LREC COLING 2024

KoCoSa: Korean Context-aware Sarcasm Detection Dataset

Yumin Kim*, Heejae Suh*, Mingi Kim*, Dongyeon Won*, Hwanhee Lee

Language Intelligence Laboratory, Chung-Ang University

Sarcasm Detection

Sarcasm: form of verbal irony characterized by saying something contrary to the text's literal meaning.

The oppositeness makes sarcasm difficult to be detected!

	Negative situation	Positive situation
Context	A: I couldn't sleep at all yesterday to finish my assignment. B: Didn't you get it a month ago? A: That's true, but I kept put it off and did it all yesterday.	A: Happy birthday! I woke up early in the morning to make a handmade cake for you.
Normal Response	B: Wow, how can you put it off that far.	B: Thank you. I'm so touched.
Sarcastic Response	B: Put it off until the deadline? Good job, good job!	B: Really? You don't need to care about my birthday!

⇩

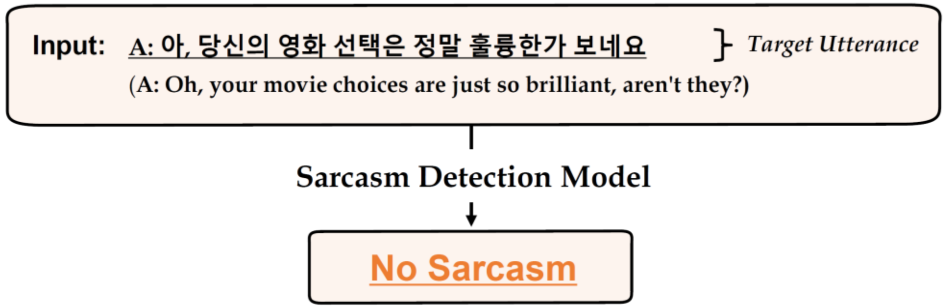
Same Literal Meaning:
Criticizing of A's laziness

⇩

Same Literal Meaning:
Grateful for the birthday cake

Context-aware Sarcasm Detection

Same Target Utterance, but **Different Detection Result**



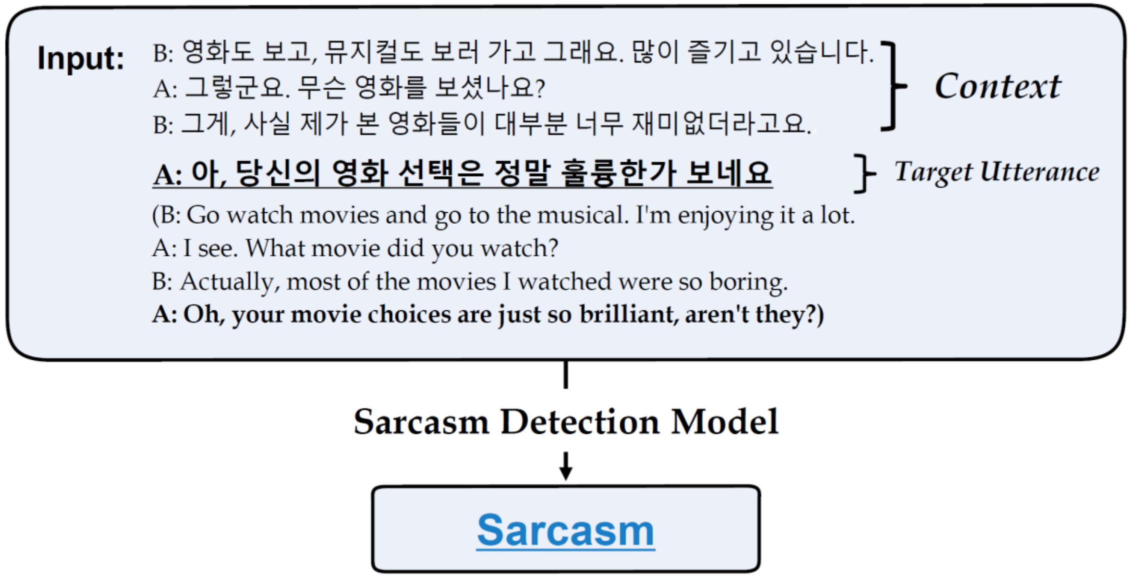
Input: <u>A: 아, 당신의 영화 선택은 정말 훌륭한가 보네요</u> } *Target Utterance*

(A: Oh, your movie choices are just so brilliant, aren't they?)

Sarcasm Detection Model

No Sarcasm

(A) Without Context: "No Sarcasm"



Input: B: 영화도 보고, 뮤지컬도 보러 가고 그래요. 많이 즐기고 있습니다.
A: 그렇군요. 무슨 영화를 보셨나요?
B: 그게, 사실 제가 본 영화들이 대부분 너무 재미없더라고요. } *Context*

<u>A: 아, 당신의 영화 선택은 정말 훌륭한가 보네요</u> } *Target Utterance*

(B: Go watch movies and go to the musical. I'm enjoying it a lot.
A: I see. What movie did you watch?
B: Actually, most of the movies I watched were so boring.
A: Oh, your movie choices are just so brilliant, aren't they?)

