Universität
Konstanz

Challenges in sarcasm handling by language models (and humans)

Hyewon Jang

Department of Linguistics, University of Konstanz

May 21, 2025

Introduction

Quiz 1. Is the following utterance sarcastic?

> *"It was such a pleasant sight to see a guy picking up used chewing gum, and he put it in his mouth."*

Introduction

Quiz 2. Is the following utterance sarcastic?

> *“Is the present inside the water can?”*

Introduction

Quiz 2.2. What about now?

> *Steve gives you a watering can on your birthday while smiling at you with a strange expression. But you don't even have a single plant.*
>
> *"Is the present inside the water can?"*

Introduction

Quiz 3.1. On a scale of 1 to 6, how sarcastic is the following utterance?

1 (not at all) – 2 (mostly not) – 3 (not so much) – 4 (somewhat) – 5 (mostly) – 6 (completely)

> *So the Scottish Government want people to get their booster shots so badly that the website doesn't even work.*

Introduction

Quiz 3.2. On a scale of 1 to 6, how sarcastic is the following utterance?

1 (not at all) – 2 (mostly not) – 3 (not so much) – 4 (somewhat) – 5 (mostly) – 6 (completely)

> *No thanks. There are other ways to meet dates. It's very easy for gays to meet dates that are not officially gay.*

Introduction

- Definition of sarcasm: The utterance of the opposite of the intended meaning (Glucksberg, 1995).

Introduction

- Definition of sarcasm: The utterance of the opposite of the intended meaning (Glucksberg, 1995).
- Operationalized definition in CL/NLP: "saying the opposite of the true message, often with the intent to be hurtful" (Cai et al., 2019; Frenda et al., 2022; A. Ghosh & Veale, 2017; Joshi et al., 2015; Pan et al., 2020).

Introduction

- Definition of sarcasm: The utterance of the opposite of the intended meaning (Glucksberg, 1995).
- Operationalized definition in CL/NLP: "saying the opposite of the true message, often with the intent to be hurtful" (Cai et al., 2019; Frenda et al., 2022; A. Ghosh & Veale, 2017; Joshi et al., 2015; Pan et al., 2020).
- Sarcasm detection work - heavy focus on data from social media (Abu Farha et al., 2022; Barbieri et al., 2014; Joshi et al., 2015; Khodak et al., 2018; Ptacek et al., 2014; Van Hee et al.,2018).

Introduction

- Definition of sarcasm: The utterance of the opposite of the intended meaning (Glucksberg, 1995).
- Operationalized definition in CL/NLP: “saying the opposite of the true message, often with the intent to be hurtful” (Cai et al., 2019; Frenda et al., 2022; A. Ghosh & Veale, 2017; Joshi et al., 2015; Pan et al., 2020).
- Sarcasm detection work – heavy focus on data from social media (Abu Farha et al., 2022; Barbieri et al., 2014; Joshi et al., 2015; Khodak et al., 2018; Ptacek et al., 2014; Van Hee et al.,2018).
- Computational work does not appear to grasp the essence of the complicated nature of sarcasm, which psycholinguistic work addresses extensively.

Framework

Overarching goal: To ameliorate ...

1. the limited focus on specific types of sarcasm data

Framework

Overarching goal: To ameliorate ...

1. the limited focus on specific types of sarcasm data
2. the lack of integration of prior (psycholinguistic) knowledge about sarcasm.

Framework

Our framework

- Find empirical evidence for the reason why humans speak sarcastically.

Framework

Our framework

- Find empirical evidence for the reason why humans speak sarcastically.
- By doing so, collect a large enough dataset that is psycholinguistically motivated.

Framework

Our framework

- Find empirical evidence for the reason why humans speak sarcastically.
- By doing so, collect a large enough dataset that is psycholinguistically motivated.
- Use the data to examine how language models process sarcasm.

Sarcasm use by humans

Focus 1: Sarcasm production

RQ. What contextual factors motivate speakers to use sarcasm?

Hypothesis. Certain contexts $\implies$ certain emotional reactions $\implies$ sarcasm.

Focus 1: Sarcasm production

Situation: Speaker A asks you to get him some coffee from a coffee shop nearby. But he doesn't seem to be particularly busy right now. When you ask him "why can't you get it yourself?"...

Focus 1: Sarcasm production

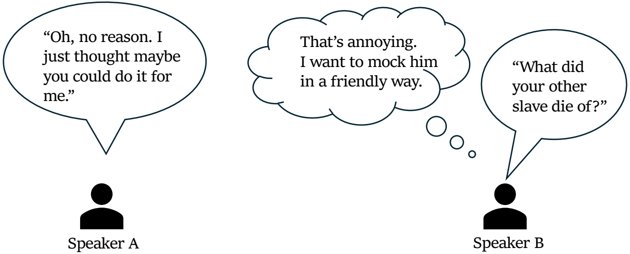
Situation: Speaker A asks you to get him some coffee from a coffee shop nearby. But he doesn't seem to be particularly busy right now. When you ask him "why can't you get it yourself?"...

Focus 1: Sarcasm production

Situation: Speaker A asks you to get him some coffee from a coffee shop nearby. But he doesn't seem to be particularly busy right now. When you ask him "why can't you get it yourself?"...

Focus 1: Sarcasm production

Situation: Speaker A asks you to get him some coffee from a coffee shop nearby. But he doesn't seem to be particularly busy right now. When you ask him "why can't you get it yourself?"...



Focus 1: Sarcasm production

Experimental design

Stimulus: [A situation similar to what we just saw.]

Task 1: Freely respond to the interlocutor (Free text).

Task 2: Answer the following questions (Likert scale/multiple choice).

1	How silly or annoying did you find the interlocutor?
2	How sarcastic is your response?
3	What were your intentions with your utterance?

Focus 1: Sarcasm production

Stimuli and participants

	# stimuli	# participants
Step 1	32	60
Step 2	40	128

- All experiments were set up on FindingFive.
- All participants were recruited online (Prolific).
- All participants were native English speakers (gender-balanced).

