

# The Difficult Case of Intended and Perceived Sarcasm: a Challenge for Humans and Large Language Models

Hyewon Jang<sup>1,3</sup> & Diego Frassinelli<sup>2</sup>

<sup>1</sup>Department of Linguistics, University of Konstanz

<sup>2</sup>MaiNLP, Center for Information and Language Processing, LMU Munich

<sup>3</sup>Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg

hyewon.jang@gu.se, frassinelli@cis.lmu.de

## Abstract

We examine the cases of failed communication in sarcasm, defined as ‘the discrepancy between what speakers and observers perceive as sarcasm’. We identify factors that are associated with such failures, and how those difficult instances affect the detection performance of encoder-only and decoder-only generative models. We find that speakers’ incongruity between their felt annoyance and sarcasm in their utterance is highly correlated with sarcasm that fails to be communicated to human observers. This factor also relates to the drop of classification performance of large language models (LLMs). Additionally, disagreement among multiple observers about sarcasm is correlated with poorer performance of LLMs. Finally, we find that generative models produce better results with ground-truth labels from speakers than from observers, in contrast to encoder-only models, which suggests a general tendency by generative models to identify with speakers’ perspective by default.

## 1 Introduction

An utterance that is intended to be sarcastic by the speaker is sometimes not understood as such by the listener or external observers, or vice versa (Fox Tree et al., 2020). Consider the example below.

About two years ago, Steve spent half a year in Japan, where he learned a lot about Japanese food culture. Ever since then, whenever Steve and John eat something together, Steve says some version of, “you know, in Japan, people do it this way.” And John says, “that’s cool to hear!”

In this situation, if an external observer thinks that John is being sarcastic, but John intended to be literal, there is a discrepancy between intended and perceived sarcasm.<sup>1</sup> This type of communication failure can occur in numerous communicative

<sup>1</sup>In this work, the discrepancy we address is between speaker and observer rather than speaker and listener. An

scenarios, especially those requiring layers of inferences, which are common features of sarcasm (Bryant, 2023). Discussing the divergence between intended and perceived sarcasm is not new. Prior work in psycholinguistics has widely discussed the differences in what speakers intend with sarcastic utterances and how listeners or observers interpret them (Pexman and Olineck, 2002). NLP tasks and datasets are also affected by such discrepancies. For instance, Oprea and Magdy (2020) demonstrated that there are many instances for which external annotators provide different sarcasm judgments from the producers of the utterances. Sarcasm detection by language models, especially BERT-like models, also show different classification performances according to ground-truth labels based on self-evaluation versus external evaluation (Abu Farha et al., 2022; Jang and Frassinelli, 2024; Oprea and Magdy, 2019; Plepi and Flek, 2021). Since the capacity of (large) language models has increased exponentially over the years with the advent of generative models, which are often placed in direct conversations with human users, it has become an important question to ask how language models navigate different perspectives in communication involving sarcasm.

Although it is evident that sarcasm judgment is contingent on the different perspectives of speakers and observers (Oprea and Magdy, 2020), there is a lack of systematic investigation on what factors contribute to the general difficulty of the task for LLMs as well as for human observers. One of the numerous keys to identifying the source of such discrepancy between speakers and observers is to think about why sarcasm is used in the first place.

observer is a non-participant of a conversation who evaluates the interaction from an external perspective. Though this is less natural for real communication, it is more relevant for computational linguistics, as data are often created with evaluation by external observers.

Sarcasm is used to convey specific communicative intentions, such as to mock the addressee (Gibbs, 2000), which in turn is motivated by speaker’s emotion in a given communicative situation (Jang et al., 2023). The strong link between emotion and sarcasm has long been identified and discussed in numerous previous studies (e.g., Filik, 2023; Jang et al., 2023; Veale, 2023). As such, we focus on the close connection between sarcasm and emotion to examine the discrepancy between speakers and external observers in the use of sarcasm. Specifically, we focus on *annoyance*, an emotion shown to strongly influence sarcasm production and identification (Jang et al., 2023). Though annoyance is not the only reason why a speaker chooses to use sarcasm, we focus on annoyance in this work based on Jang et al. (2023), who report that a strong connection is observed between speaker’s annoyance and the level of sarcasm in their utterance, and that external observers are also able to capture this connection. The availability of such information in the dataset described in Section 3.1 also motivates such research design.<sup>2</sup> We demonstrate which factors are associated with the divergence of sarcasm judgment between speakers and observers and how this affects (L)LM performance on sarcasm detection.

## 2 Related work

### 2.1 Intended vs. perceived sarcasm

Numerous previous studies exist on sarcasm detection, but very few of them address the perspective divergence between speakers and external observers (Dadu and Pant, 2020; Khodak et al., 2018; Kumar and Anand, 2020; Misra and Arora, 2023). In fact, an absolute majority of sarcasm datasets contain labels annotated by third-party annotators (Castro et al., 2019; Khodak et al., 2018; Oraby et al., 2016), or a combination of self-labels and third-party labels (Khodak et al., 2018; Van Hee et al., 2018). Some datasets provide only author labels without third-party labels (Oprea and Magdy, 2020). Only a small body of work addresses the difference between intended and perceived sarcasm (Jang et al., 2023; Jang and Frassinelli, 2024; Oprea