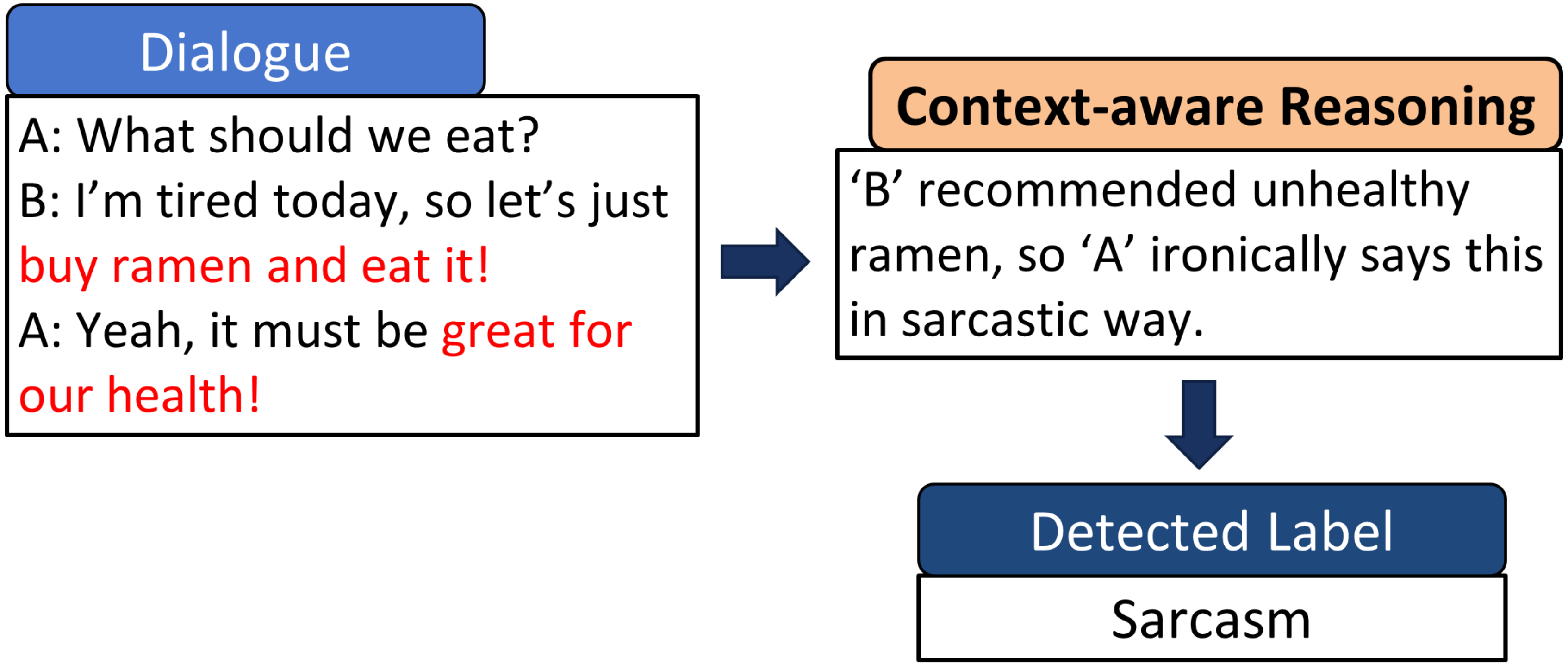
Sarcasm Detection Model

Sarcasm

(B) With Context: "Sarcasm"

Context-aware Sarcasm Detection

- Efficacy of sarcasm detection significantly improves with the consideration of context.
- Incorporating both the last utterance in the context and the target response proved to be the most effective approach in enhancing sarcasm detection performance.



Dialogue

A: What should we eat?
B: I'm tired today, so let's just buy ramen and eat it!
A: Yeah, it must be great for our health!

Context-aware Reasoning

'B' recommended unhealthy ramen, so 'A' ironically says this in sarcastic way.

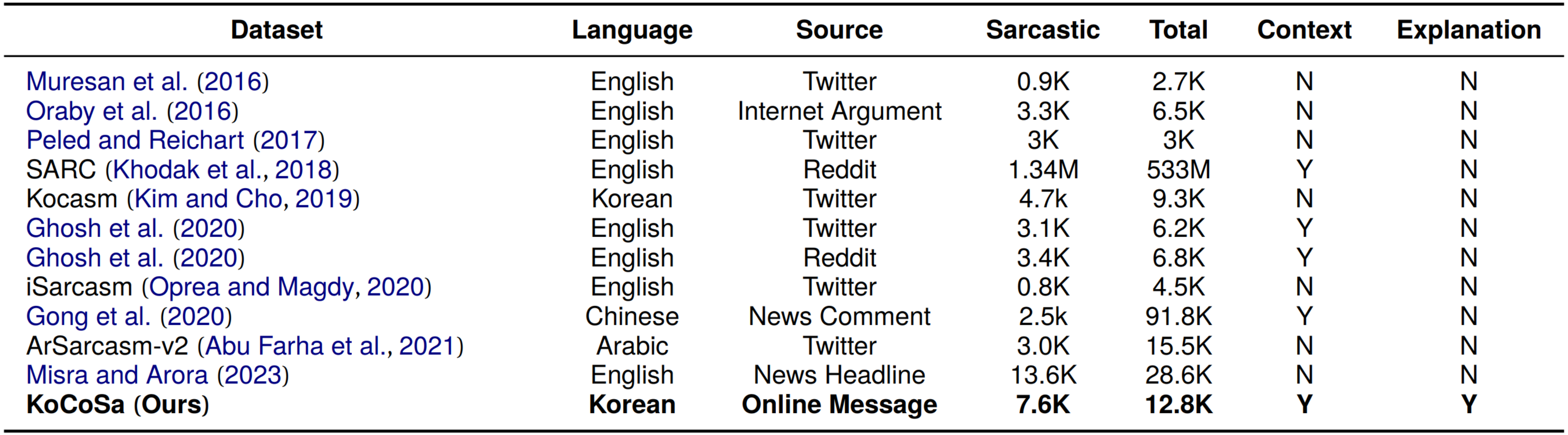
Detected Label

Sarcasm

Lack of Korean Context-aware Sarcasm Dataset

1. Non-English sarcasm datasets are not abundant. Sarcasm reflects cultural differences, so relying solely on translationese may not capture the linguistic nuances in sarcasm detection.

