i.e., generating more CoT reasons). In our implementation, Sarcasm-GRM uses majority voting based on 16 votes, meaning $K = 16$ in the equation 6.

3.3 Sarcasm-Reason Dataset

We use five mainstream sarcasm detection benchmark datasets and synthesize the sarcasm reasoning process based on these datasets in three dimensions: language, context, and emotion. Specifically include: *Ghosh* (Ghosh and Veale, 2017) which is collected from Twitter and annotated automatically, *SARC* (Khodak et al., 2018) that only contain political content, *IAC-V2* (Oraby et al., 2016) which is obtained from Internet Argument Corpus, *iSarcasm* (Oprea and Magdy, 2020) that encompass tweets which are written by online users and *SemEval2018* (Hee et al., 2018), which is collected from SemEval 2018 Task 3.

Following the approach in Wang et al. (2024), our data pair construction process consists of three main steps: Standard Reasoning Generation, Negative Sample Generation, and Data Validation. In our methodology, we use GPT-4o to synthesize the positive and negative sample pairs for sarcasm reasoning. Subsequently, we employ Claude-3.5 Sonnet, Gemini 2.0 Flash, and Qwen-Max to evaluate the quality of these pairs.

3.3.1 Standard Reasoning Generation

We leverage LLMs to synthesize sarcastic reasoning processes. In the **linguistic** dimension, we focus on features such as keywords, rhetorical devices, polysemy, punctuation, and homophones in the original text. In the **contextual** dimension, we consider elements including the topic, cultural background, social consensus, and common knowledge of the original text. In the **emotional** dimension, we primarily analyze emotional words and instances of emotional reversal. Based on this analysis, we design a group of prompts and utilize LLMs to annotate sarcastic reasoning results for the original dataset: $\mathbf{y}^{+}_{i,j} = LLM(\mathbf{x}_i, \mathbf{p}_j)$, where $\mathbf{x}_i$ represents the $i-\text{th}$ data entry in the dataset, and $\mathbf{p}_j$ denotes the $j-\text{th}$ prompt from the prompt group. Figure 3 shows a prompt example.

3.3.2 Negative Sample Generation

We generate "noisy" versions of the original prompts to construct low-quality reasoning negative samples. There are two principles for constructing negative samples: (1) the reasoning do not accurately express the reasoning content of this dimension, and (2) the reasoning confused with the content of other dimensions. Specifically, we employ the following strategies to create "noisy" prompts: adding irrelevant content, removing critical information, or partially replacing content across different dimensions. Additionally, we record the modifications made to the original prompts and analyze why the standard reasoning process outperforms the negative samples: $\mathbf{y}^{-}_{i,k} = LLM(\mathbf{x}_i, \mathbf{p}'_j)$, where $\mathbf{x}_i$ represents the $i-\text{th}$ data entry in the dataset, and $\mathbf{p}'_k$ denotes the $k-\text{th}$ prompt from the noisy prompt group.

> Here is a text that is {not} a sarcastic text. Please deduce the reason why it is {not} sarcastic from the perspectives of (1)language, (2)context, and (3)emotion in this text. Please answer in the following format:
> <reason>
> (1)language: like keywords, rhetorical de vices, punctuation and language style
> (2)context: like topic, cultural background, social consensus, common knowledge
> (3)emotion: like emotional words, special symbols and emotional reversal
> </reason>
> Here is the text: {text}

Figure 3: This is our original prompt. The other prompts in the prompt group are slightly modified from this.

3.3.3 Data Validation

To verify the correctness of positive and negative reasoning samples, we randomly combine them corresponding to the same $\mathbf{x}_i$ and submit them to multiple distinct LLMs for evaluation. If all LLMs can accurately distinguish between the positive and negative samples, the positive-negative sample pair, along with the reason $\mathbf{J}$ why the positive sample outperforms the negative one. Finally, for each original text $\mathbf{x}_i$, we retain a standard sarcastic reasoning positive sample $\mathbf{y}^{+}_{i}$ and several negative samples $\mathbf{y}^{-}_{i,k}$, as well as the corresponding positive and negative sample judgment reasoning process $\mathbf{J}_{i,k}$. Table 1 gives the information about the original dataset including the standard sarcastic reasoning process.