<sup>2</sup>We further tested the validity of annoyance as a relevant emotion to sarcasm in a separate preliminary experiment using an emotion classification model ([https://huggingface.co/bsingh/roberta\\_goEmotion](https://huggingface.co/bsingh/roberta_goEmotion)) fine-tuned on the GoEmotions dataset (Demszky et al., 2020). When using this model, the logits of the top 20% of most important emotions (out of 28 categories) for sarcastic utterances from CSC were *annoyance*, *admiration*, *amusement*, *approval*, and *curiosity*.

and Magdy, 2020; Plepi and Flek, 2021; Shmueli et al., 2020). They report that there is a noticeable difference in LM performance depending on the source of ground-truth labels. But the discussion of which factors may contribute to such difference, or how the difference can be used to evaluate LLM performance has not been extensively addressed in the literature.

### 2.2 The connection between sarcasm and emotion

Previous work has identified numerous reasons for which human communicators use sarcasm. Sarcasm can be used to express an attitude (Colston, 2023), to cause certain emotional reactions in the listener (Filik, 2023), or to achieve specific communicative goals such as to be humorous (Gibbs, 2000), appear emotionally controlled (Dews et al., 1995), or mock the addressee (Pexman and Olineck, 2002). These communicative functions are often motivated by the emotion in reaction to an experience (Jang et al., 2023). Sarcasm as such is strongly related to emotions, whether sarcasm serves as the trigger for emotional reactions or is itself triggered by them.

## 3 Method

### 3.1 Data

We used the publicly available Conversational Sarcasm Corpus (CSC; Jang and Frassinelli, 2024)<sup>3</sup> to analyze misaligned cases between intended and perceived sarcasm. CSC provides a good opportunity to examine divergences in sarcasm judgment because it provides evaluations of two concepts (sarcasm & emotion) reported by both speakers and multiple external observers (4-6 per speaker). Specifically, it contains contexts and utterances (N = 7,036), ratings for *sarcasm* and *annoyance felt by the speaker* that are judged by two parties (speaker & observers). The original ratings provided in the dataset are text-coded as 1 (*not at all*) - 2 (*mostly not*) - 3 (*not so much*) - 4 (*somewhat*) - 5 (*mostly*) - 6 (*completely*), which makes both numerical manipulation and binarization possible.

### 3.2 Hypotheses

We identified two potential sources of gap for which observers reach a different judgment about sarcasm than the speakers:

<sup>3</sup><https://github.com/CoPsyN/CSC>

Type	Text	Sarc(S)	Sarc(O)	Ann(S)	Ann(O)
<b>H1: Speaker’s annoyance-sarcasm incongruity</b>	Context: You got a date this evening. When you tell Steve you got a date, he asks, “ooh, what’s the plan?” Response: <i>We’re going for Malaysian and then a gig.</i>	<u>6</u>	1	<u>2</u>	1
<b>H2: Speaker-observer annoyance misalignment</b>	Context: About two years ago, Steve spent half a year in Japan, where he learned a lot about Japanese food culture. Ever since then, whenever you eat something together, Steve says some version of, “you know, in Japan, people do it this way.” Response: <i>That’s cool to hear!</i>	1	6	<u>1</u>	<u>6</u>

Table 1: Examples of *speaker’s annoyance-sarcasm incongruity* (boxed) and *speaker-observer annoyance misalignment* (underlined) associated with sarcasm failure (6 vs. 1). Sarc=Sarcasm ratings, Ann=Annoyance ratings, S=Speaker, O=Observer.

- **H1: speakers’ annoyance-sarcasm incongruity:** An incongruity between a speaker’s annoyance and the level of sarcasm in the output utterance causes misalignment between self-rated and other-rated sarcasm.
- **H2: observers’ failure to detect annoyance:** A failure by observers to identify the annoyance a speaker felt in a given situation causes the misaligned sarcasm judgment between speakers and observers.

Table 1 shows the cases from the data that exemplify either **H1** or **H2**. In both cases, sarcasm has failed to be communicated, since the ratings given by the speakers and observers are substantially different (6 vs. 1). However, in each case, different factors stand out as being associated with the failure. In **H1**, we hypothesize that the gap between the sarcasm rating and annoyance rating given by the speaker (6 vs. 2) may be associated with the failure of communicating sarcasm (*speaker’s annoyance-sarcasm incongruity*). In **H2**, we hypothesize that the discrepancy between the annoyance ratings given by the speaker and observer (1 vs. 6) is linked to the failure of communication (*observers’ failure to identify speaker’s annoyance*).

## 4 Experiment 1: Sarcasm detection by human observers

The first experiment inspected the factors related to sarcasm communication failure between human speakers and observers, by testing two hypotheses described in Section 3.2.

### 4.1 Quantifying variables

**Sarcasm alignment:** We quantified the alignment between the sarcasm scores given by a speaker and multiple observers using the inverse of normalized

mean absolute error (MAE).<sup>4</sup> We define alignment as:

$$1 - \frac{1}{n} \sum_{i=1}^n |y - \hat{y}|$$

where  $y$  is the rating by the speaker,  $\hat{y}$  is the rating from an observer for the same instance, and  $n$  is the total number of speaker-observer pairs considered. Values closer to 1 indicate stronger agreement. Sarcasm alignment score is the dependent variable in our following statistical analysis.

