Could you BE more sarcastic? A Cognitive Approach to Bidirectional Sarcasm Understanding in Language Models

Veer Chheda[1] **Avantika Sankhe**[1] **Atharva Sankhe**[2]

[1] Dwarkadas J. Sanghvi College of Engineering, Mumbai, India
[2] Sardar Patel Institute of Technology, Mumbai, India
{veerchheda3525, avantikasankhe1, atharvasankhe984}@gmail.com

Abstract

Sarcasm is a specific form of ironic speech which can often be hard to understand for language models due to its nuanced nature. Recent improvements in the ability of such models to detect and generate sarcasm motivate us to try a new approach to help language models perceive sarcasm as a speech style, through a human cognitive perspective. In this work, we propose a multi-hop Chain of Thought (CoT) methodology to understand the context of an utterance that follows a dialogue and to perform bidirectional style transfer on that utterance, leveraging the Theory of Mind. We use small language models (SLMs) due to their cost-efficiency and fast response-time. The generated utterances are evaluated using both LLM-as-a-judge and human evaluation, suitable to the open-ended and stylistic nature of the generations. We also evaluate scores of automated metrics such as DialogRPT, BLEU and SBERT; drawing valuable insights from them that support our evidence. Based on this, we find that our cognitive approach to sarcasm is an effective way for language models to stylistically understand and generate sarcasm with better authenticity.

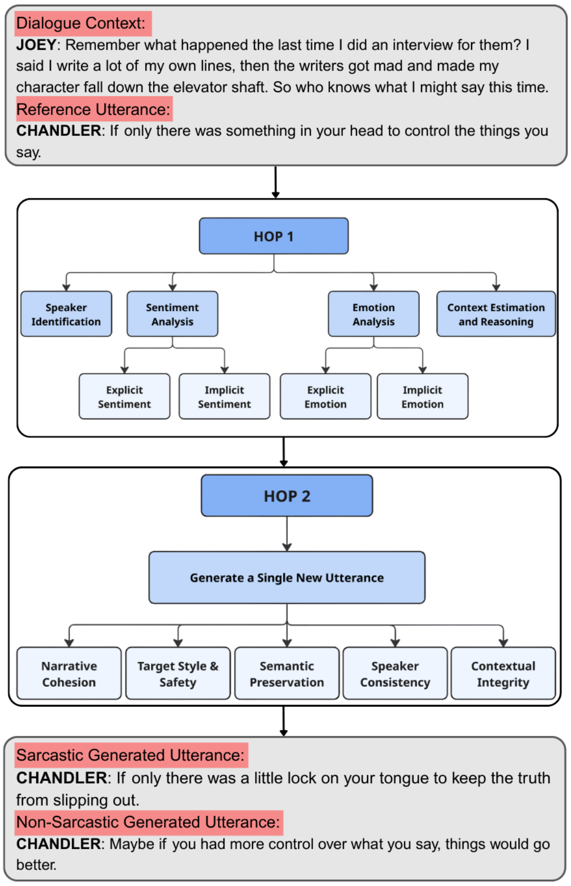

Figure 1: Illustration of our framework for bidirectional sarcasm understanding.

1 Introduction

Sarcasm is a form of verbal irony used to mock or convey contempt toward a person or subject. It is often used as a form of aggressive humour critical in tone indicating playful teasing (Pexman and Olineck, 2002; Frenda et al., 2022). Sarcasm is a communicative act rooted in social cognition and emotional intelligence. It heavily relies on contextual and linguistic cues, including preceding discourse (Campbell, 2012), conversational tone, and linguistic markers such as negation or inversion of literal meaning and use of interjections like 'gee' or 'yeah, right.'

Since sarcasm relies on implied meaning and situational cues, it can often be structurally indistinguishable from non-sarcastic speech, having the same or similar forms (Campbell, 2012). It is observed that language models can often struggle to understand the exact nuances that characterize sarcastic speech (Sharma et al., 2022), such as incongruity between the literal and intended meaning of a statement which particularly marks the presence of sarcasm (Kader et al., 2023; Mishra et al., 2019). Addressing these difficulties and helping language models overcome them is important to improve the natural, human-like quality of text generated by them. This would benefit the ability of language

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 259–276
December 20-24, 2025 ©2025 Association for Computational Linguistics

models to generate and understand humour and double entendres in speech, useful to chatbots, social media analytics and content moderation.