2. There is only one available Korean sarcasm dataset, *Kocasm*, lacks the necessary contexts.

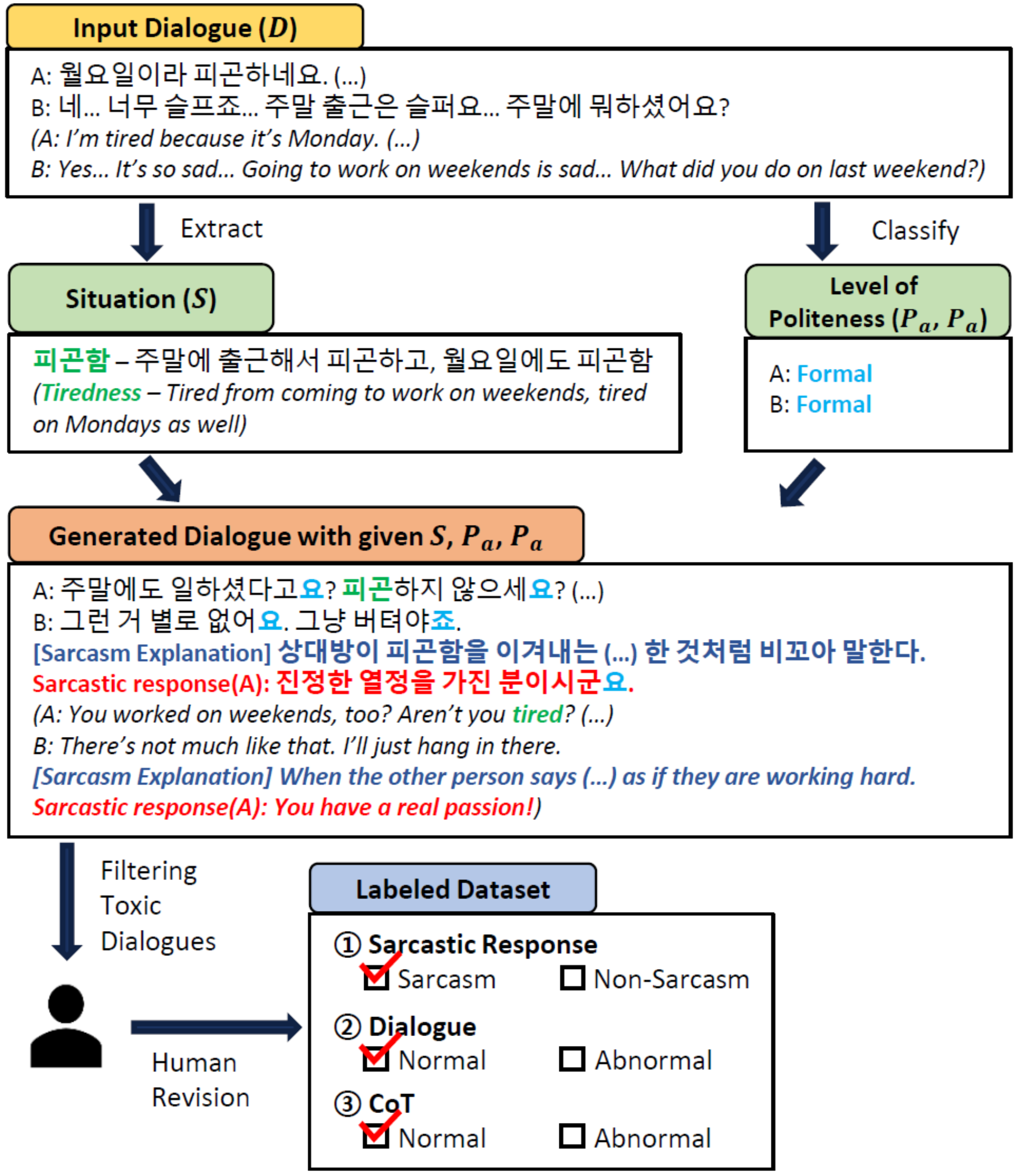


Dataset	Language	Source	Sarcastic	Total	Context	Explanation
Muresan et al. (2016)	English	Twitter	0.9K	2.7K	N	N
Oraby et al. (2016)	English	Internet Argument	3.3K	6.5K	N	N
Peled and Reichart (2017)	English	Twitter	3K	3K	N	N
SARC (Khodak et al., 2018)	English	Reddit	1.34M	533M	Y	N
Kocasm (Kim and Cho, 2019)	Korean	Twitter	4.7k	9.3K	N	N
Ghosh et al. (2020)	English	Twitter	3.1K	6.2K	Y	N
Ghosh et al. (2020)	English	Reddit	3.4K	6.8K	Y	N
iSarcasm (Oprea and Magdy, 2020)	English	Twitter	0.8K	4.5K	N	N
Gong et al. (2020)	Chinese	News Comment	2.5k	91.8K	Y	N
ArSarcasm-v2 (Abu Farha et al., 2021)	Arabic	Twitter	3.0K	15.5K	N	N
Misra and Arora (2023)	English	News Headline	13.6K	28.6K	N	N
KoCoSa (Ours)	**Korean**	**Online Message**	**7.6K**	**12.8K**	**Y**	**Y**

KoCoSa Dataset Construction Pipeline

Two-stage approach using both LLMs and human

1. Extract Situation & Classify Level of Politeness from Input Dialogue
2. Generate Sarcasm Dialogues based on Situation & Level of Politeness
3. Filtering Toxic Dialogues
4. Human Revision
5. Human Labeling



Input Dialogue (D)

A: 월요일이라 피곤하네요. (...)
B: 네... 너무 슬프죠... 주말 출근은 슬퍼요... 주말에 뭐하셨어요?
(A: I'm tired because it's Monday. (...)
B: Yes... It's so sad... Going to work on weekends is sad... What did you do on last weekend?)

Situation (S)

피곤함 – 주말에 출근해서 피곤하고, 월요일에도 피곤함
(Tiredness – Tired from coming to work on weekends, tired on Mondays as well)

Level of Politeness (P_a, P_a)

A: Formal
B: Formal

Generated Dialogue with given S, P_a, P_a

A: 주말에도 일하셨다고요? 피곤하지 않으세요? (...)
B: 그런 거 별로 없어요. 그냥 버텨야죠.
[Sarcasm Explanation] 상대방이 피곤함을 이겨내는 (...) 한 것처럼 비꼬아 말한다.
Sarcastic response(A): 진정한 열정을 가진 분이시군요.
(A: You worked on weekends, too? Aren't you tired? (...)
B: There's not much like that. I'll just hang in there.
[Sarcasm Explanation] When the other person says (...) as if they are working hard.
Sarcastic response(A): You have a real passion!)

Labeled Dataset

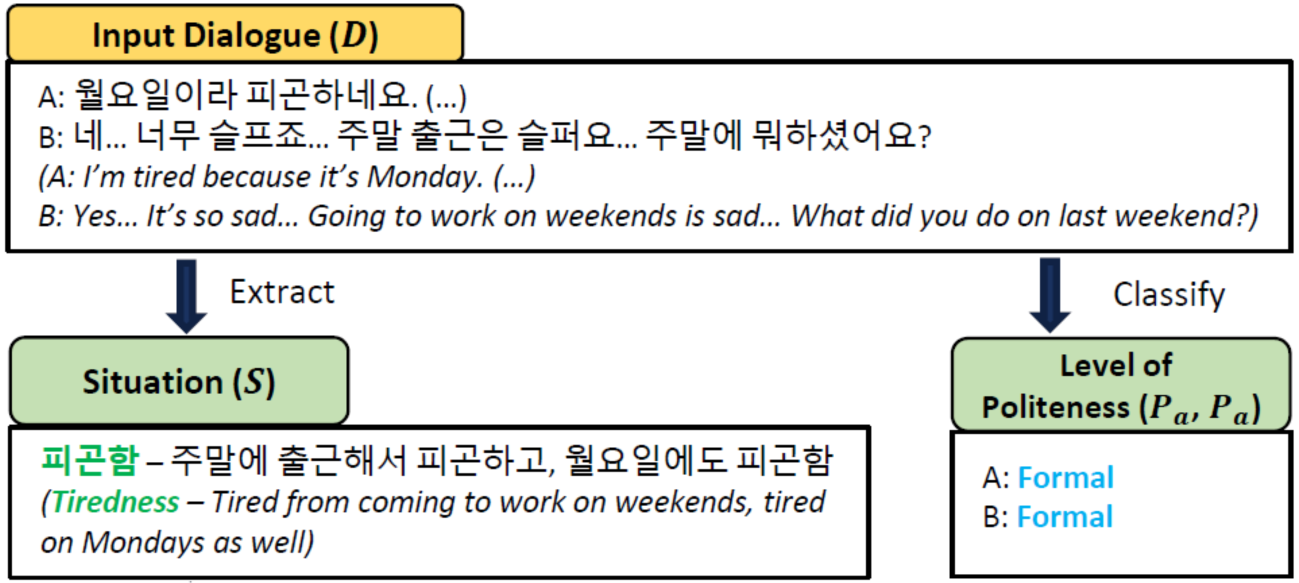
① Sarcastic Response
☑ Sarcasm □ Non-Sarcasm
② Dialogue
☑ Normal □ Abnormal
③ CoT
☑ Normal □ Abnormal

Collecting Source Dialogue

1. Simply instructing LLM to create a new dialogue which includes the sarcastic response for the last utterance results in almost similar theme.
2. Adding only a sarcastic response at the end of an existing dialogue with LLM often generates an unnatural sarcastic utterance that doesn't match the context.