**Annoyance alignment:** The alignment between annoyance scores given by a speaker and multiple observers is computed using the same formula as above. This measure is one of the main predictors in our statistical analyses.

**Speaker congruity:** To quantify the congruity between sarcasm and annoyance expressed by the speaker we assigned a value of 1 (*congruous*) if the speaker rated sarcasm and annoyance levels as both negative (1-not at all, 2-mostly not, 3-not so much) or positive (4-somewhat, 5-mostly, 6-completely). If the speaker gave a negative rating to sarcasm (1, 2, 3) but a positive rating to annoyance (4, 5, 6), and vice versa, we assigned a value of 0 (*incongruous*). Together with annoyance alignment, this is the second main predictor in our statistical analysis.

	SP	OB1	OB2	OB3	OB4	OB5	OB6	Avg	Alignment
Ex.1	4	5	4	5	4	4	1	3.83	0.86
Ex.2	4	5	6	4	3	2	3	3.83	0.81

Table 2: Examples with the same average score (Avg) but with different alignment scores between speaker and observers.

<sup>4</sup>Though a conventional measure for quantifying errors is the mean squared error (MSE), the mean absolute error (MAE) aligns with the purpose of our task better, because the MAE does not penalize outliers among observers as harshly as the MSE. A single outlier is not much of a communication failure as long as the majority of the observers make judgments similar to the speaker’s original intention.

The two examples in Table 2 have the same average score, but in Example 1, most observers agreed with the speaker except for one major outlier, while Example 2 shows less alignment overall between the observers and the speaker. Therefore, Example 1 gets a higher alignment score of 0.86 and Example 2 gets a lower score of 0.81.

We tested our hypothesis using a linear mixed-effects model (Barr et al., 2013) that predicted **sarcasm alignment** given the **annoyance alignment** in interaction with **speaker’s annoyance-sarcasm congruity**, with by-item and by-participant random intercepts.

## 4.2 Results

The speaker’s annoyance-sarcasm **congruity** showed a statistically significant positive effect on speaker-observer **sarcasm alignment** ( $\beta = 0.15, p < 0.001$ ). Importantly, we found a strong positive interaction effect between the two predictors: In cases where the speaker’s annoyance-sarcasm **congruity** was preserved, the **annoyance alignment** judgment between speakers and observers led to higher **sarcasm alignment** judgment ( $\beta = 0.42, p < 0.001$ ). However, when this congruity was not maintained, the observer’s correct identification of speaker’s annoyance no longer contributed to the alignment in sarcasm judgment between speaker and observers.

To summarize, when the utterance of the speaker does not seem matched with the level of annoyance they may have felt in that context, observers are more likely to provide a sarcasm judgment that diverges from the speaker’s own judgment (**H1**). If the speaker’s underlying annoyance is congruous with their sarcastic utterance, the correct identification of speaker’s annoyance by observers helps align observers’ judgment of sarcasm with that of the speaker (**H2**). Therefore, in the next sections, based on **H1**, we conduct experiments using LLMs to examine the influence of speaker’s annoyance-sarcasm **congruity** on sarcasm detection.

## 5 Experiment 2: Sarcasm detection by LLMs

In Section 4, we showed that speaker’s annoyance-sarcasm congruity is highly correlated with sarcasm being correctly transmitted to human observers. Based on these results, we examined whether the same factor influences the sarcasm detection performance of LLMs. We conducted a classification experiment with encoder-only mod-

els and decoder-only models. The encoder-only models are classical observer models suitable for the task of sarcasm detection, and the decoder-only models are generative models that have shown their impressive capabilities to handle numerous NLP tasks. We used the fine-tuning settings for the encoder-only models, because they tend to require task-specific tuning to ensure a reasonable level of performance (Lyu et al., 2024). We used zero-shot settings for the generative models, without additional fine-tuning that requires substantial computational resources. We binary-coded the original sarcasm ratings by both sources – speakers and observers (averaged) – by using the midpoint of the scale (3.5) as the cut-off point. We downsampled CSC to have an equal number of sarcastic and non-sarcastic instances (N = 2,210 vs. 2,398).

### 5.1 Encoder-only models

We fine-tuned `bert-base-uncased` (110M parameters; Devlin et al., 2019) and `roberta-base` (125M parameters; Liu et al., 2019) on CSC to perform binary sarcasm classification (See Appendix A for setup details). For each language model, we obtained predictions on the test set.

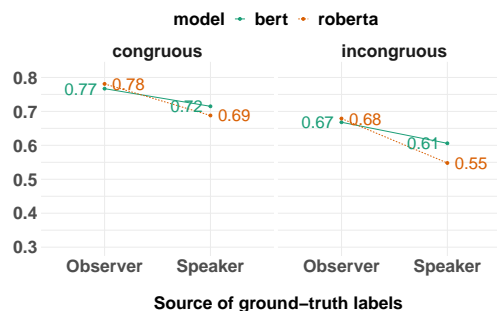


Figure 1: Macro F-scores (y-axis) for sarcasm detection by encoder-only models, according to ground-truth labels by observers or speakers. Results shown in two conditions - instances with speaker’s annoyance matching the level of sarcasm (congruous) or not (incongruous).