Recent advancements have been made in helping language models detect the presence or absence of sarcasm as a stepping stone towards developing this understanding (Jang and Frassinelli, 2024). The emergence of datasets specifically annotated for sarcasm detection, (Oraby et al., 2016; Jang and Frassinelli, 2024; Oprea and Magdy, 2020; Castro et al., 2019), coupled with improvement in the ability of language models to reason and understand broader contexts (Srivastava et al., 2025) has made it possible to consider helping language models comprehend the nature of sarcasm from a human cognitive perspective. It is interpreted as a dynamic communicative act rather than a speech label, drawing from Theory of Mind (Shamay-Tsoory et al., 2005; Zhu and Wang, 2020).

We approach the task of perceiving sarcasm as a two-step process: 1) the ability of language models to comprehend sarcasm, and 2) to perform bidirectional transformation on the utterance to generate sarcastic or non-sarcastic utterances within an existing context. Using the MUStARD dataset (Castro et al., 2019), we prompt six small language models (SLMs) to generate both sarcastic and non-sarcastic utterances when given the preceding dialogue as context. We use three different prompting methods: 1) zero-shot prompting, 2) few-shot prompting, and 3) Chain of Thought (CoT) prompting. The utterances are generated as alternatives to existing utterances in the MUStARD dataset.

The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 details on our task; Section 4 describes our methodology; Section 5 covers experimental setup and evaluation; Section 6 delineates human evaluation; Section 7 presents results.

2 Related Works

While there has been significant breakthrough in sarcasm detection tasks (Castro et al., 2019; Oprea and Magdy, 2020; Gole et al., 2024), sarcasm generation remains an underexplored task. Recent works focus on highlighting the importance of context (Lunando and Purwarianti, 2013). One such work proposed an unsupervised, modular framework for generating sarcastic outputs by introducing contextual incongruity, setting a benchmark for style transfer techniques without paired data (Mishra et al., 2019). 'Chandler' is not only a sarcasm response generator but also provides explanations for why each response is sarcastic (Oprea et al., 2021). Evaluation of large language models (LLMs) and smaller 7B/8B models on the emerging Sarcasm Explanation in Dialogue (SED) task shows that larger parameter size is an effective factor for superior human language comprehension and reasoning capabilities (Zhang et al., 2024). Investigation of people's preferences on generated sarcasm showed that even when sarcasm was considered highly appropriate, non-sarcastic responses were still preferred (Oprea et al., 2022), which provided a backbone to the concept of sarcasm style transfer according to user preference.

While existing works either focus on detection or one-directional generation, our work stands out as the first to explore bidirectional sarcasm style transfer and incorporating contextual incongruity. We also direct our focus on evaluating our approach on lightweight SLMs.

3 Sarcasm Understanding

From a cognitive perspective, research has shown that sarcasm comprehension engages additional inferential processes compared to literal language (Fanari et al., 2023; McDonald, 1999). Prior work in sarcasm detection highlighted the importance of sarcastic cues for detection, but they don't assess whether a model actually understands sarcasm beyond recognition. To comprehend sarcasm, one must grasp the incongruity between literal meaning and intended meaning, drawing on contextual knowledge and theory of mind (Shamay-Tsoory et al., 2005; Zhu and Wang, 2020). Similarly, production of sarcasm requires speakers to manipulate linguistic cues to insert an incongruity while ensuring the underlying context remains interpretable (Ghosh et al., 2018; Ghosh and Veale, 2017). Motivated by this, we suggest a bidirectional framework with complementary tasks of comprehension and production, related but distinct cognitive processes that are necessary for demonstrating sarcasm perception.

3.1 Sarcasm Generation

Given a dialogue, the model must generate a sarcastic counterpart that retains the context while introducing pragmatic cues such as exaggeration, polarity reversal or context-dependent irony (Chakrabarty et al., 2020). This task reflects a