Two Korean source dialogue corpora to construct KoCoSa

- NIKL Messenger corpus (24.4%)
- NIKL Online Text Message corpus (75.6%)



Input Dialogue (D)

A: 월요일이라 피곤하네요. (...)
B: 네... 너무 슬프죠... 주말 출근은 슬퍼요... 주말에 뭐하셨어요?
(A: I'm tired because it's Monday. (...)
B: Yes... It's so sad... Going to work on weekends is sad... What did you do on last weekend?)

Situation (S)

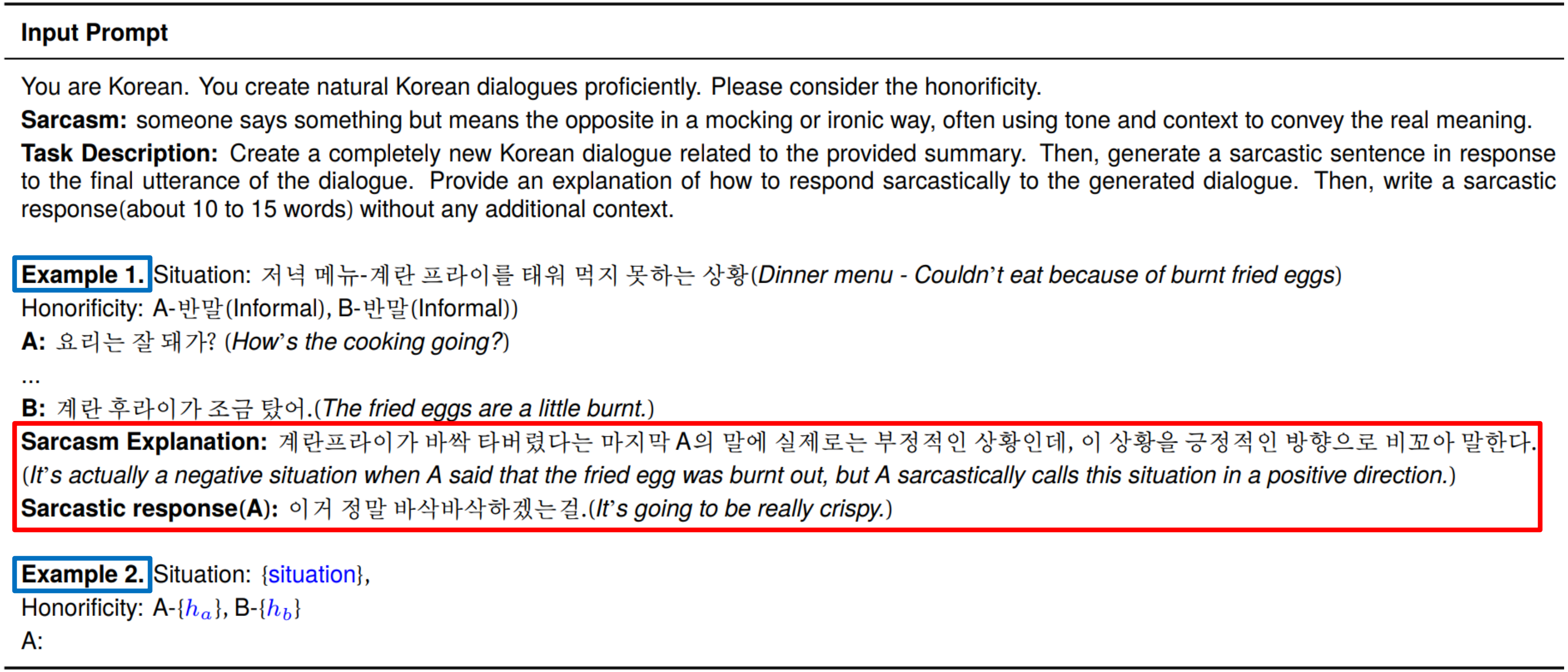
피곤함 – 주말에 출근해서 피곤하고, 월요일에도 피곤함
(Tiredness – Tired from coming to work on weekends, tired on Mondays as well)

Level of Politeness (P_a, P_a)

A: Formal
B: Formal

Dataset Generation using LLM Chain-of-Thought

- Employ Chain-of-Thought (CoT) prompting and two-shot generation approach.
- Instruct GPT-4 to generate ***Sarcasm Explanation*** that explains why the last utterance would be sarcastic considering the **context** before creating a sarcastic response in dialogue generation process.



Input Prompt

You are Korean. You create natural Korean dialogues proficiently. Please consider the honorificity.

Sarcasm: someone says something but means the opposite in a mocking or ironic way, often using tone and context to convey the real meaning.

Task Description: Create a completely new Korean dialogue related to the provided summary. Then, generate a sarcastic sentence in response to the final utterance of the dialogue. Provide an explanation of how to respond sarcastically to the generated dialogue. Then, write a sarcastic response(about 10 to 15 words) without any additional context.

Example 1. Situation: 저녁 메뉴-계란 프라이를 태워 먹지 못하는 상황(*Dinner menu - Couldn't eat because of burnt fried eggs*)
Honorificity: A-반말(Informal), B-반말(Informal))
A: 요리는 잘 돼가? (*How's the cooking going?*)
...
B: 계란 후라이가 조금 탔어.(*The fried eggs are a little burnt.*)
Sarcasm Explanation: 계란프라이가 바싹 타버렸다는 마지막 A의 말에 실제로는 부정적인 상황인데, 이 상황을 긍정적인 방향으로 비꼬아 말한다.
(*It's actually a negative situation when A said that the fried egg was burnt out, but A sarcastically calls this situation in a positive direction.*)
Sarcastic response(A): 이거 정말 바삭바삭하겠는걸.(*It's going to be really crispy.*)

Example 2. Situation: {situation},
Honorificity: A-{h_a}, B-{h_b}
A:

Data Filtering

To improve the quality and moderate the dataset, we **automatically filter toxic or abnormal dialogues** in advance to the annotation process.