Focus 1: Sarcasm production

Dependent variable: sarcasm ratings (collected)

Focus 1: Sarcasm production

Independent variables

	Independent variables		
	Manipulated	**Collected**	
Step 1	Context Types (2 levels)	Affect (silly/annoying)	Intentions
Step 2	Context Types (5 levels)	Affect (funny)	Affect (annoying)

- All collected variables (-intentions): on a 1 (*not at all*) - 6 (*completely*) scale and z-normalized ($m = 0, sd = 1$) for analysis

Focus 1: Sarcasm production

Speaker intentions

Intentions	
To criticize interlocutor harsher	To criticize interlocutor softer
To mock interlocutor hilariously	To mock interlocutor friendly
To appear clever	To be direct
To be nice	To be natural

Multiple-choice selection from 8 given options
(0 vs. 1 for each intention; multiple selection possible).

Focus 1: Sarcasm production

Analysis using linear mixed-effect model (LMER):

predict the level of sarcasm given the predictors, accounting for random effects from stimuli and participants.

Focus 1: Sarcasm production

Results

1. Emotions:

Focus 1: Sarcasm production

Results

1. Emotions:
    - Silliness/annoyance to the context $\Longrightarrow$ sarcastic responses (Step 1: $p < 0.001$).

Focus 1: Sarcasm production

Results

1. Emotions:
   - Silliness/annoyance to the context $\Longrightarrow$ sarcastic responses (Step 1: $p < 0.001$).
   - Annoyance $\Longrightarrow$ sarcastic responses (Step 2: $p < 0.001$).

Focus 1: Sarcasm production

Results

1. Emotions:
   - Silliness/annoyance to the context $\Longrightarrow$ sarcastic responses (Step 1: $p < 0.001$).
   - Annoyance $\Longrightarrow$ sarcastic responses (Step 2: $p < 0.001$).
   - Funniness boosts sarcastic responses for certain types of situations (Step 2).

Focus 1: Sarcasm production

Results

1. Emotions:
   - Silliness/annoyance to the context $\implies$ sarcastic responses (Step 1: $p < 0.001$).
   - Annoyance $\implies$ sarcastic responses (Step 2: $p < 0.001$).
   - Funniness boosts sarcastic responses for certain types of situations (Step 2).
2. Intentions:

Focus 1: Sarcasm production

Results

1. Emotions:
   - Silliness/annoyance to the context $\Longrightarrow$ sarcastic responses (Step 1: $p < 0.001$).
   - Annoyance $\Longrightarrow$ sarcastic responses (Step 2: $p < 0.001$).
   - Funniness boosts sarcastic responses for certain types of situations (Step 2).
2. Intentions:
   - Intent to mock $\Longrightarrow$ sarcastic responses ($p < 0.001$).

Focus 1: Sarcasm production

Results

1. Emotions:
   - Silliness/annoyance to the context $\implies$ sarcastic responses (Step 1: $p < 0.001$).
   - Annoyance $\implies$ sarcastic responses (Step 2: $p < 0.001$).
   - Funniness boosts sarcastic responses for certain types of situations (Step 2).
2. Intentions:
   - Intent to mock $\implies$ sarcastic responses ($p < 0.001$).
   - Intent to speak cleverly $\implies$ sarcastic responses ($p < 0.001$).

Focus 1: Sarcasm production

Results

1. Emotions:
   - Silliness/annoyance to the context $\Longrightarrow$ sarcastic responses (Step 1: $p < 0.001$).
   - Annoyance $\Longrightarrow$ sarcastic responses (Step 2: $p < 0.001$).
   - Funniness boosts sarcastic responses for certain types of situations (Step 2).
2. Intentions:
   - Intent to mock $\Longrightarrow$ sarcastic responses ($p < 0.001$).
   - Intent to speak cleverly $\Longrightarrow$ sarcastic responses ($p < 0.001$).
   - Intent to criticize $\not\Longrightarrow$ sarcastic responses ($p > 0.25$).

Focus 1: Sarcasm production

Hypotheses confirmed.

- Certain emotional reactions will trigger sarcasm. ✓

Focus 1: Sarcasm production

Hypotheses confirmed.

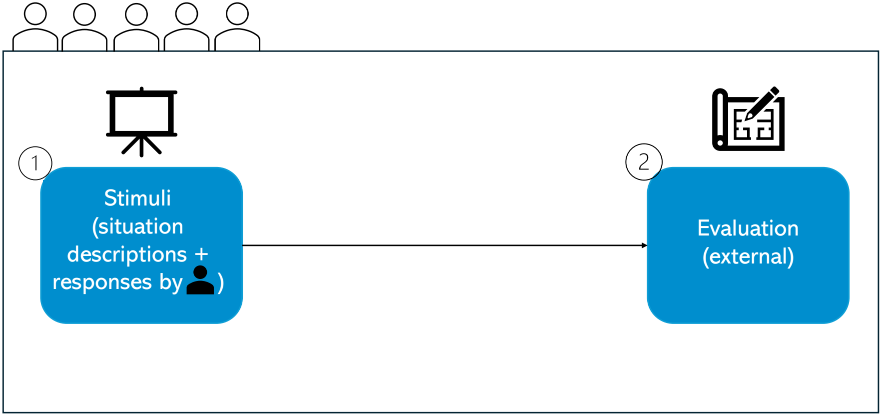
- Certain emotional reactions will trigger sarcasm. ✓
- Certain contexts will trigger such reactions, thereby causing more frequent sarcasm. ✓

Focus 2: Sarcasm comprehension

RQ. What commonalities, and what differences, do speakers and observers have when identifying a remark as sarcastic?

Focus 2: Sarcasm comprehension

Experimental design



Focus 2: Sarcasm comprehension

Task: Answer the following questions.

1	How silly or annoying did *the speaker* find the interlocutor?
2	How sarcastic is *the speaker's* response?
3	What were *the speaker's* intentions?