Human review is an integral part of our data synthesis process. Our specific procedure is as follows: for each generated positive-negative pair, we consider it a high-quality sample only when all three judge models make the correct assessment. This initial step filters out 80% of the originally generated pairs. For the remaining samples, we conduct a manual review, focusing on whether the emotional reversal between the positive and negative samples is reasonable, thereby ensuring the quality

Model	Ghosh		SARC		IAC-V2		iSarcasm		SemEval2018	
	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1
RoBERTa	72.2	72.4	66.7	66.7	76.6	76.7	78.6	56.8	70.2	69.1
DC-Net	80.2	78.6	72.9	72.4	78.0	77.9	78.8	58.7	70.8	69.6
SD-APRR	82.6	82.3	-	-	78.8	78.8	80.3	61.2	72.2	70.7
SensoryT5	86.1	**86.1**	-	-	83.0	83.0	-	-	77.7	77.9
SarcasmCue	83.0	82.9	74.1	73.7	73.4	72.3	79.4	60.3	74.0	74.0
EICR	86.2	84.3	77.2	75.3	84.5	83.8	83.3	70.4	80.1	80.3
Gemini 2.0 Flash	84.5	83.3	72.5	70.6	74.1	72.7	78.6	63.8	74.4	72.6
GPT-4o	82.4	82.1	74.1	70.0	73.0	72.0	76.0	59.4	69.0	68.2
Claude 3.5 Sonnet	82.9	81.4	73.1	72.4	76.8	76.6	74.7	61.2	75.1	75.1
Qwen-Max	84.1	82.9	77.4	75.8	78.4	76.3	80.4	65.7	76.9	75.6
LLaMA SFT	79.3	77.1	70.6	68.9	75.3	74.9	74.9	61.4	71.9	70.4
LLaMA 0-shot	74.2	72.2	67.7	65.3	69.7	67.9	69.2	55.4	62.4	61.5
LLaMA 5-shot	76.9	74.1	69.2	66.7	70.1	68.8	70.1	55.9	62.4	61.9
Qwen SFT	78.2	75.1	71.9	70.2	74.1	72.0	71.7	62.4	70.9	70.4
Qwen 0-shot	73.9	71.5	68.7	65.9	67.7	66.9	70.2	56.4	61.6	61.2
Qwen 5-shot	75.2	73.9	70.2	67.9	71.9	69.1	69.3	55.1	61.9	62.0
Sarcasm-R1	**87.3**	85.4	**79.2**	**77.3**	**85.6**	**84.4**	**85.4**	**72.3**	**82.1**	**81.9**

Table 2: Performance comparison of different methods on datasets. Acc denotes Accuracy(%) and Ma-F1 denotes Macro-F1(%). The input of Gemini 2.0 Flash, GPT-4o, Claude 3.5 Sonnet and Qwen-Max models is 0-shot CoT prompt. The best results are represented in bold. The second-best results are underlined.LLaMA is LLaMA 3.1 8B Instruct and Qwen is Qwen2-7B-Instruct.

Method	Ghosh		SARC		IAC-V2		iSarcasm		SemEval2018	
	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1
w/o Lin	80.9	74.6	75.1	70.5	82.6	77.3	81.8	62.7	80.2	78.3
w/o Con	86.1	81.2	77.9	74.3	80.7	74.3	83.4	63.0	79.1	74.3
w/o Emo	82.7	79.1	73.3	71.2	81.9	76.2	78.9	63.3	78.1	77.2
Sarcasm-R1	87.3	85.4	79.2	77.3	85.6	84.4	85.4	72.3	82.1	81.9

Table 3: Result of ablation experiments on reward models in three dimensions: language, context, and emotion.

of the pairs. This process is inspired by the data synthesis method proposed by Meta FAIR (Wang et al., 2024). The rationale for this approach is our focus on the relative quality of the positive-negative pairs. This data is primarily used to train our pairwise reward models: LinGRM, ConGRM, and EmoGRM. Therefore, during data synthesis and filtering, we prioritize the relative quality of pairs, using a combination of multi-LLM screening and human review to ensure data integrity.

4 Experiments

We conduct extensive experiments to evaluate the performance of our method. On this basis, we evaluate the contribution of the quality of reasoning in different dimensions to the final sarcasm prediction through ablation experiments. The details of the experiment implementation are shown in the Appendix A, including the training details of Sar-GRM and Sarcasm-R1, and the model used for data synthesis.