	Method	**Dataset Size**
Among 17,073 generated dialogues		
Filtering Toxic Dialogue	1) Apply the **Moderation API of OpenAI** to all data.	Remove 23 (0.0013%) data
	2) Filter samples including **swear words**, using a pre-defined *frequently used toxic words* list.	Remove 5 data
Filtering Abnormal Dialogue	Apply automatic abnormal dialogue **detection by instructing GPT-3.5** to check if the given dialogue is incongruent	Remove 2,257 data

Human Annotation

1. **Context Abnormality Detection (Normal/Abnormal):**

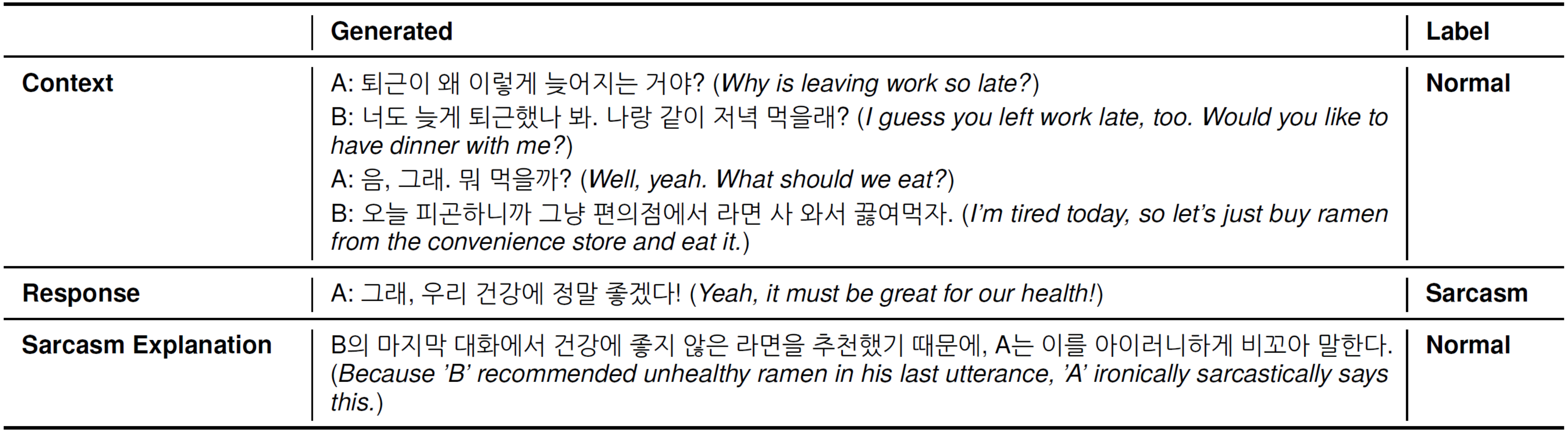
   Does the provided context seem to be a natural dialogue?

2. **Sarcasm Detection (Sarcasm/Non-Sarcasm/Abnormal):**

   Does the last response have sarcastic nuance natural with context?

3. **Sarcasm Explanation Revision (Normal/Abnormal):**

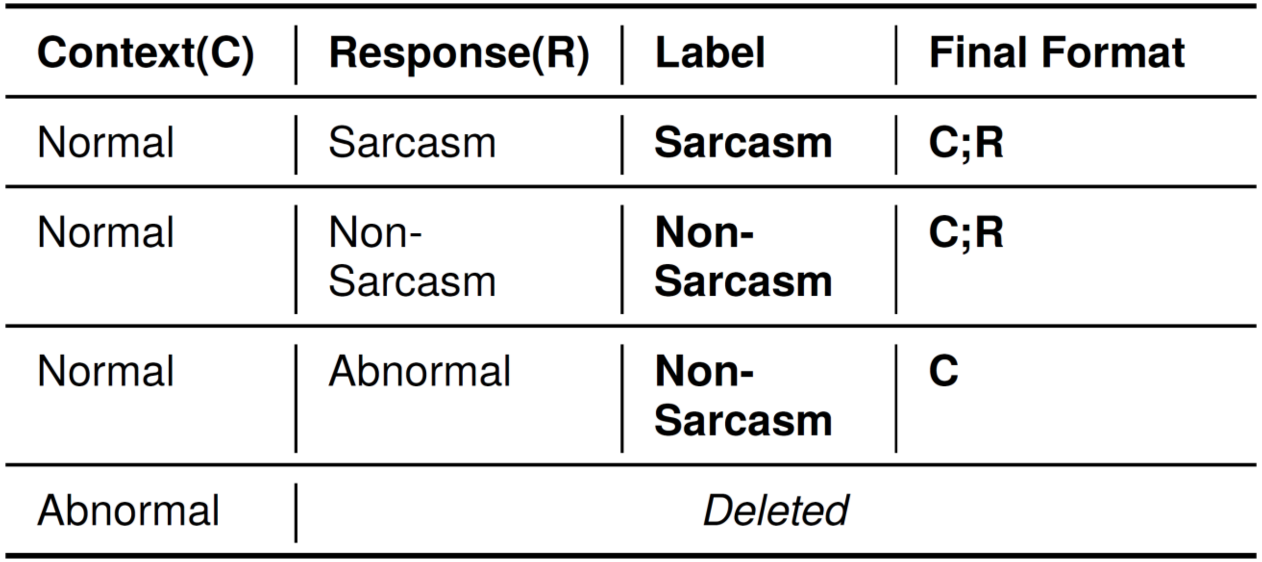
   Does the explanation offer a suitable description of why the response is sarcasm?



	Generated	Label
Context	A: 퇴근이 왜 이렇게 늦어지는 거야? (*Why is leaving work so late?*) B: 너도 늦게 퇴근했나 봐. 나랑 같이 저녁 먹을래? (*I guess you left work late, too. Would you like to have dinner with me?*) A: 음, 그래. 뭐 먹을까? (*Well, yeah. What should we eat?*) B: 오늘 피곤하니까 그냥 편의점에서 라면 사 와서 끓여먹자. (*I’m tired today, so let’s just buy ramen from the convenience store and eat it.*)	**Normal**
Response	A: 그래, 우리 건강에 정말 좋겠다! (*Yeah, it must be great for our health!*)	**Sarcasm**
Sarcasm Explanation	B의 마지막 대화에서 건강에 좋지 않은 라면을 추천했기 때문에, A는 이를 아이러니하게 비꼬아 말한다. (*Because ’B’ recommended unhealthy ramen in his last utterance, ’A’ ironically sarcastically says this.*)	**Normal**