Figure 1 shows sarcasm prediction performance by encoder-only models given the different conditions of speakers’ annoyance-sarcasm congruity. In general, the instances for which speaker’s annoyance level was not matching the level of sarcasm of their subsequent responses (incongruous) show lower F-scores. These results are compatible with the results about human observers described in Section 4: Cases in which sarcasm fails to be communicated to observers are related to cases in which the speaker says something that is disproportion-

ate to their emotional motivation (low annoyance-sarcasm congruity). Likewise, also for encoder-only models, sarcasm is more difficult to detect when the speaker’s annoyance level is unmatched with the output utterance. We further find that these models show better detection with observer ground-truth labels than speaker ground-truth labels, which suggests their inclination to play the observer’s role (Jang and Frassinelli, 2024).

## 5.2 Decoder-only models

We prompted smaller ( $\approx 3\text{B}$  parameters) and bigger ( $\approx 7\text{-}8\text{B}$  parameters) open-source instruction-tuned generative LLMs, in zero-shot settings<sup>5</sup>: Llama3.2-3B, Llama3.1-8B, Qwen2.5-3B, Qwen2.5-7B. (See Appendix B for full prompts).

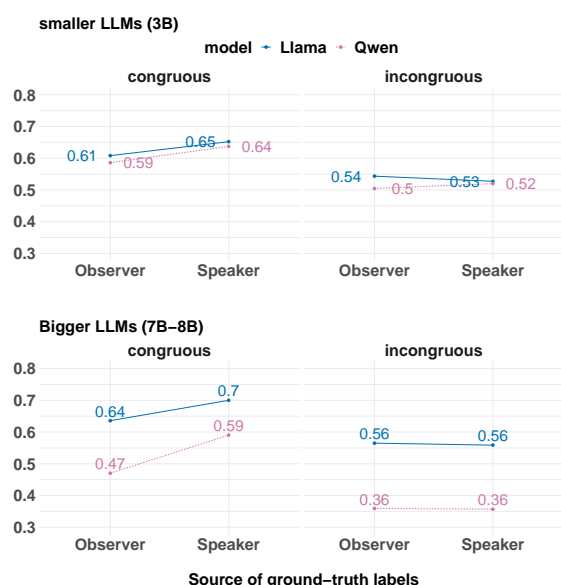


Figure 2: F-scores for sarcasm detection by generative LLMs in zero-shot settings (y-axis). Models with 3B parameters (top) or 7-8B parameters (bottom), according to ground-truth labels by observers or speakers. Results shown in two conditions - instances with speaker’s annoyance matching the level of sarcasm (congruous; left) or not (incongruous; right).

Figure 2 shows sarcasm prediction performance by the generative models given different conditions of speakers’ annoyance-sarcasm congruity. The best performing model is Llama3.1-8B, with speaker ground-truth, in the congruous condition. In general, all the generative models also struggle to detect sarcasm when the utterance is incongruous

<sup>5</sup>Though we also experimented with few-shot prompting, we only report results from the zero-shot experiments, as the results were comparable in both settings.

with the speaker’s annoyance level, in line with the previous results from Section 5.1.

On the other hand, we see an interesting difference between the generative models and the encoder-only models. In the congruous condition, the generative models perform better with speaker ground-truth than observer ground-truth (e.g., F-scores of 0.64 vs. 0.59 for Qwen2.5-3B in Figure 2). This is in contrast to the results in Section 5.1, in which the encoder-only models perform better with observer ground-truth (F-scores of 0.77 vs. 0.72 for bert-base-uncased in Figure 1). In the incongruous condition, though, the performance of the generative models drops to about the same level between speaker ground-truth and observer ground-truth (e.g., F-score of 0.56 for observer and speaker ground-truth for Llama3.1-8B in Figure 2).

We observe model-specific variations as well. The Llama3.1-8B performs better than its smaller version Llama3.2-3B, whereas Qwen2.5-7B underperforms its smaller version Qwen2.5-3B. Also, between the congruous versus incongruous conditions, the performance drop by Qwen2.5-7B is steeper (0.59 to 0.36) than that by Qwen2.5-3B (0.64 to 0.52), which suggests its relatively lower robustness against speaker’s incongruity.

## 6 Decoder-only vs. encoder-only: Identification with speaker’s perspective

In Sections 6 and 7, we conduct further experiments to examine the difference between encoder-only and decoder-only models. In Section 5.2, the decoder-only generative models showed better performance with speaker ground-truth labels than observer ground-truth labels in the congruous condition. This pattern is in contrast to the pattern we observed with encoder-only models, which demonstrated better performance with labels judged by the observers (Section 5.1).

One possible explanation for such difference is that generative models are more sensitive in interpreting speakers’ “point-of-view” than encoder-only models. In Section 5.2, the prompt for the generative models asking “how sarcastic is someone’s response” could have biased the models to take the speaker’s perspective by default. We investigated whether prompting the LLMs with more explicit instructions to take the perspective of an observer

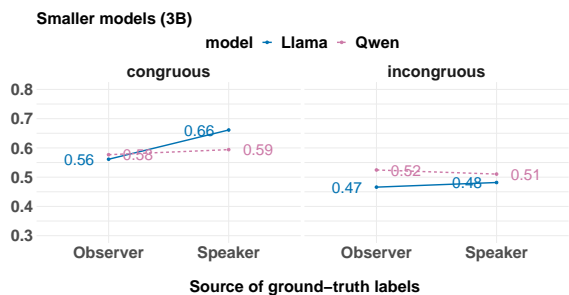


Figure 3: Macro F-scores (y-axis) for sarcasm detection by generative LLM models when explicitly prompted to take observer’s perspective. Results shown in two conditions (congruous vs. incongruous).

would provide more information about why the generative LLMs perform better with speaker ground-truth than observer ground-truth (See Appendix B for the full prompt). We prompted Llama3.2-3B and Qwen2.5-3B with the new prompt.