Focus 2: Sarcasm comprehension

Stimuli and participants

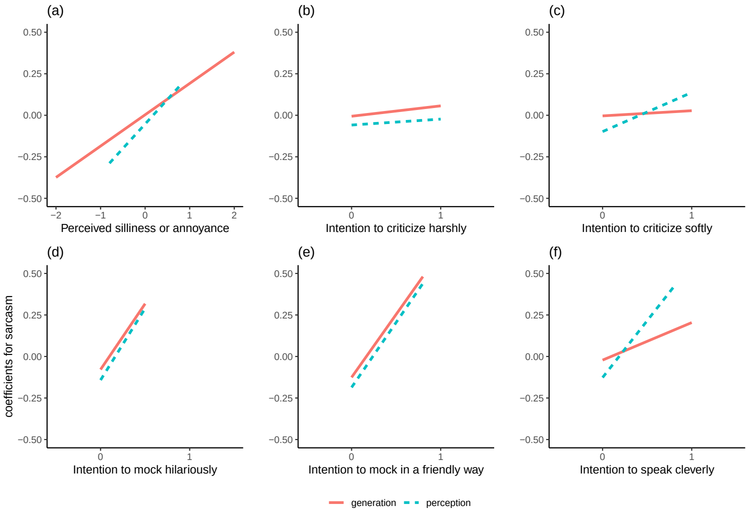
	# stimuli	# participants
Step 1	32	60×6
Step 2	40	128×4

Focus 2: Sarcasm comprehension

Analysis using LMER: predict the level of sarcasm given the predictors.

Focus 2: Sarcasm comprehension

Results



Sarcasm production & comprehension

Discussion – food for thought in the next section.

- The emotional reaction that a situation causes has a strong effect in triggering sarcasm.

Sarcasm production & comprehension

Discussion – food for thought in the next section.

- The emotional reaction that a situation causes has a strong effect in triggering sarcasm.
- Sarcasm is generally associated with negative attitudes (i.e., being upset or annoyed), but there is also an undertone of humor to it.

Sarcasm production & comprehension

Discussion – food for thought in the next section.

- The emotional reaction that a situation causes has a strong effect in triggering sarcasm.
- Sarcasm is generally associated with negative attitudes (i.e., being upset or annoyed), but there is also an undertone of humor to it.
- Observers can mostly infer the speaker's underlying motivation behind a sarcastic utterance, though not perfectly (i.e., critical intention).

Sarcasm use by humans

Publications:

1. Hyewon Jang, Bettina Braun, Diego Frassinelli, **Intended and Perceived Sarcasm Between Close Friends: What Triggers Sarcasm and What Gets Conveyed?**, *Proceedings of the 45th Annual Conference of the Cognitive Science Society (CogSci 2023)*.
2. Hyewon Jang, Bettina Braun, Diego Frassinelli, **Contextual Factors that Trigger Sarcasm**, *under final review at Metaphor and Symbol*.

Conversational Sarcasm Corpus (CSC)

Conversational Sarcasm Corpus (CSC)

As a result of four experiments...

Context: You are helping Steve move into a new apartment. After an hour, you realize that Steve is only carrying light stuff and you are doing all the heavy lifting. Steve says, “ugh, moving is always so stressful and chaotic...”

Response: “Yeah it is, especially when you are doing the bare minimum”

Conversational Sarcasm Corpus (CSC)

As a result of four experiments...

Context: You are helping Steve move into a new apartment. After an hour, you realize that Steve is only carrying light stuff and you are doing all the heavy lifting. Steve says, “ugh, moving is always so stressful and chaotic...”

Response: “Yeah it is, especially when you are doing the bare minimum”

- Sarcasm rating – speaker: 6

Conversational Sarcasm Corpus (CSC)

As a result of four experiments...

Context: You are helping Steve move into a new apartment. After an hour, you realize that Steve is only carrying light stuff and you are doing all the heavy lifting. Steve says, “ugh, moving is always so stressful and chaotic...”

> **Response**: “Yeah it is, especially when you are doing the bare minimum”

- Sarcasm rating – speaker: 6
- Sarcasm ratings – multiple observers: [4, 5, 4, 5, 5, 6]

Conversational Sarcasm Corpus (CSC)

As a result of four experiments...

Context: You are helping Steve move into a new apartment. After an hour, you realize that Steve is only carrying light stuff and you are doing all the heavy lifting. Steve says, “ugh, moving is always so stressful and chaotic...”

> **Response**: “Yeah it is, especially when you are doing the bare minimum”

- Sarcasm rating – speaker: 6
- Sarcasm ratings – multiple observers: [4, 5, 4, 5, 5, 6]
- Affect rating (silly-annoying) – speaker: 5

Conversational Sarcasm Corpus (CSC)

As a result of four experiments...

Context: You are helping Steve move into a new apartment. After an hour, you realize that Steve is only carrying light stuff and you are doing all the heavy lifting. Steve says, “ugh, moving is always so stressful and chaotic...”

> **Response**: “Yeah it is, especially when you are doing the bare minimum”

- Sarcasm rating – speaker: 6
- Sarcasm ratings – multiple observers: [4, 5, 4, 5, 5, 6]
- Affect rating (silly-annoying) – speaker: 5
- Presumed affect ratings (silly-annoying) – multiple observers: [6, 5, 4, 6, 5, 4]

Conversational Sarcasm Corpus (CSC)

CSC statistics

		Total	%
Speaker eval (bin)	Not sarc	4,826	69
	Sarc	2,210	**31**
Observer eval (bin)	Not sarc	4,638	66
	Sarc	2,398	**34**
Total # of context+utterance		7,036	

Challenge for LLMs 1: Generalizability of sarcasm detection

Challenge 1: Generalizability of sarcasm detection

RQ: "Can sarcasm detection models detect sarcasm of various styles?"