Labeled Dataset

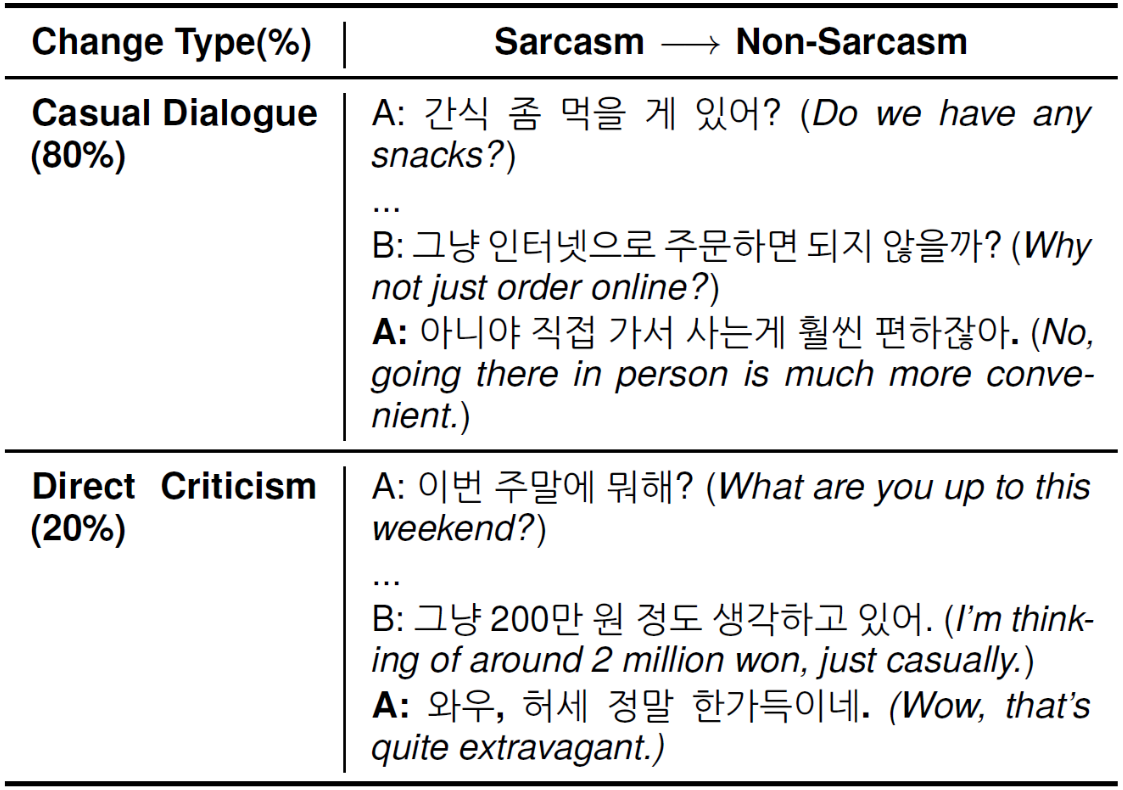
With the annotated data, we construct the dataset with the following construction criteria.



Context(C)	Response(R)	Label	Final Format
Normal	Sarcasm	**Sarcasm**	**C;R**
Normal	Non-Sarcasm	**Non-Sarcasm**	**C;R**
Normal	Abnormal	**Non-Sarcasm**	**C**
Abnormal	*Deleted*		

Unintentional Data Case Study: Non-Sarcasm

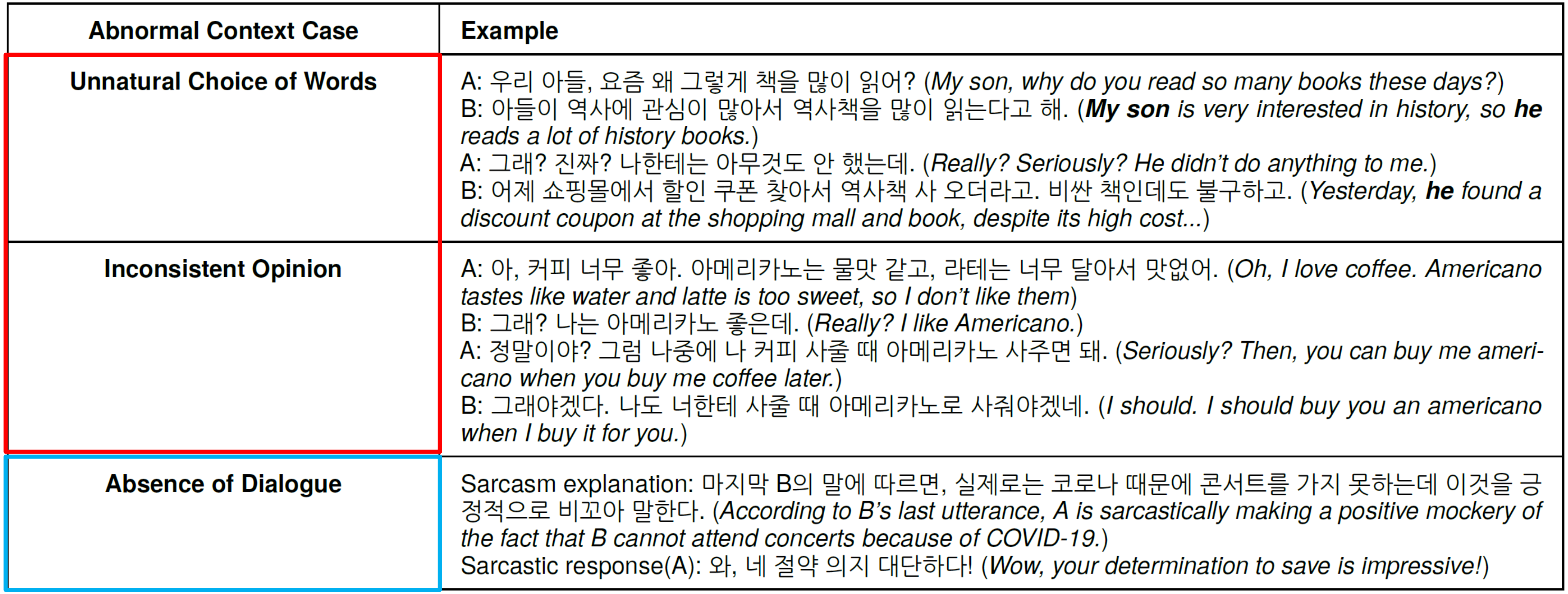
Non-sarcasm data consists of 80% casual dialogue and 20% direct criticism.



Change Type(%)	Sarcasm ⟶ Non-Sarcasm
Casual Dialogue (80%)	A: 간식 좀 먹을 게 있어? (*Do we have any snacks?*) ... B: 그냥 인터넷으로 주문하면 되지 않을까? (*Why not just order online?*) **A:** 아니야 직접 가서 사는게 훨씬 편하잖아**.** (*No, going there in person is much more convenient.*)
Direct Criticism (20%)	A: 이번 주말에 뭐해? (*What are you up to this weekend?*) ... B: 그냥 200만 원 정도 생각하고 있어. (*I'm thinking of around 2 million won, just casually.*) **A:** 와우**,** 허세 정말 한가득이네**.** *(Wow, that's quite extravagant.)*

Unintentional Data Case Study: Abnormal

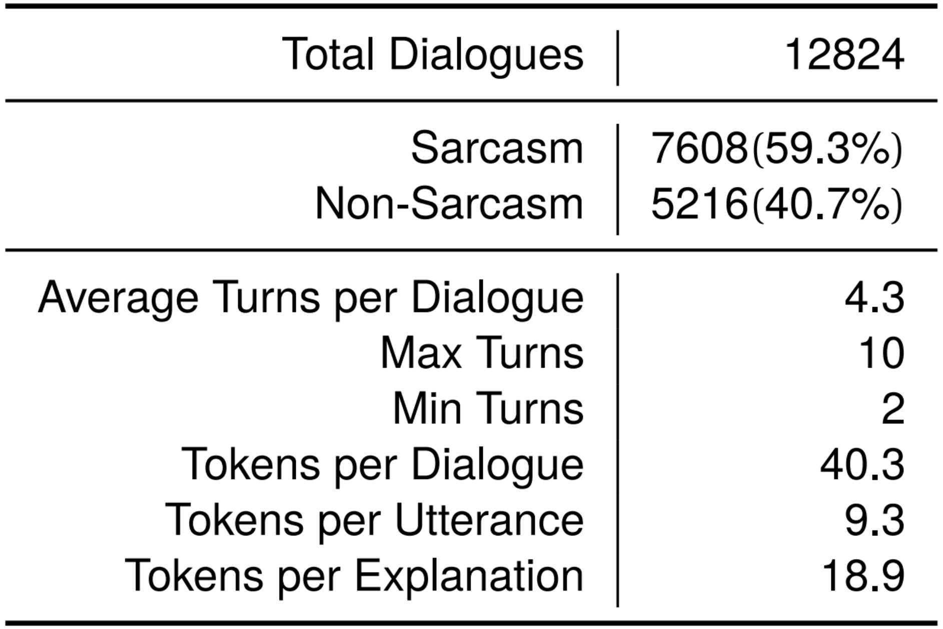
Abnormal data includes 97% contextual awkwardness and 3% format misalignment.