Figure 3 shows the prediction results of Llama3.2-3B and Qwen2.5-3B explicitly prompted to take the perspective of an external observer. For both models, the prediction performance does not increase with the new prompting method. When we manually inspect the responses by the LLMs, the general tendency of these models is that they provide plenty of descriptions about the emotions that the speaker would be experiencing, or the message that the speaker is trying to express (see Appendix C for sample responses by the LLMs). Even when instructed to take the observers’ perspective, the models still focus on the speaker’s experience in the conversation, and use the conclusion about this as a basis to determine an external observer’s sarcasm judgment.<sup>6</sup> Given these results, we detect a tendency that generative models identify with the speaker’s perspective by default rather than observer’s perspective.

## 7 Decoder-only vs. encoder-only: Sensitivity to disagreement among observers

In investigating the reason why generative models perform better with speaker ground-truth labels, another possibility is that generative models are more sensitive to the disagreement among multiple observers and therefore may struggle to per-

<sup>6</sup>Though a human observer is also expected to speculate about the speaker’s emotions and communicative intentions before judging the level of sarcasm in their utterances, we think this may partially depend on the theory of mind capacity, which varies across individuals (Zhu and Wang, 2020).

form at their best when the ground-truth labels are the result of simple averaging. If true, the sensitivity would have influenced the results in Section 5.2, in which the sarcasm labels by observers were averaged and binary-coded, which discards information about potential disagreement among the observers. Annotator-wide disagreement in general is considered as an important topic in NLP, as ground-truth labels cannot always come down to one single judgment (Cabitza et al., 2023; Knupleš et al., 2023; de Marneffe and Manning, 2012; Plank et al., 2014; Weber-Genzel et al., 2024). To test this, we examined whether the disagreement among the observers influences the performance of the generative LLMs more than that of the encoder-only models.

We quantified the level of disagreement among different observers using the normalized MAE described in Section 4.1. For the purpose of visual inspection, we split the scores we obtained using this formula at the mean value into *low* versus *high*. We inspected F-scores of both encoder-only models and generative models in the two groups of disagreement (low vs. high).

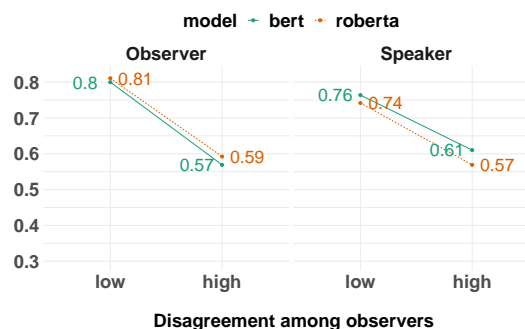


Figure 4: Macro F-scores (y-axis) for sarcasm detection by encoder-only models, per ground-truth labels by observers (left panel) and speakers (right panel). Results divided according to disagreement among observers (low vs. high).

Both encoder-only models and generative models show better performance when human annotators agree on the sarcasm judgment (Figures 4 and 5). For encoder-only models, the difference in F-score between the two groups (low vs. high) is comparable for both ground-truth labels (e.g., 0.80 vs. 0.57 in the left panel  $\approx$  0.76 vs. 0.61 in the right panel of Figure 4). In contrast, for all generative models, the difference in F-score is larger for observer ground-truth when the disagreement is low versus high (e.g., 0.71 vs. 0.53 on bottom left panel in Figure 5), than it is for speaker ground-truth

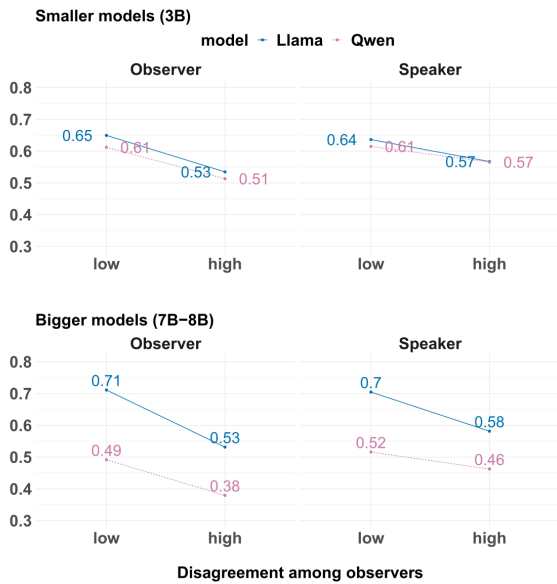


Figure 5: Macro F-scores (y-axis) for sarcasm detection by generative models, per ground-truth labels by observers (left panel) and speakers (right panel). Results divided by disagreement among observers (low vs. high).