4.1 Baseline

We compare our method against 10 mainstream models, including 6 baseline methods for sarcasm detection and 4 high-performance general models. The baseline methods include: (1)*RoBERTa* (Liu et al., 2019), which served as a strong baseline by capturing nuanced contextual and linguistic features; (2)*DC-Net* (Liu et al., 2022), which modeled literal and implied sentiments separately to recognize sentiment conflict; (3)*SD-APRR* (Min et al., 2023), an incongruity reasoning model that employed a denoising module based

data	1	2	3	4	5	6	7	8	avg	p_{value}	statistic
Ghosh w/o Lin	78.5	79.0	81.5	79.9	83.7	83.6	81.9	79.4	80.9	0.165	-1.55
Ghosh w/o Emo	81.1	80.9	83.0	84.4	80.7	81.5	83.7	86.1	82.7		
iSarcasm w/o Lin	80.8	82.6	82.7	81.8	82.2	81.6	80.9	82.2	81.8	5.7e-06	12.2
iSarcasm w/o Emo	78.9	78.9	79.8	78.4	79.7	79.4	77.9	78.5	79.9		

Table 4: We select two extreme cases as our test objects, namely **Ghosh** and **iSarcasm**. For the former, when the language dimension is ablated, the average performance drops more than emotion dimension, while for the latter, the opposite is true. The test involved conducting 8 ablation experiments with the same settings on these two datasets and analyzing the significance of the results.

on a commonsense-augmented dependency graph; (4)*SensoryT5* (Zhao et al., 2025), which integrated sensory knowledge into the T5 framework's attention mechanism to facilitate sensory emotional interactions; (5)*SarcasmCue* (Yao et al., 2025), which introduced a prompting framework that elicited LLMs to detect sarcasm by considering sequential and non-sequential prompting methods; (6)*EICR* (Qiu et al., 2025), which performs incongruous reasoning based on commonsense enhancement and adopts adversarial contrastive learning to improve the robustness of the detector. The high-performance general models include: (1)*Gemini 2.0 Flash* (Google, 2025), (2)*GPT-4o* (OpenAI, 2025), (3)*Claude 3.5 Sonnet* (Anthropic, 2025), (4)*Qwen-Max* (Tongyi, 2025).

4.2 Analysis

We present our experimental results in this subsection and design three research questions (RQs) to aid our analysis.

RQ1: Is focusing on the sarcasm reasoning process effective for sarcasm detection?

Our method is evaluated against the baseline methods shown in Table 2. We observe that: (1)Sarcasm-R1 achieves the best performance on five public benchmarks in most case, demonstrating the effectiveness of improving sarcasm detection performance by focusing on the sarcasm reasoning process; (2)The baseline methods exhibit relatively low Macro-F1 scores on the iSarcasm dataset, which may be caused by the imbalanced label distribution. In contrast, Sarcasm-R1 shows strong robustness in handling this imbalance; (3)The performance improvement observed compared to SarcasmCue suggests that guiding LLM thinking through prompts alone will lead to a certain degree of hallucination. We use RL algorithm to improve the model's reasoning ability, which can effectively alleviate the phenomenon of hallucinations in the process of sarcastic reasoning. The comparison results with the general artificial intelligence models Gemini 2.0 Flash, GPT-4o, Claude 3.5 Sonnet and Qwen-Max also illustrate this point; (4)Compared with EICR, SensoryT5 and SD-APRR, the experimental results show that our comprehensive consideration of sarcasm from the three dimensions of language context and emotion is more comprehensive.

RQ2: Is it reasonable to decompose sarcasm reasoning into three dimensions of language, context, and emotion?

We observe that eliminating any of the three dimensions would have a significant impact on the final model performance. (1) In the ablation experiment of the linguistic dimension, the average accuracy drop by 3.8% and the average Macro-F1 drop by 7.6%. And the Macro-F1 drop in this experiment is the largest, which reflects the importance of the linguistic dimension in determining the balance between positive and negative samples. This was also demonstrated by the ablation experiment on the iSarcasm dataset with an imbalance of positive and negative samples. (2) In the contextual dimension, the average accuracy drop by 2.5% and the average Macro-F1 drop by 6.8%. When ignoring context reasoning, the drop in Macro-F1 value is the largest relative to the drop in accuracy, which also illustrates the importance of context dimension reasoning in distinguishing the balance between positive and negative samples. (3) In the emotional dimension, the average accuracy drop by 4.9% and the average Macro-F1 drop by 6.7%. As can be seen in 4, although the average performance drops more when the language dimension is ablated, the p-value = 0.165 indicates that the result is not significant. In iSarcasm, the average performance drop is smaller when the language dimension is ablated, and the p-value = 5.7e-06 indicates that the result is sufficiently significant.