Challenge 1: Generalizability of sarcasm detection

Motivation:

- Prior work focuses on the critical aspect of sarcasm only (Frenda et al., 2022).

Challenge 1: Generalizability of sarcasm detection

Motivation:

- Prior work focuses on the critical aspect of sarcasm only (Frenda et al., 2022).
- Datasets of sarcasm contain different styles of sarcasm coming from different domains (Castro et al., 2019; Oprea & Magdy, 2019; Khodak et al., 2018).

Challenge 1: Generalizability of sarcasm detection

Motivation:

- Prior work focuses on the critical aspect of sarcasm only (Frenda et al., 2022).
- Datasets of sarcasm contain different styles of sarcasm coming from different domains (Castro et al., 2019; Oprea & Magdy, 2019; Khodak et al., 2018).
- There is a need to evaluate the new dataset – CSC.

Challenge 1: Generalizability of sarcasm detection

Datasets:

- Conversational Sarcasm Corpus (CSC)

Challenge 1: Generalizability of sarcasm detection

Datasets:

- Conversational Sarcasm Corpus (CSC)
- MUStARD

Challenge 1: Generalizability of sarcasm detection

Datasets:

- Conversational Sarcasm Corpus (CSC)
- MUStARD
- Sarcasm Corpus (SC)

Challenge 1: Generalizability of sarcasm detection

Datasets:

- Conversational Sarcasm Corpus (CSC)
- MUStARD
- Sarcasm Corpus (SC)
- iSarcasm

Challenge 1: Generalizability of sarcasm detection

Datasets:

- Conversational Sarcasm Corpus (CSC)
- MUStARD
- Sarcasm Corpus (SC)
- iSarcasm

- They vary in *original source domain*, *size*, *label source*, *modality*, and *presence of context*.

Challenge 1: Generalizability of sarcasm detection

Dataset comparison (quantitative differences)

	CSC	MUStARD	SC	iSarcasm
Original source domain	Sim. conversations	TV series	Online debates	Social media
Original label type	Multi (1-6)	Binary	Binary	Binary
Annotator agreement	Moderate (Kendall's W 0.56)	Low (Kappa 0.23)	High (Percent agreement 0.80)	N/A
# of sarcastic sentences	2,210 (A) / 2,398 (T)	345	4,693	1,067
Author labels exist	Y	N	N	Y
Third-party labels exist	Y	Y	Y	N
Is multimodal	N	Y	N	N
Context exists	Y	Y	N	N

Challenge 1: Generalizability of sarcasm detection

Qualitative differences

Dataset	Examples from each dataset
CSC	*Context*: Steve gives you a watering can on your birthday while smiling at you with a strange expression. But you don't even have a single plant. *Response*: Maybe I will use it as an outside shower.
MUStARD	*Context*: 'How do I look?', 'Could you be more specific?', "Can you tell I'm perspiring a little?" *Response*: No. The dark crescent-shaped patterns under your arms conceal it nicely.
SC	Ever hear of artificial ensemination? Why is that heteros only think there is one way to produce children? I find hetero sex disturbing, and an unnatural lifestyle choice.
iSarcasm	Imagine going to university for 4 years when you could just follow Elon Musk on Twitter for free.

Challenge 1: Generalizability of sarcasm detection

Method:

1. Fine-tune encoder-only language models (BERT, RoBERTa, DeBERTa) on different datasets.

Challenge 1: Generalizability of sarcasm detection

Method:

1. Fine-tune encoder-only language models (BERT, RoBERTa, DeBERTa) on different datasets.
2. Intra-dataset prediction.

Challenge 1: Generalizability of sarcasm detection

Method:

1. Fine-tune encoder-only language models (BERT, RoBERTa, DeBERTa) on different datasets.
2. Intra-dataset prediction.
3. Cross-dataset prediction.

Challenge 1: Generalizability of sarcasm detection

Method:

1. Fine-tune encoder-only language models (BERT, RoBERTa, DeBERTa) on different datasets.
2. Intra-dataset prediction.
3. Cross-dataset prediction.
4. Posthoc analysis: What linguistic features are important for detecting sarcasm in each dataset?

Challenge 1: Generalizability of sarcasm detection

Intra-dataset predictions

(A: with author labels, T: with third-party labels, +CONT: context + utterance)

	CSC-A+CONT	**CSC-T+CONT**	**MUS+CONT**	**SC**	**iSarcasm**
BERT	68	<u>73</u>	63	**77**	59
RoBERTa	68	<u>72</u>	44	**80**	42
DeBERTa	67	<u>72</u>	44	**78**	41

- LMs fine-tuned on <u>SC</u> show the best intra-dataset predictions.
- LMs fine-tuned on <u>CSC</u> with third-party labels show the second best performance.
- Ground-truth by <u>observers</u> lead to better sarcasm detection by LMs compared to speaker ground-truth.

Challenge 1: Generalizability of sarcasm detection

Cross-dataset predictions:

- All the LMs struggle to generalize (F-score: 0.80 vs. 0.59).

Challenge 1: Generalizability of sarcasm detection

Cross-dataset predictions:

- All the LMs struggle to generalize (F-score: 0.80 vs. 0.59).
- LMs fine-tuned on CSC show the most stable generalizations to other datasets, though CSC is not the biggest dataset or with the highest intra-dataset predictions.