Model	Method	BLEU	SBERT	DialogRPT	Accuracy
Gemma 3 1B	Zero-shot	0.017 ± 0.001	0.439 ± 0.002	0.596 ± 0.001	0.902 ± 0.093
	Few-shot	0.046 ± 0.001	0.453 ± 0.001	0.634 ± 0.001	0.862 ± 0.087
	Ours	$\mathbf{0.057 \pm 0.001}$	0.462 ± 0.003	$\mathbf{0.651 \pm 0.001}$	$\mathbf{0.930 \pm 0.060}$
	Few-shot + Ours	0.052 ± 0.003	$\mathbf{0.469 \pm 0.001}$	0.642 ± 0.002	0.906 ± 0.068
Gemma 3 4B	Zero-shot	0.109 ± 0.002	0.443 ± 0.001	0.636 ± 0.001	0.924 ± 0.061
	Few-shot	0.150 ± 0.002	0.451 ± 0.001	0.638 ± 0.001	0.937 ± 0.045
	Ours	$\mathbf{0.189 \pm 0.003}$	0.466 ± 0.001	$\mathbf{0.643 \pm 0.001}$	$\mathbf{0.945 \pm 0.039}$
	Few-shot + Ours	0.167 ± 0.003	$\mathbf{0.467 \pm 0.001}$	0.640 ± 0.002	0.934 ± 0.035
LlaMa 3.2 1B	Zero-shot	0.047 ± 0.004	0.411 ± 0.004	0.602 ± 0.005	$\mathbf{0.839 \pm 0.103}$
	Few-shot	0.052 ± 0.004	0.407 ± 0.004	0.616 ± 0.010	0.706 ± 0.324
	Ours	0.057 ± 0.004	$\mathbf{0.435 \pm 0.003}$	$\mathbf{0.630 \pm 0.013}$	0.723 ± 0.224
	Few-shot + Ours	$\mathbf{0.059 \pm 0.008}$	0.412 ± 0.003	0.598 ± 0.017	0.592 ± 0.397
LlaMa 3.2 3B	Zero-shot	0.044 ± 0.003	0.424 ± 0.003	0.637 ± 0.002	0.906 ± 0.071
	Few-shot	0.072 ± 0.002	$\mathbf{0.453 \pm 0.001}$	0.646 ± 0.002	0.895 ± 0.051
	Ours	0.075 ± 0.002	0.437 ± 0.001	$\mathbf{0.648 \pm 0.001}$	$\mathbf{0.921 \pm 0.052}$
	Few-shot + Ours	$\mathbf{0.094 \pm 0.002}$	0.438 ± 0.002	0.646 ± 0.003	0.899 ± 0.061
Qwen 3 1.7B	Zero-shot	0.250 ± 0.005	0.417 ± 0.002	0.638 ± 0.001	0.773 ± 0.127
	Few-shot	0.271 ± 0.007	$\mathbf{0.432 \pm 0.001}$	0.640 ± 0.001	$\mathbf{0.786 \pm 0.095}$
	Ours	0.224 ± 0.008	0.425 ± 0.001	$\mathbf{0.653 \pm 0.002}$	0.781 ± 0.098
	Few-shot + Ours	$\mathbf{0.291 \pm 0.012}$	0.422 ± 0.001	0.642 ± 0.002	0.752 ± 0.080
Qwen 3 4B	Zero-shot	0.144 ± 0.004	0.424 ± 0.002	0.659 ± 0.001	0.933 ± 0.045
	Few-shot	0.153 ± 0.002	0.438 ± 0.001	0.648 ± 0.001	0.973 ± 0.018
	Ours	0.147 ± 0.004	$\mathbf{0.449 \pm 0.001}$	$\mathbf{0.664 \pm 0.001}$	$\mathbf{0.975 \pm 0.025}$
	Few-shot + Ours	$\mathbf{0.162 \pm 0.004}$	0.436 ± 0.002	0.650 ± 0.002	0.972 ± 0.019
GPT-4o	Zero-shot	0.019 ± 0.001	0.411 ± 0.000	0.671 ± 0.000	0.984 ± 0.017
	Few-shot	0.020 ± 0.001	0.413 ± 0.000	0.672 ± 0.000	0.988 ± 0.006
	Ours	0.020 ± 0.001	$\mathbf{0.423 \pm 0.000}$	$\mathbf{0.675 \pm 0.000}$	0.994 ± 0.006
	Few-shot + Ours	$\mathbf{0.021 \pm 0.002}$	0.422 ± 0.000	0.674 ± 0.000	$\mathbf{0.996 \pm 0.004}$

Table 1: Comparison of different models and methods across automatic evaluation metrics. Scores are reported as mean $\pm$ standard deviation over 5 runs. Best performing scores are highlighted in bold.

model's ability not only to recognize sarcastic cues but also to recreate it intentionally by understanding the inherent context. This includes both stylistic paraphrasing (sarcastic to sarcastic) and style transfer (non-sarcastic to sarcastic).

3.2 Sarcasm Removal

The model must produce a non-sarcastic utterance for a dialogue that retains the intended meaning by recognizing cues and resolving incongruity to recover the reader's intent (Pexman and Olineck, 2002). This includes both style neutralization (sarcastic to non-sarcastic) as it evaluates the model's ability to disentangle the core semantic content of an utterance from its style and factual paraphrasing (non-sarcastic to non-sarcastic) as an anchor point of literal communication for complete bi-directionality. A comprehensive understanding of sarcasm necessarily requires an equally robust understanding of non-sarcasm, since the recognition of irony depends on contrasting it with cases where intent and expression remain aligned.