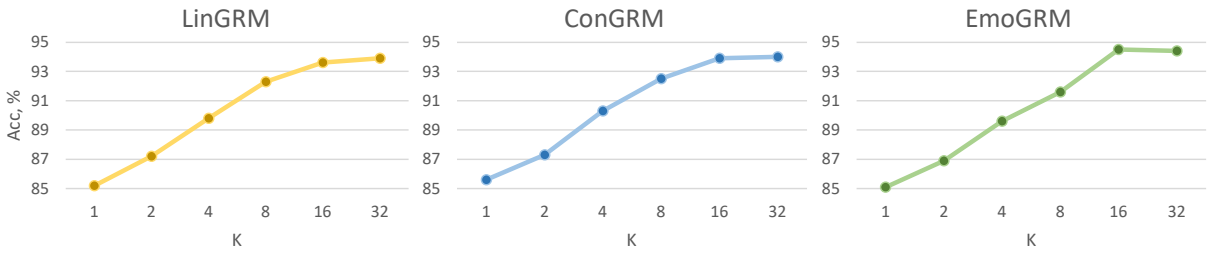

Figure 4: We test the performance change trend of LinGRM, ConGRM and EmoGRM reward models as the change of sampling number K. The horizontal axis is the value of K, and the vertical axis is the accuracy of the reward models classification, in %.

This reflects the core role of emotional reasoning in sarcasm detection, which is mainly based on the reasoning of emotion reversal. This observation demonstrates the importance of the three dimensions of language, context, and emotion for sarcasm detection, and also proves the rationality of our decomposition of sarcasm into these three dimensions.

RQ3: Are the judgments of the three reward models reliable?

In this research problem, we test the performance of three reward models under different K values, as shown in Figure 4. The experiment shows that as the number of $\mathbf{v_{CoT}}$ samples gradually increases, the performance of the reward model shows a certain scalability. This shows that we can improve the performance of the reward model by increasing the inference cost. This proves the reliability of our reward model.

Through the analysis of the above three RQs, we can draw the following three conclusions:

1. By focusing on the sarcastic reasoning process, the sarcasm detection ability of model can be effectively improved.
2. It is reasonable to decompose sarcasm into three dimensions: language, context, and emotion.
3. Rewarding the rationality of model design is helpful to improve the quality of sarcastic reasoning.

To summarize these three research questions, we find that the emotional dimension dominated by emotional reversal is at the core, but the accurate expression of emotional reversal depends on the joint effect of language clues and common context. The coordinated thinking of these three dimensions is the core of detecting sarcasm.

5 Conclusion

This study aims to study the working mechanism of sarcasm. First, we decompose sarcasm into three dimensions: language context and emotion. By using reinforcement learning algorithms, we design reward models for each of these three dimensions to enhance the model's reasoning ability for sarcasm. In order to improve the reliability of the reward model, we propose the Sarcasm-Reason dataset and rely on repeated sampling methods to expand the performance of the reward model. Our method outperforms traditional methods and achieves the best accuracy on all five datasets, which proves the rationality of our method. Finally, we rely on experimental observations to summarize the working mechanism of sarcasm from the above three dimensions.

Limitations

Sarcasm is highly dependent on human subjective judgment. Although we have experimentally proved that it is reasonable to decompose sarcasm into three dimensions: language, context, and emotion, it is undeniable that there are still other angles that we have not noticed, which can better explain sarcasm in some cases.

Acknowledgments

We thank reviewers for their comments, which provided some insights on this research that will further influence our future work. This work is partially supported by grants from the Key R&D Projects in Liaoning Province award numbers (2023JH26/10200015), the Natural Science Foundation of China award numbers (62376051, 62366040, 62076046, 62066044, 61976036, 61702080) and the Fundamental Research Funds for the Central Universities award number (DUT24LAB123).

References

Anthropic. 2025. Claude 3.5 Sonnet.

Songlin Chen, Weicheng Wang, Xiaoliang Chen, Peng Lu, Zaiyan Yang, and Yajun Du. 2024. Llama-lora neural prompt engineering: A deep tuning framework for automatically generating chinese text logical reasoning thinking chains. *Data Intell.*, 6(2):375–408.

Cameron L Davis, Kenichi Oishi, Andreia V Faria, John Hsu, Yessenia Gomez, Susumu Mori, and Argye E Hillis. 2016. White matter tracts critical for recognition of sarcasm. *Neurocase*, 22(1):22–29.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, and Junxiao Song et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Rachele Fanari, Sergio Melogno, and Roberta Fadda. 2023. An experimental study on sarcasm comprehension in school children: The possible role of contextual, linguistics and meta-representative factors. *Brain Sciences*, 13(6):863.