Challenge 1: Generalizability of sarcasm detection

Cross-dataset predictions: BERT

fine-tuned on	Predicted on							
	CSC+A+CONT	**CSC+T+CONT**	**CSC+A-CONT**	**CSC+T-CONT**	**MUS+CONT**	**MUS-CONT**	**SC**	**iSarcasm**
CSC+A+CONT	–	–	–	–	0.54	0.56	0.42	0.50
CSC+T+CONT	–	–	–	–	0.55	0.57	0.51	**0.53**
CSC+A-CONT	–	–	–	–	**0.57**	**0.58**	0.39	0.43
CSC+T-CONT	–	–	–	–	0.56	0.56	0.46	0.47
MUS+CONT	0.45	0.46	0.51	0.50	–	–	0.39	0.44
MUS-CONT	0.47	0.47	**0.53**	**0.52**	–	–	0.40	0.45
SC	0.44	0.44	0.44	0.44	0.39	0.46	–	0.45
iSarcasm	**0.48**	**0.48**	0.52	0.51	0.44	0.50	**0.59**	–

Challenge 1: Generalizability of sarcasm detection

Cross-dataset predictions: RoBERTa

fine-tuned on	Predicted on							
	CSC+A+CONT	**CSC+T+CONT**	**CSC+A-CONT**	**CSC+T-CONT**	**MUS+CONT**	**MUS-CONT**	**SC**	**iSarcasm**
CSC+A+CONT	–	–	–	–	**0.59**	0.55	0.48	0.52
CSC+T+CONT	–	–	–	–	0.57	**0.57**	**0.56**	**0.54**
CSC+A-CONT	–	–	–	–	0.55	0.56	0.42	0.44
CSC+T-CONT	–	–	–	–	0.56	**0.57**	0.51	0.51
MUS+CONT	0.35	0.35	0.39	0.38	–	–	0.37	0.38
MUS-CONT	0.35	0.35	0.41	0.40	–	–	0.36	0.40
SC	**0.47**	**0.49**	**0.52**	**0.53**	0.39	0.49	–	**0.54**
iSarcasm	0.36	0.35	0.38	0.39	0.36	0.37	0.44	–

Challenge 1: Generalizability of sarcasm detection

Cross-dataset predictions: DeBERTa

fine-tuned on	Predicted on							
	CSC+A+CONT	**CSC+T+CONT**	**CSC+A-CONT**	**CSC+T-CONT**	**MUS+CONT**	**MUS-CONT**	**SC**	**iSarcasm**
CSC+A+CONT	–	–	–	–	**0.55**	**0.57**	0.44	**0.52**
CSC+T+CONT	–	–	–	–	**0.55**	0.56	0.53	**0.52**
CSC+A-CONT	–	–	–	–	0.54	0.55	**0.56**	0.48
CSC+T-CONT	–	–	–	–	0.53	0.54	0.55	0.50
MUS+CONT	0.37	0.37	0.40	0.40	–	–	0.45	0.39
MUS-CONT	0.35	0.35	0.43	0.41	–	–	0.36	0.40
SC	**0.53**	**0.53**	**0.50**	**0.50**	0.37	0.47	–	0.49
iSarcasm	0.34	0.34	0.38	0.37	0.45	0.50	0.35	–

Challenge 1: Generalizability of sarcasm detection

Posthoc analysis: *Why low generalizability?*

- Linguistic features that enable sarcasm detection are different across datasets.

Challenge 1: Generalizability of sarcasm detection

Posthoc analysis: *Why low generalizability?*

- Linguistic features that enable sarcasm detection are different across datasets.
- **Sarcasm Corpus**: Words about negative emotion, social issues, swearwords, and online-style words;

Challenge 1: Generalizability of sarcasm detection

Posthoc analysis: *Why low generalizability?*

- Linguistic features that enable sarcasm detection are different across datasets.
- **Sarcasm Corpus**: Words about negative emotion, social issues, swearwords, and online-style words;
- **MUStARD**: Words related to family and drives (i.e., achievement, rewards, etc.);

Challenge 1: Generalizability of sarcasm detection

Posthoc analysis: *Why low generalizability?*

- Linguistic features that enable sarcasm detection are different across datasets.
- **Sarcasm Corpus**: Words about negative emotion, social issues, swearwords, and online-style words;
- **MUStARD**: Words related to family and drives (i.e., achievement, rewards, etc.);
- **CSC**: Words related to agreement (i.e., Ok, yes..), and religion (i.e., oh my god, Jesus Christ...);

Challenge 1: Generalizability of sarcasm detection

Discussion:

- Sarcasm comes in different styles in different datasets, which poses challenges for language models.

Challenge 1: Generalizability of sarcasm detection

Discussion:

- Sarcasm comes in different styles in different datasets, which poses challenges for language models.
- Sarcasm is not as templated as demonstrated in prior work (Joshi et al., 2015; Chakrabarty et al., 2022).

Challenge 1: Generalizability of sarcasm detection

Discussion:

- Sarcasm comes in different styles in different datasets, which poses challenges for language models.
- Sarcasm is not as templated as demonstrated in prior work (Joshi et al., 2015; Chakrabarty et al., 2022).
- Sarcasm detection models that boast 0.90+ accuracy (e.g., Maynard & Greenwood, 2013) should be evaluated in context.

Challenge 1: Generalizability of sarcasm detection

Publication:
Hyewon Jang & Diego Frassinelli, **Generalizable Sarcasm Detection is Just Around the Corner, of Course!**, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024)*.

Challenge for LLMs 2: Failure of sarcasm in communication

Challenge 2: Failure of sarcasm in communication

RQ: “What factors cause sarcasm failure and do they affect LLM performance?”

Challenge 2: Failure of sarcasm in communication

Sarcasm failure: Intended sarcasm not being understood as such, vice versa (Oprea & Magdy, 2020).

Challenge 2: Failure of sarcasm in communication

Motivation:

- About 75% of sarcasm judgments in CSC align between speakers and observers.

Challenge 2: Failure of sarcasm in communication

Motivation:

- About 75% of sarcasm judgments in CSC align between speakers and observers.
- LLM performance different against speaker vs. observer ground-truth (Jang & Frassinelli, 2024).

Challenge 2: Failure of sarcasm in communication

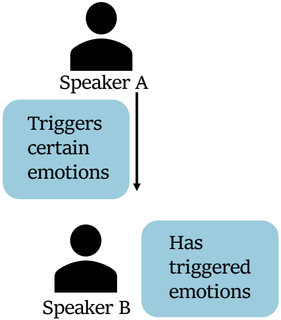
Motivation:

- About 75% of sarcasm judgments in CSC align between speakers and observers.
- LLM performance different against speaker vs. observer ground-truth (Jang & Frassinelli, 2024).