(0.70 vs. 0.58 on bottom right panel in Figure 5). This difference is bigger for LLMs with a larger number of parameters, which suggests their higher sensitivity to disagreement among observers.

These results suggest that both types of language models are influenced by the disagreement among the observers. But in the face of this challenge, the generative models, especially those with more parameters, show a somewhat higher sensitivity by reacting against observer ground-truth to a greater extent than speaker ground-truth, which is in principle an expected behavior (i.e., in an ideal scenario, the performance with speaker ground-truth should see no change). This contrasts with encoder-only models, which show an equal drop against both sources of ground-truth labels and therefore demonstrates fragility to challenges stemming from human disagreement.

## 8 Experiment 3: Sarcasm detection by LLMs with additional information

Sections 4 and 5 showed that *speaker’s annoyance-sarcasm congruity* influences the judgment of sarcasm both by human observers and LLMs. Here we tested whether adding information about the speaker’s annoyance to LLMs would then improve the classification results.

### 8.1 Encoder-only models

We added information about speaker’s annoyance in the form of logits to sarcasm detection models. We assessed if the added information contributes to better sarcasm detection to different degrees in congruous versus incongruous conditions. We fine-tuned `bert-base-uncased` on CSC for annoyance detection (*annoying* vs. *not annoying*). We obtained the prediction logits for annoyance on the test set, and concatenated them to the embeddings obtained from the sarcasm detection models described in Section 5.1. This concatenation strategy was inspired from the experiment in Yeo et al. (2024), which combined information about multiple dimensions into a single prediction model based on theoretical grounds. We made sure that sarcasm fine-tuning and annoyance fine-tuning would be done with the same training and test split settings to avoid the models being exposed to the same fine-tuning data for annoyance detection and sarcasm detection. We used the fine-tuned sarcasm detection models to extract embeddings as text representation, to which we added annoyance information in the form of logits. We then used a logistic regression classifier (with a ‘liblinear’ solver that works better for high-dimension data, and the maximum iteration of 500) on the remaining test set with a 5-fold cross-validation.

Table 3 shows the results on sarcasm classification and the improvement in performance with the addition of annoyance information. Additional annoyance information is not helpful for the encoder-only models when predicting sarcasm based on observer labels, regardless of the congruity between the sarcasm and the underlying annoyance. In contrast, when the models predict sarcasm based on speaker labels, in the incongruous condition, helping the models with additional annoyance information leads to better results (5-6%).

### 8.2 Decoder-only models

We prompted `Llama3.2-3B`, `Llama3.1-8B`, `Qwen2.5-3B` and `Qwen2.5-7B` with direct information about speaker’s annoyance level (See Appendix B for the full prompt).

The patterns by which the added information about the speaker’s annoyance helped the models varied across LLMs (See Table 3). `Qwen2.5-3B` showed comparable patterns with encoder-only models, in which adding the annoyance information increased the F-score for speaker ground-truth by a bigger margin (+0.12/+0.14) compared to ob-

G.T	Congruity	Encoder-only		Generative			
		BERT	RoBERTa	Llama-3B	Llama-8B	Qwen-3B	Qwen-7B
Speaker	Congruous	+0.01	+0.02	-0.04	+0.02	<b>+0.12</b>	<b>+0.05</b>
	Incongruous	<b>+0.05</b>	<b>+0.06</b>	+0.02	-0.07	<b>+0.14</b>	-0.05
Observer	Congruous	+0.00	-0.01	-0.01	+0.02	<b>+0.05</b>	+0.00
	Incongruous	+0.00	+0.00	+0.02	-0.02	<b>+0.07</b>	-0.03

Table 3: Improvement in F-score for sarcasm detection performance by different LLMs when annoyance information was additionally supplied in the form of logits (encoder-only models) and prompting (generative models). Improvement of 5% and higher marked in bold.

server ground-truth (+0.05/+0.07), and the increase being higher for incongruous condition (+0.14) than the congruous condition (+0.12). None of the other models show any consistent improvement when information about the underlying annoyance of the speakers was supplied in the prompt. We suspect that it may be because models with a larger number of parameters are less likely to be influenced by added information from one dimension only (annoyance). Nevertheless, given the inconclusive results of this experiment, further examination would be needed about the influence of assistive information for LLM performance.

## 9 General discussion

When speakers use sarcasm without any noticeable emotional cues, external observers lose an important source of information for judging the level of sarcasm in the provided utterance (**H1**). This tendency in human observers is also reflected in LLMs. For both humans and LLMs, sarcasm is difficult to detect when the speakers’ annoyance seems unmatched with the output utterance. In contrast, in sarcastic utterances where proportional annoyance can be perceived as an underlying motivation, models are better at detecting sarcasm. Some differences are observed between encoder-only models and generative models, in terms of which source of ground-truth labels (speaker vs. observer) they match better. Encoder-only models show better performance with observer ground-truth, in line with prior work (Abu Farha et al., 2022; Jang and Frassinelli, 2024). However, generative models show better performance with speaker ground-truth. A further analysis suggests that generative models may impersonate speakers’ perspective by default compared to encoder-only models. This aligns with the capabilities that these models are expected to have, exemplified by one of the evaluation suites for Llama-3 models “inhabiting a character/persona”.<sup>7</sup>

<sup>7</sup><https://ai.meta.com/blog/meta-llama-3/>

Nevertheless, speaker’s incongruity in their underlying emotion and utterance still poses a challenge for LLMs. This is a factor worth considering for the inspection of linguistic competence of LLMs, because investigating a linguistic output by humans often requires understanding the factors that led up to it (e.g., speaker’s motivation). Another obstacle that hinders good performance by LLMs is higher disagreement among observers (annotators). It is important for the evaluation of LLM capabilities to investigate the patterns by which LLMs navigate through varying linguistic judgments by humans, especially on heavily subjective topics such as sarcasm. The findings in this work also suggest that future research should address both perspectives of conversational partners (speaker vs. listener) when evaluating LLM output. Examining which perspective is reflected in the output of LLMs would help understand the competence of LLMs in more depth.