Ruth Filik, Alexandra Ţurcan, Dominic Thompson, Nicole Harvey, Harriet Davies, and Amelia Turner. 2016. Sarcasm and emoticons: Comprehension and emotional impact. *Quarterly Journal of Experimental Psychology*, 69(11):2130–2146.

Zeyu Gan, Yun Liao, and Yong Liu. 2025. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning. *CoRR*, abs/2501.15602.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 482–491. Association for Computational Linguistics.

Google. 2025. Gemini 2.0 Flash.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 39–50. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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.

Saima Kanwal, Ali Raza, Chunyan Bai, Dawei Zhang, Jing Wen, and Dileep Kumar. 2025. An effective machine learning approach with hyper-parameter tuning for sentiment analysis. *Data Intell.*, 7(1):70–94.

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, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

Hejing Liu, Qiudan Li, Zaichuan Tang, and Jie Bai. 2021. An attention based multi-view model for sarcasm cause detection (student abstract). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 15833–15834. AAAI Press.

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. *CoRR*, abs/1907.11692.

Yiyi Liu, Yequan Wang, Aixin Sun, Xuying Meng, Jing Li, and Jiafeng Guo. 2022. A dual-channel framework for sarcasm recognition by detecting sentiment conflict. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1670–1680. Association for Computational Linguistics.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1739–1749. ACM.

Tomoko Matsui, Tagiru Nakamura, Akira Utsumi, Akihiro T Sasaki, Takahiko Koike, Yumiko Yoshida, Tokiko Harada, Hiroki C Tanabe, and Norihiro Sadato. 2016. The role of prosody and context in sarcasm comprehension: Behavioral and fmri evidence. *Neuropsychologia*, 87:74–84.

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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10172–10183. Association for Computational Linguistics.

Shwe Zin Su Naing and Piyachat Udomwong. 2024. Public opinions on chatgpt : An analysis of reddit discussions by using sentiment analysis, topic modeling, and SWOT analysis. *Data Intell.*, 6(2):344–374.

Ojas Nimase and Sanghyun Hong. 2024. When do "more contexts" help with sarcasm recognition? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 17537–17543. ELRA and ICCL.

Rosel Oida-Onesa and Melvin A. Ballera. 2024. Fine tuning language models: A tale of two low-resource languages. *Data Intell.*, 6(4):946–967.

OpenAI. 2025. GPT-4o.

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, ACL 2020, Online, July 5-10, 2020*, pages 1279–1289. Association for Computational Linguistics.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn A. Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 31–41. The Association for Computer Linguistics.

Qianjun Pan, Wenkai Ji, Yuyang Ding, Junsong Li, Shilian Chen, Junyi Wang, Jie Zhou, Qin Chen, Min Zhang, Yulan Wu, and 1 others. 2025. A survey of slow thinking-based reasoning llms using reinforced learning and inference-time scaling law. *arXiv preprint arXiv:2505.02665*.

Jinhui Pang, Xinyun Yang, Xiaoyao Qiu, Zixuan Wang, and Huang Tai Sheng. 2024. MMAF: masked multimodal attention fusion to reduce bias of visual features for named entity recognition. *Data Intell.*, 6(4):1114–1133.

Penny M Pexman. 2018. How do we understand sarcasm? *Frontiers for Young Minds*, 6.

Bethany Pickering, Dominic Thompson, and Ruth Filik. 2018. Examining the emotional impact of sarcasm using a virtual environment. *Metaphor and Symbol*, 33(3):185–197.

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, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 9062–9073. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, and Léonard Hussenot et al. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

John Schulman. Approximating kl divergence.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. *URL https://arxiv.org/abs/2402.03300*.

Tongyi. 2025. Qwen-Max.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Rui Wang, Qianlong Wang, Bin Liang, Yi Chen, Zhiyuan Wen, Bing Qin, and Ruifeng Xu. 2022. Masking and generation: An unsupervised method for sarcasm detection. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2172–2177. ACM.

Tianlu Wang, Ilia Kulikov, and Olga Golovneva et al. 2024. Self-taught evaluators. *CoRR*, abs/2408.02666.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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 *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Changsong Wen, Guoli Jia, and Jufeng Yang. 2023. DIP: dual incongruity perceiving network for sarcasm detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2540–2550. IEEE.