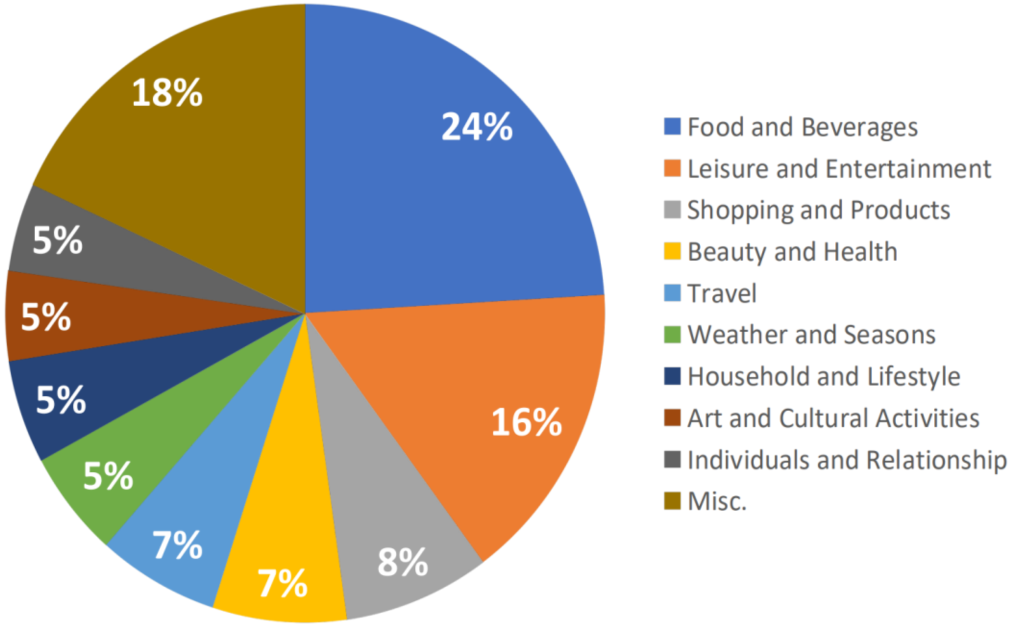
Abnormal Context Case	Example
Unnatural Choice of Words	A: 우리 아들, 요즘 왜 그렇게 책을 많이 읽어? (*My son, why do you read so many books these days?*) B: 아들이 역사에 관심이 많아서 역사책을 많이 읽는다고 해. (***My son** is very interested in history, so **he** reads a lot of history books.*) A: 그래? 진짜? 나한테는 아무것도 안 했는데. (*Really? Seriously? He didn't do anything to me.*) B: 어제 쇼핑몰에서 할인 쿠폰 찾아서 역사책 사 오더라고. 비싼 책인데도 불구하고. (*Yesterday, **he** found a discount coupon at the shopping mall and book, despite its high cost...*)
Inconsistent Opinion	A: 아, 커피 너무 좋아. 아메리카노는 물맛 같고, 라테는 너무 달아서 맛없어. (*Oh, I love coffee. Americano tastes like water and latte is too sweet, so I don't like them*) B: 그래? 나는 아메리카노 좋은데. (*Really? I like Americano.*) A: 정말이야? 그럼 나중에 나 커피 사줄 때 아메리카노 사주면 돼. (*Seriously? Then, you can buy me americano when you buy me coffee later.*) B: 그래야겠다. 나도 너한테 사줄 때 아메리카노로 사줘야겠네. (*I should. I should buy you an americano when I buy it for you.*)
Absence of Dialogue	Sarcasm explanation: 마지막 B의 말에 따르면, 실제로는 코로나 때문에 콘서트를 가지 못하는데 이것을 긍정적으로 비꼬아 말한다. (*According to B's last utterance, A is sarcastically making a positive mockery of the fact that B cannot attend concerts because of COVID-19.*) Sarcastic response(A): 와, 네 절약 의지 대단하다! (*Wow, your determination to save is impressive!*)

Overall Statistics

- **Number of utterances:** minimum of 2 to a maximum of 10 and an average of 4.3
- **Only 13.28% data loss from abnormality:** 14,788 → 12,824 datasets
- **Diverse Source Dialogue Topics:** similar with source dialogues, KoCoSa also contains diverse topics.



Total Dialogues	12824
Sarcasm	7608(59.3%)
Non-Sarcasm	5216(40.7%)
Average Turns per Dialogue	4.3
Max Turns	10
Min Turns	2
Tokens per Dialogue	40.3
Tokens per Utterance	9.3
Tokens per Explanation	18.9