## 10 Conclusion

We showed that speaker’s incongruity between their utterance and the annoyance they felt is associated with their judgment of sarcasm diverging from the judgment by external observers. This factor, as well as disagreement among the observers, also presented challenges to language models (both generative and encoder-only). Lastly, we discovered that the generative models are more likely to impersonate speakers’ perspective more than observers’ perspective, in contrast to encoder-only models.

## Limitations

The main limitation of this work is that only two factors were considered as intermediary elements contributing to sarcasm failure, as mentioned in Section 3.2. We acknowledge that sarcasm can fail to be communicated for several reasons other than the mismatch between annoyance and sarcasm, given its complexity mentioned in the literature e.g., Gibbs and Colston (2023). Examining more factors



such as multimodal and other contextual factors in addressing the causes for failure of sarcastic communication is left to future work.

Another limitation of this work is that only one dataset was used for our experiments, because this is the only dataset we found to have both speaker and observer labels on multiple related dimensions (e.g., sarcasm and annoyance). Replications of our findings with other datasets and topics would strengthen our findings about communication failure in general. The use of larger language models than reported in this paper may also be considered for more conclusive insights about this topic.

Lastly, the ways of integrating annoyance information to LLMs were limited. Annoyance information was integrated with a simple concatenation of embeddings and logits for encoder-only models, and with prompting for generative models. With the results from this preliminary work, other forms of information integration will need to be tested in future work.

## Acknowledgments

We are grateful to Bettina Braun for her advice on the methodology for the initial version of this work. The dissemination of this work has been supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg, Sweden.

## Ethics Statement

We see little ethical issue related to this work. All our modeling experiments were conducted with open-source libraries, which received due citations. We did not rely on any AI-assistant tools for manuscript creation. But we acknowledge that at times, sarcasm in itself can be a sensitive topic including offensive language and content depending on the circumstances.

## References

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. [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.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for

confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Gregory A. Bryant. 2023. Vocal strategies in verbal irony. In Raymond W. Gibbs and Herbert L. Colston, editors, *The Cambridge Handbook of Irony and Thought*, pages 197–215. Cambridge University Press, New York.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. 37:6860–6868.

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.

Herbert L. Colston. 2023. Irony as social work: Opposition, expectation violation, and contrast. In Raymond W. Gibbs and Herbert L. Colston, editors, *The Cambridge Handbook of Irony and Thought*, pages 81–95. Cambridge University Press, New York.

Tanvi Dadu and Kartikey Pant. 2020. [Sarcasm detection using context separators in online discourse](#). In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 51–55, Online. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. 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.

Shelly Dews, Joan Kaplan, and Ellen Winner. 1995. Why not say it directly? The social functions of irony. *Discourse Processes*, 19(3):347–367.

Ruth Filik. 2023. Emotional responses to sarcasm. In Raymond W. Gibbs and Herbert L. Colston, editors, *The Cambridge Handbook of Irony and Thought*, pages 255–271. Cambridge University Press, New York.

Jean E. Fox Tree, J. Trevor D’Arcey, Alicia A. Hammond, and Alina S. Larson. 2020. The sarcasm: Sarcasm production and identification in spontaneous conversation. *Discourse Processes*, 57(5-6):507–533.