An Yang, Baosong Yang, and Beichen Zhang et al. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang, Baosong Yang, Binyuan Hui, and Bo Zheng et al. 2024b. Qwen2 technical report. *CoRR*, abs/2407.10671.

Ben Yao, Yazhou Zhang, Qiuchi Li, and Jing Qin. 2025. Is sarcasm detection a step-by-step reasoning process in large language models? In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 25651–25659. AAAI Press.

Zhe Yu, Di Jin, Xiaobao Wang, Yawen Li, Longbiao Wang, and Jianwu Dang. 2023. Commonsense knowledge enhanced sentiment dependency graph for sarcasm detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 2423–2431. ijcai.org.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025. Generative verifiers: Reward modeling as next-token prediction. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Qingqing Zhao, Yuhan Xia, Yunfei Long, Ge Xu, and Jia Wang. 2025. Leveraging sensory knowledge into text-to-text transfer transformer for enhanced emotion analysis. *Inf. Process. Manag.*, 62(1):103876.

A Implementation Details

For the three reward models LinGRM, ConGRM and EmoGRM, we use the Lora method (Hu et al., 2022; Oida-Onesa and Ballera, 2024) to train Gemma 7B (Rivière et al., 2024). The Lora Adapter is configured on all linear layers of Gemma 7B, and both rank and alpha are set to 32. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $2e-6$. We use a linear warm up with 1000 gradient steps and a cosine decay scheme that decays to 10% of the peak learning rate after the decay period.

We conduct RL based on based on Qwen2.5-7B-Instruct (Yang et al., 2024a). The learning rate of the policy model is set to 1e-6. The KL coefficient is 0.04. For each problem, we sample 8 outputs, the maximum length is set to 2048, and the batch size is set to 8. The policy model is updated only once in each exploration phase.

In the data synthesis stage, we use the GPT-4o to synthesize our positive and negative reasoning samples. In the data validation, we use Claude 3.5 Sonnet, Gemini 2.0 Flash and Qwen-Max.

Our training is divided into three stages.

Stage 1: We train the model for approximately 50 steps using only a format reward function. This stage aims to familiarize the model with the correct response format, which facilitates the subsequent extraction of its reasoning and predicted answer. In this process, if the model's response conforms to the predefined structure, the reward score is 1; otherwise, it is 0.

Stage 2: We use the format reward combined with our three specialized reward models. This stage is primarily focused on enhancing the model's sarcastic reasoning ability. If the model's response conforms to the predefined structure, the reward is 0; otherwise, it is -1. For the reward models, if the model's reasoning for a corresponding dimension is superior to the default reasoning, the reward is 0.5; otherwise, it is 0.

Stage 3: We use the format reward, a correctness reward, and the three specialized reward models to ultimately boost the model's performance. If the response conforms to the predefined structure, the reward is 0; otherwise, it is -1. If the model's reasoning for a corresponding dimension is superior to the default reasoning, the reward is 0; otherwise, it is -0.5. If the model's final answer is correct, it receives a reward of 1.5; otherwise, the reward is 0.

During our training process, we set the maxi-

mum prompt length for the model to 1024 tokens and the maximum length for the reasoning path to 2048 tokens. Based on our observations and statistics, the model's response rarely exceeds 2048 tokens in the vast majority of cases. This indicates that the length of the reasoning path is not the primary factor limiting the model's performance.

Our analysis suggests this is because the input sarcastic prompts are relatively short (typically under 150 tokens) and do not qualify as long-text inputs. Furthermore, the reasoning required for the dimensions of language, context, and emotion does not necessitate excessively long responses to be high-quality. Specifically: Language reasoning involves analyzing elements like keywords, rhetorical devices, intonation, linguistic style, punctuation, and emojis. Contextual reasoning involves analyzing the topic, social consensus, cultural background, and common sense. Emotional reasoning involves analyzing sentiment words and emotional incongruity. For all these aspects, a high-quality explanation can be generated without being particularly verbose.

The training of the reward model took a total of 15 hours on a single L40 GPU. For the rl phase, we use a server with 8 A100 GPUs, and the training took approximately 60 hours. The GPU allocation is : 3 A100s are used to host the 3 reward models. 1 A100 is used for the reference model during the GRPO training process. The remaining 4 A100s are dedicated to updating the policy model's parameters. The training for our ablation studies is similar, with each experiment requiring approximately 2/3 to 3/4 of the compute resources of the full training process.

For inference, we utilize the vLLM acceleration framework. This enables our model to achieve real-time responses (within a second) on a single 4090 GPU.