- Annoyance is highly correlated with sarcasm (Jang et al., 2023).

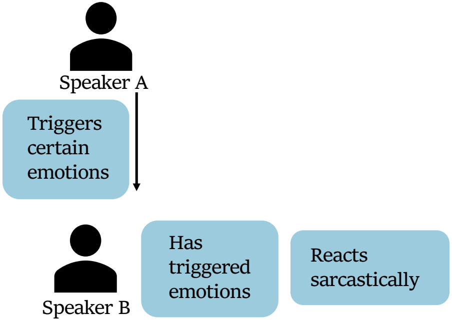
Challenge 2: Failure of sarcasm in communication

Hypothesis:



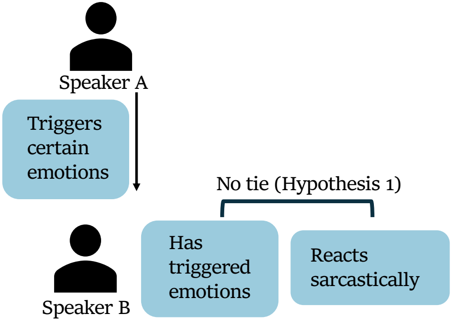
Challenge 2: Failure of sarcasm in communication

Hypothesis:



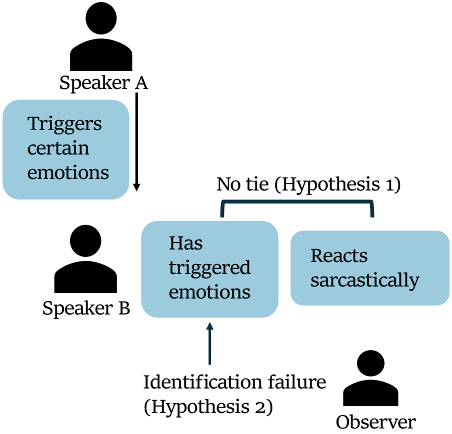
Challenge 2: Failure of sarcasm in communication

Hypothesis:



Challenge 2: Failure of sarcasm in communication

Hypothesis:



Challenge 2: Failure of sarcasm in communication

Hypothesis:

H1. Annoyance-sarcasm incongruity $\Longrightarrow$ failure.

Challenge 2: Failure of sarcasm in communication

Hypothesis:

H1. Annoyance-sarcasm incongruity $\implies$ failure.

H2. Speaker-observer annoyance judgment misalignment $\implies$ failure.

Challenge 2: Failure of sarcasm in communication

Examples of sarcasm failure from CSC

*Annoyance is a type of affect that we focused on for this study.

Type	Sarcasm(Speaker)	Sarcasm(Observer)	Annoyance(Speaker)	Annoyance(Observer)
Speaker's annoyance-sarcasm incongruity	**6**	1	**2**	1

Challenge 2: Failure of sarcasm in communication

Examples of sarcasm failure from CSC

**Affect* used for this study = annoyance

Type	Sarcasm(Speaker)	Sarcasm(Observer)	Annoyance(Speaker)	Annoyance(Observer)
Speaker's annoyance-sarcasm incongruity	**6**	1	**2**	1
Speaker-observer annoyance misalignment	6	1	**5**	**1**

Challenge 2: Failure of sarcasm in communication

Examples of sarcasm failure from CSC

**Affect* used for this study = annoyance

Type	Sarcasm(Speaker)	Sarcasm(Observer)	Annoyance(Speaker)	Annoyance(Observer)
Speaker's annoyance-sarcasm incongruity	**6**	1	**2**	1
Speaker-observer annoyance misalignment	6	1	**5**	**1**

Congruity: 1 if (Sarc $\geq$ 4 & Annoyance $\geq$ 4) or (Sarc $\leq$ 3 & Annoyance $\leq$ 3) else 0

Challenge 2: Failure of sarcasm in communication

Speaker-observer judgment alignment & observers' agreement

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

- y: speaker score
- $\hat{y}$: observer score
- n: number of observers

Challenge 2: Failure of sarcasm in communication

Examples

C + R	SP	OB1	OB2	OB3	OB4	OB5	OB6	**Avg**	**SP-OB alignment**	**OBs agreement**
Ex.1	4	5	4	5	4	4	1	3.86	0.86	0.74
Ex.2	4	5	6	4	3	2	3	3.86	0.81	0.70

Challenge 2: Failure of sarcasm in communication

Analysis using LMER: predict the sarcasm alignment between speakers and observers given the predictors.

Challenge 2: Failure of sarcasm in communication

LMER results

- annoyance-sarcasm congruity $\implies$ sarcasm alignment ($p < 0.001$)
- speaker-observer annoyance alignment:annoyance-sarcasm congruity $\implies$ sarcasm alignment ($p < 0.001$)

Interpretation: Speaker's congruity between annoyance and sarcasm is a very important hint for observers. In this case, when observers correctly identify speaker's annoyance, they will likely identify speaker's sarcasm correctly.

Challenge 2: Failure of sarcasm in communication

Method

1. Task: sarcasm detection (binary)

Challenge 2: Failure of sarcasm in communication

Method

1. Task: sarcasm detection (binary)
2. Data: CSC

Challenge 2: Failure of sarcasm in communication

Method

1. Task: sarcasm detection (binary)
2. Data: CSC
3. Ground-truth sarcasm judgment: from speakers & observers (averaged).

Challenge 2: Failure of sarcasm in communication

Method

1. Task: sarcasm detection (binary)
2. Data: CSC
3. Ground-truth sarcasm judgment: from speakers & observers (averaged).
4. Metric: macro-F1

Challenge 2: Failure of sarcasm in communication

Method

1. Task: sarcasm detection (binary)
2. Data: CSC
3. Ground-truth sarcasm judgment: from speakers & observers (averaged).
4. Metric: macro-F1
5. Encoder-only models: fine-tuned `bert-base-uncased` and `roberta-base`.