- Raymond W. Gibbs. 2000. Irony in Talk Among Friends. *Metaphor and Symbol*, 15(1-2):5–27.
- Raymond W. Gibbs and Herbert L. Colston. 2023. Irony and thought: The state of the art. In Raymond W. Gibbs and Herbert L. Colston, editors, *The Cambridge Handbook of Irony and Thought*, pages 3–14. Cambridge University Press, New York.
- Hyewon Jang, Bettina Braun, and Diego Frassinelli. 2023. Intended and perceived sarcasm between close friends: What triggers sarcasm and what gets conveyed? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.
- 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.
- 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. European Language Resources Association (ELRA).
- Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde. 2023. Investigating the nature of disagreements on mid-scale ratings: A case study on the abstractness-concreteness continuum. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 70–86, Singapore. Association for Computational Linguistics.
- Amardeep Kumar and Vivek Anand. 2020. [Transformers on sarcasm detection with context](#). In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 88–92, Online. 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*.
- Youngang Lyu, Lingyong Yan, Shuaiqiang Wang, Haibo Shi, Dawei Yin, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2024. KnowTuning: Knowledge-aware fine-tuning for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14535–14556, Miami, Florida, USA. Association for Computational Linguistics.
- Marie-Catherine de Marneffe and Christopher D. Manning. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, (2):301–333.
- Rishabh Misra and Prahal Arora. 2023. Sarcasm detection using news headlines dataset. *AI Open*, 4:13–18.
- Silviu Oprea and Walid Magdy. 2019. [Exploring author context for detecting intended vs perceived sarcasm](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.
- 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*, pages 1279–1289, Online. 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.
- Penny M. Pexman and Kara M. Olineck. 2002. Does Sarcasm Always Sting? Investigating the Impact of Ironic Insults and Ironic Compliments. *Discourse Processes*, 33(3):199–217.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. [Linguistically debatable or just plain wrong?](#) In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.
- Joan Plepi and Lucie Flek. 2021. Perceived and intended sarcasm detection with graph attention networks. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 97–105, Online. Association for Computational Linguistics.
- Boaz Shmueli, Lun-Wei Ku, and Soumya Ray. 2020. [Reactive Supervision: A New Method for Collecting Sarcasm Data](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2553–2559, Online. Association for Computational Linguistics.
- 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*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.
- Tony Veale. 2023. Great expectations and epic fails: A computational perspective on irony and sarcasm. In Raymond W. Gibbs and Herbert L. Colston, editors, *The Cambridge Handbook of Irony and Thought*, pages 225–234. Cambridge University Press, New York.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Gerard Yeo, Shaz Furniturewala, and Kokil Jaidka. 2024. [Beyond text: Leveraging multi-task learning and cognitive appraisal theory for post-purchase intention analysis](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12353–12360, Bangkok, Thailand. Association for Computational Linguistics.

Ning Zhu and Zhenlin Wang. 2020. The paradox of sarcasm: Theory of mind and sarcasm use in adults. *Personality and Individual Differences*, 163:110035.

## A Fine-tuning details for Section 5.1

We used 80% of the downsampled data for fine-tuning in a 5-fold split (model initialization seeds: 10, 20). We fine-tuned each model for 2 epochs with a batch size of 16, and a learning rate of 5e-5. The fine-tuning was implemented using the Transformers class from the Hugging Face library, and was conducted on an NVIDIA A100 GPU with a total memory of 40GB. The reported results in Section 5.1 are the average of all seeds and folds.

## B Prompts for generative models

The prompting was implemented using the Transformers class from the Hugging Face library, and was conducted on an NVIDIA A100 GPU with a total memory of 40GB.

### B.1 Base prompt

*“Analyze each situation between Steve and John. In each situation, judge if John’s response is sarcastic or not. Say 1 if John’s response is sarcastic and 0 if it is not sarcastic. Only respond with 1 or 0 and nothing else: [context]. And John says [response].”*

### B.2 Prompt with explicit instructions to take the observer’s perspective

*\*\*We also used a prompt with the proper name of Sarah removed, and the results were comparable to using the above prompt.*

*“Sarah is analyzing conversations between Steve and John. In each situation, Sarah will judge if John’s response is sarcastic or not. Sarah will say 1 if John’s response is sarcastic and 0 if it is not sarcastic. What would Sarah say in this situation, 1 or 0?: [context]. And John says [response].”*

### B.3 Prompt with annoyance information

*“Read each situation between Steve and John. Judge if John’s response is sarcastic or not. Say 1 if John’s response is sarcastic and 0 if it is not sarcastic: [context]. And John says [response]. In this situation, on a scale of 1 to 6, John’s annoyance was at [annoyance-rating].”*

## C Reasoning provided by LLMs in Section 6

### Llama3.2-3B:

- In this situation, Sarah would say 0. John’s response seems to be a supportive and understanding comment, rather than a sarcastic one. He’s acknowledging Steve’s frustration and offering a sympathetic ear, which suggests that John is not being sarcastic.
- Based on the conversation, Sarah would say 1. Although John’s response seems to be a straightforward explanation (“it’s just because I want to pick your girlfriend up later!”), the initial response (“come on, why can’t I borrow it? Don’t you trust me?”) has a tone of sarcasm and implies that John is questioning Steve’s trust, which is a common characteristic of sarcastic responses.
- In this situation, Sarah would say 0. Although John’s response seems to be slightly negative towards social media, it does not appear to be sarcastic. John’s statement “Life’s too short” is a genuine expression of his opinion, and his offer to help Steve with posting a picture is not sarcastic.

### Qwen2.5-3B:

- In this scenario, John’s statement “In about an hour” does not appear to be sarcastic. Steve’s initial response indicates that he is genuinely interested in joining the party and wants to know when it starts. John’s reply simply confirms the timing of the party without any indication of sarcasm. Therefore, Sarah would say: 0.
- In this scenario, John’s response seems genuine rather than sarcastic. John is complimenting the watch and even going so far as to suggest that it has become even more shiny since he last saw it. This implies that he finds the watch impressive and possibly in better condition now. Therefore, Sarah would judge this response as **\*\*0\*\*** (not sarcastic).
- In this situation, John’s response can be interpreted as somewhat sarcastic. John acknowledges Steve’s offer to lend him a shirt but then suggests he already has another shirt that isn’t

tight. This could imply that John feels his current shirt is tight and that he doesn't need to accept the offer, which might be seen as a subtle rejection or at least an attempt to downplay Steve's offer. Given this context, Sarah would likely judge John's response as sarcastic. So, Sarah would say \*\*1\*\*.