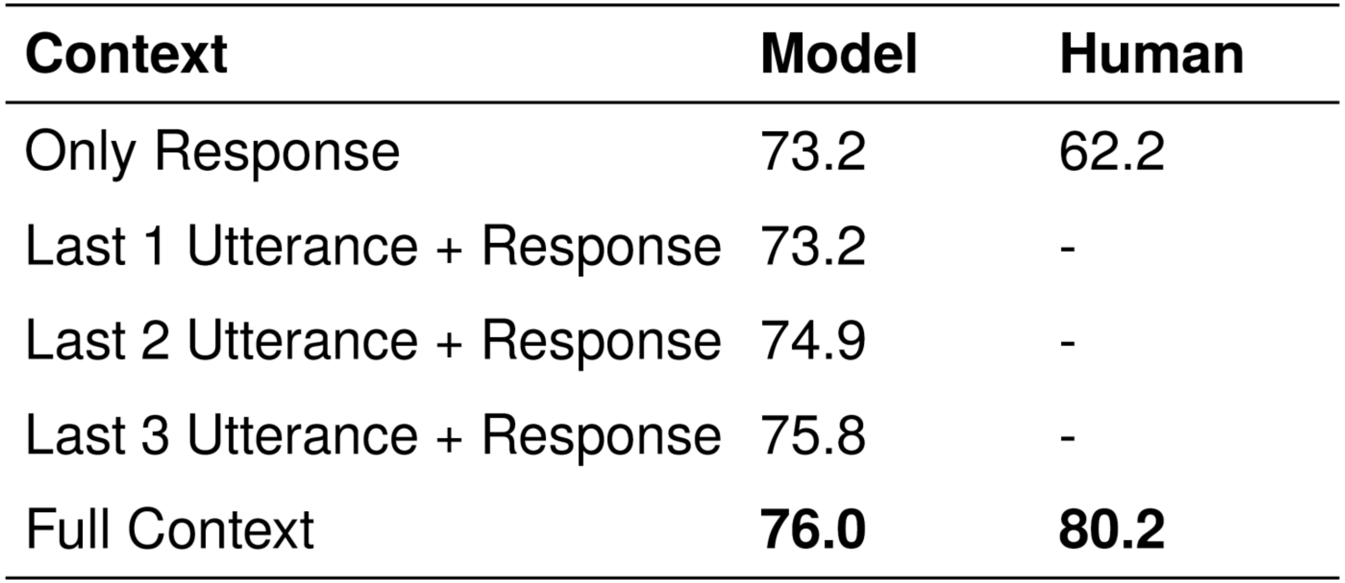
Experiments: Baseline Results

- GPT-3.5 shows about 50% on the Balanced Accuracy and lowest Recall-S in all 0/4/8-shot learning.
- GPT-4 shows the competitive score to Human evaluation.
- Fine-tuning models show competitive performance compared to GPT models, despite the huge gap in the model size.

Model	BA	M-F1	W-F1	Precision-S	Recall-S	Precision-N	Recall-N
Zero/Few-shot							
GPT-3.5(zero-shot)	53.5	43.4	40.6	71.2	20.5	40.2	86.5
GPT-3.5(4-shot)	51.8	42.1	39.4	66.4	20.0	39.2	83.5
GPT-3.5(8-shot)	49.4	32.3	27.2	57.7	6.0	37.8	**92.9**
GPT-4(zero-shot)	73.2	71.7	72.6	**83.3**	68.8	60.5	77.6
GPT-4(4-shot)	75.0	75.1	76.6	80.5	82.3	70.2	67.7
GPT-4(8-shot)	74.5	73.9	75.1	81.9	76.3	65.4	72.6
Fine-tuning							
KLUE-RoBERTa$_{base}$	74.0(±0.2)	74.1(±0.6)	74.7(±0.2)	71.5(±0.2)	**93.4**(±0.5)	**87.2**(±0.7)	54.7(±0.7)
KLUE-RoBERTa$_{large}$	74.9(±0.3)	75.1(±1.0)	75.5(±0.3)	74.6(±0.3)	85.0(±0.6)	80.0(±0.6)	64.8(±0.6)
Human Evaluation	**80.2**	**80.1**	**80.3**	83.0	80.5	77.1	79.9

Experiments: Context Length Dependency

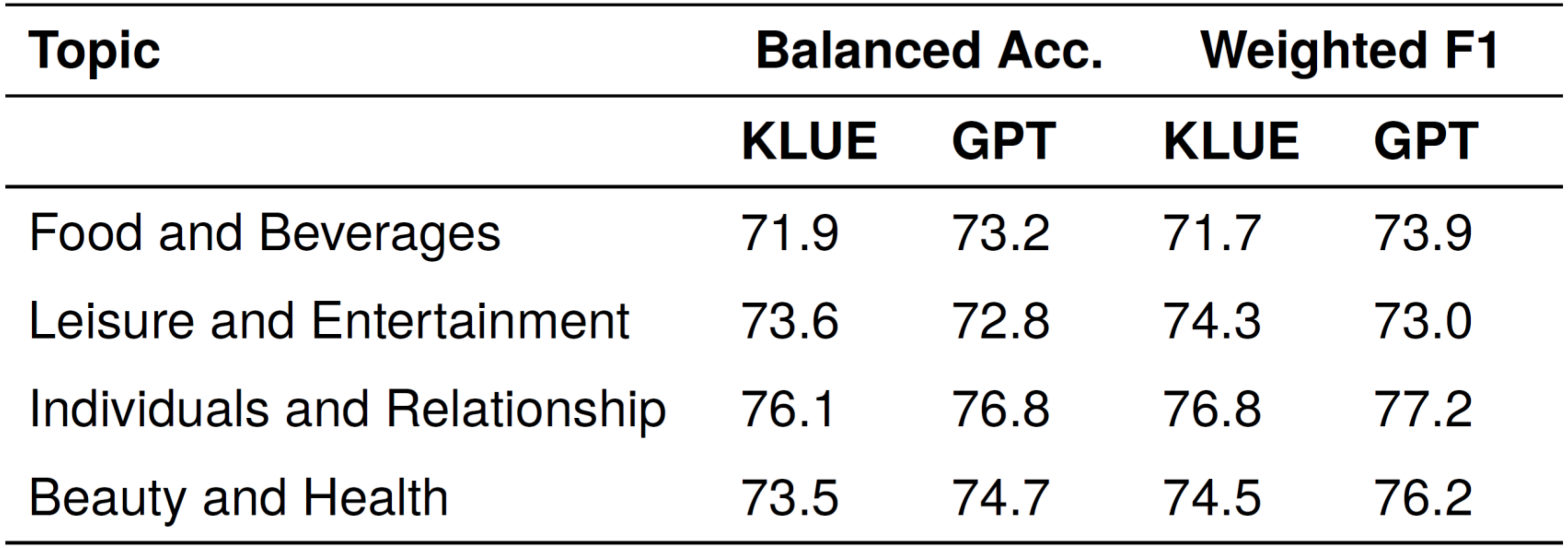
- Model: KLUE-RoBERTa-large
- From the lowest Balanced Accuracy of 73.2 (only the response is used), the performance steadily improves as more context utterances are incorporated.
- Language models benefit from sufficient contextual information to enhance the accuracy of sarcasm detection



Context	Model	Human
Only Response	73.2	62.2
Last 1 Utterance + Response	73.2	-
Last 2 Utterance + Response	74.9	-
Last 3 Utterance + Response	75.8	-
Full Context	**76.0**	**80.2**

Experiments: Topic Dependency

- In KLUE-RoBERTa-large, the gaps between the highest and lowest scores are 4.2%p and 5.1%p in each Balanced Accuracy and Weigthed F1.
- Sarcasm detection performance does not vary with the dialogue's topic.



Topic	Balanced Acc.		Weighted F1	
	KLUE	GPT	KLUE	GPT
Food and Beverages	71.9	73.2	71.7	73.9
Leisure and Entertainment	73.6	72.8	74.3	73.0
Individuals and Relationship	76.1	76.8	76.8	77.2
Beauty and Health	73.5	74.7	74.5	76.2

Closing Remarks

- We propose a comprehensive **dataset generation pipeline** for the context-aware sarcasm detection task using **both LLMs and human revision**.

- We introduce a new **large-scale Korean Context-aware Sarcasm detection dataset (KoCoSa)** through the proposed pipeline, comprising 12.8K daily dialogues.

- We provide a decent analysis of the Korean context-aware sarcasm detection task through this dataset, including the **strong baseline system** for the task.