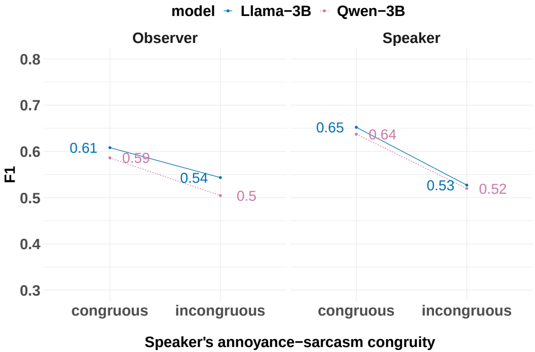
Challenge 2: Failure of sarcasm in communication

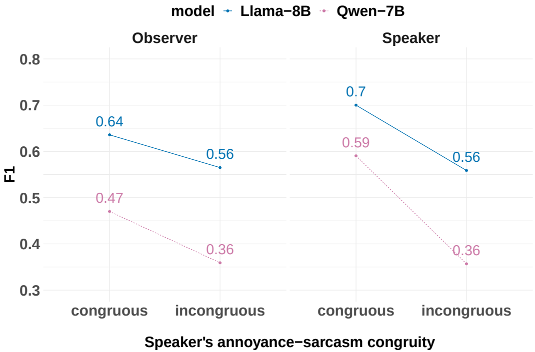
Method

1. Task: sarcasm detection (binary)
2. Data: CSC
3. Ground-truth sarcasm judgment: from speakers & observers (averaged).
4. Metric: macro-F1
5. Encoder-only models: fine-tuned `bert-base-uncased` and `roberta-base`.
6. Generative LLMs: prompted `Llama-3.2-3B-Instruct`, `Llama-3.1-8B-Instruct`, `Qwen2.5-3B-Instruct`, `Qwen2.5-7B-Instruct` in zero-shot settings.

Challenge 2: Failure of sarcasm in communication

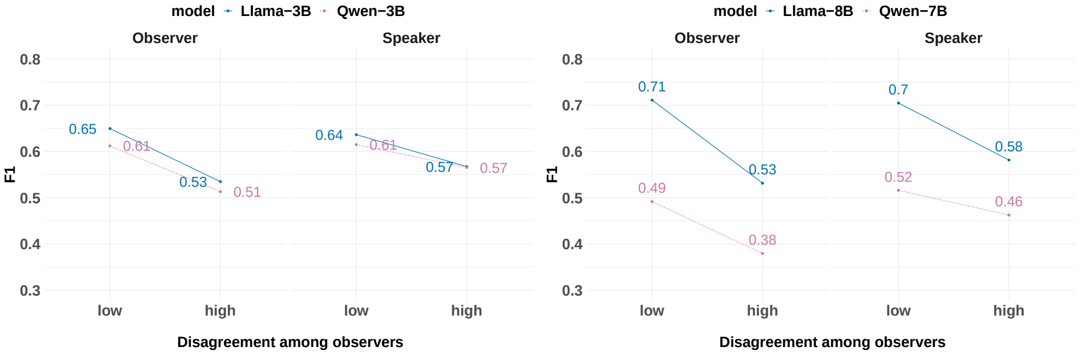
Finding 1. All LLMs struggled to detect sarcasm when the utterance is incongruous with the speaker's annoyance level.





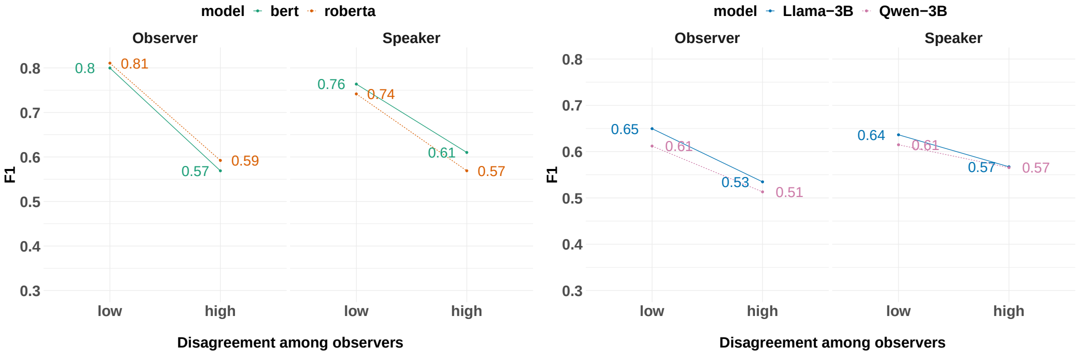
Challenge 2: Failure of sarcasm in communication

Finding 2. All LLMs showed poorer performance when multiple human annotators disagreed on sarcasm label.



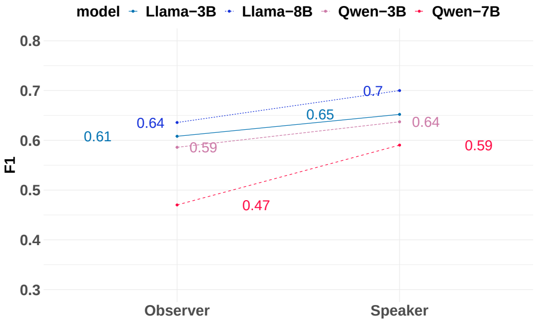
Challenge 2: Failure of sarcasm in communication

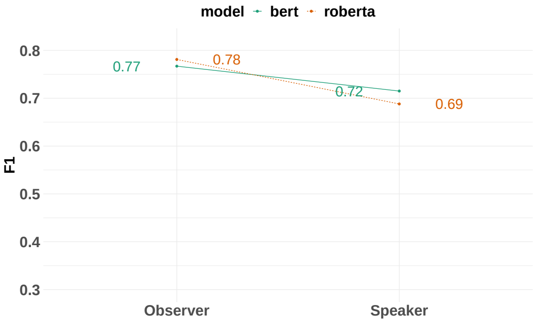
Finding 3. Generative models were generally more robust to disagreement among human annotators than encoder-only models.



Challenge 2: Failure of sarcasm in communication

Finding 4. Generative models tended to perform better with speaker ground-truth than observer ground-truth, in contrast to the encoder-only models.





** congruous condition

Challenge 2: Failure of sarcasm in communication

- To understand the difference between encoder-only and generative models, we explicitly instructed generative LLMs to take the perspective of an external observer.

Challenge 2: Failure of sarcasm in communication

- To understand the difference between encoder-only and generative models, we explicitly instructed generative LLMs to take the perspective of an external observer.
- Even so, their performance with observer ground-truth did not improve.

Challenge 2: Failure of sarcasm in communication

- To understand the difference between encoder-only and generative models, we explicitly instructed generative LLMs to take the perspective of an external observer.
- Even so, their performance with observer ground-truth did not improve.
- We observed the general tendency for them to provide descriptions about the emotions and motivations of the speaker, showing a speaker-focused behavior.

Challenge 2: Failure of sarcasm in communication

- The phrase "What were they called, do you think we bored us!" implies that John believes Steve has been talking about a topic he finds boring. This is a clear indication of sarcasm, **as John is expressing frustration or annoyance** with Steve's long-winded discussion on dinosaurs that John clearly does not find interesting (Qwen-3B).

Challenge 2: Failure of sarcasm in communication

- The phrase ”What were they called, do you think we bored us!” implies that John believes Steve has been talking about a topic he finds boring. This is a clear indication of sarcasm, **as John is expressing frustration or annoyance** with Steve’s long-winded discussion on dinosaurs that John clearly does not find interesting (Qwen-3B).
- In this scenario, John’s response appears to be sincere rather than sarcastic. Despite the accidental damage, John acknowledges his mistake and offers to help clean up. The tone and content of his response indicate that **he is genuinely apologetic and willing to assist**. Therefore, Sarah would likely judge this as **0** (not sarcastic) (Qwen-3B).

Challenge 2: Failure of sarcasm in communication

- 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 (Llama-3B).

Challenge 2: Failure of sarcasm in communication

- 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 (Llama-3B).
- In this situation, John's response is sarcastic. **He's being facetious** when he says "I'm not your waiter". The tone implies that **John is tired of** being treated like a servant and is being sarcastic in his response to Steve's request. So, Sarah would say 1 (Llama-3B).

Challenge 2: Failure of sarcasm in communication

Discussion:

- Sarcasm with hard-to-understand motivation is difficult to detect for both humans and LLMs.

Challenge 2: Failure of sarcasm in communication

Discussion:

- Sarcasm with hard-to-understand motivation is difficult to detect for both humans and LLMs.
- When human observers disagree, LLMs also struggle more in detecting sarcasm.

Challenge 2: Failure of sarcasm in communication

Discussion:

- Sarcasm with hard-to-understand motivation is difficult to detect for both humans and LLMs.
- When human observers disagree, LLMs also struggle more in detecting sarcasm.
- Generative LLMs impersonate speakers' perspective by default, compared to encoder-only ones.

Challenge 2: Failure of sarcasm in communication

Discussion:

- Sarcasm with hard-to-understand motivation is difficult to detect for both humans and LLMs.
- When human observers disagree, LLMs also struggle more in detecting sarcasm.
- Generative LLMs impersonate speakers' perspective by default, compared to encoder-only ones.
- In contrast, observer ground-truth is easier for encoder-only models, consistent with prior work (Oprea & Magdy, 2020).

Challenge 2: Failure of sarcasm in communication

Discussion:

- Sarcasm with hard-to-understand motivation is difficult to detect for both humans and LLMs.
- When human observers disagree, LLMs also struggle more in detecting sarcasm.
- Generative LLMs impersonate speakers' perspective by default, compared to encoder-only ones.
- In contrast, observer ground-truth is easier for encoder-only models, consistent with prior work (Oprea & Magdy, 2020).
- This work illustrates the importance of addressing different perspectives in communication for the assessment of LLM capabilities in future work.

Challenge 2: Failure of sarcasm in communication

Publication:
Hyewon Jang & Diego Frassinelli, **The difficult case of divergence between intended and perceived sarcasm: why it happens and how it challenges LLM performance**, *under review at CoNLL*.

Contributions

1. New experimental findings about sarcasm production and comprehension.

Contributions

1. New experimental findings about sarcasm production and comprehension.
2. New findings about the information encoded in sarcasm detection models.

Contributions

1. New experimental findings about sarcasm production and comprehension.
2. New findings about the information encoded in sarcasm detection models.
3. A new framework that connects (psycholinguistic) experimental methodologies with computational research.

Going forward

- Applying the framework of connecting human data with computational modeling to various linguistic phenomena.

Going forward

- Applying the framework of connecting human data with computational modeling to various linguistic phenomena.
- Investigation of multimodal influence on triggering sarcasm.

Going forward

- Applying the framework of connecting human data with computational modeling to various linguistic phenomena.
- Investigation of multimodal influence on triggering sarcasm.
- Speaker vs. listener vs. observer dynamics in communication & their influence on LLMs.

Thank you for your attention!

Key points.

P1. Sarcasm often occurs because of a certain affect (emotional reaction to a situation) that a context motivates speakers to have.

P2. Observers can mostly identify sarcasm used by speakers as well as the underlying affect of the speakers.

P3. Factors that influence the use of sarcasm in human communication can be used as keys to access computational sarcasm models and to reveal hidden facts about how they detect it.

P4. Sarcasm is broader and more complex than is claimed in previous computational work.

P5. Miscommunications involving sarcasm occur partially due to the broken link between the speakers' affect and their utterance, which poses a significant difficulty both for humans and language models.