4 Proposed Methodology

To model bidirectional understanding of sarcasm as a style, we propose a multi-hop framework to decompose the task into sequential stages of contextual understanding for intent and incongruity,

Model	Method	Context	Creativity	Meaning	Rank	Sarcasticness
Gemma 3 1B	Zero-shot	2.9541 ± 0.0608	2.2189 ± 0.0175	2.4931 ± 0.0777	3.0061 ± 0.0468	2.0257 ± 0.0468
	Few-shot	3.9237 ± 0.1063	2.5157 ± 0.0851	3.4439 ± 0.0631	2.4160 ± 0.1028	2.5942 ± 0.0761
	Ours	**3.9526 ± 0.1119**	**2.9104 ± 0.0776**	**3.4686 ± 0.0387**	**2.1986 ± 0.0559**	**3.0318 ± 0.1428**
	Few-shot + Ours	3.6230 ± 0.0996	2.4063 ± 0.0546	3.2265 ± 0.1106	2.5144 ± 0.0825	2.6474 ± 0.0677
Gemma 3 4B	Zero-shot	4.3400 ± 0.0494	2.9900 ± 0.0703	3.6700 ± 0.0650	2.7433 ± 0.0917	2.8700 ± 0.1139
	Few-shot	4.4033 ± 0.0845	2.8800 ± 0.0628	**4.0867 ± 0.0639**	2.3667 ± 0.1130	2.7667 ± 0.1196
	Ours	**4.5133 ± 0.0681**	**3.0133 ± 0.1959**	3.9010 ± 0.1014	**2.2267 ± 0.1782**	**3.0200 ± 0.2253**
	Few-shot + Ours	3.7067 ± 0.1475	2.6033 ± 0.1102	3.4767 ± 0.0535	2.9633 ± 0.0893	2.7333 ± 0.0825
LlaMa 3.2 1B	Zero-shot	3.1867 ± 0.1070	2.3333 ± 0.1359	2.7200 ± 0.0820	2.9167 ± 0.0577	2.0667 ± 0.1173
	Few-shot	**3.6333 ± 0.1541**	2.4767 ± 0.0962	3.2467 ± 0.1221	2.5200 ± 0.1023	2.4400 ± 0.1045
	Ours	3.6281 ± 0.1219	**2.8429 ± 0.1756**	**3.2742 ± 0.1291**	**2.2791 ± 0.0807**	**2.6605 ± 0.2488**
	Few-shot + Ours	2.6391 ± 0.2746	2.2533 ± 0.1958	2.4104 ± 0.2521	2.7238 ± 0.1469	2.3158 ± 0.2385
LlaMa 3.2 3B	Zero-shot	4.2267 ± 0.1116	2.9467 ± 0.0691	3.3933 ± 0.0673	2.3067 ± 0.1134	2.7767 ± 0.0894
	Few-shot	**4.3833 ± 0.1646**	3.0500 ± 0.1419	3.3067 ± 0.1489	2.5901 ± 0.1038	2.8300 ± 0.2053
	Ours	4.3331 ± 0.0987	**3.1759 ± 0.1598**	**3.5788 ± 0.1013**	**2.0888 ± 0.1127**	**2.8526 ± 0.1108**
	Few-shot + Ours	4.1644 ± 0.1336	3.0356 ± 0.1071	3.2875 ± 0.0879	3.0099 ± 0.0839	2.4581 ± 0.1302
Qwen 3 1.7B	Zero-shot	4.2400 ± 0.0596	2.6933 ± 0.0418	4.0300 ± 0.0861	2.3342 ± 0.0877	2.8767 ± 0.0723
	Few-shot	4.0700 ± 0.0606	2.6767 ± 0.0976	3.6167 ± 0.0850	2.4333 ± 0.0920	2.5900 ± 0.1294
	Ours	**4.3633 ± 0.0869**	**2.7467 ± 0.0938**	**4.2167 ± 0.1550**	**2.1333 ± 0.0717**	**2.9500 ± 0.2062**
	Few-shot + Ours	3.7225 ± 0.0736	2.4018 ± 0.2033	3.5190 ± 0.2180	2.9028 ± 0.0410	2.6254 ± 0.1585
Qwen 3 4B	Zero-shot	4.2633 ± 0.0861	2.9533 ± 0.0691	3.7633 ± 0.0477	2.3167 ± 0.1034	2.6900 ± 0.0450
	Few-shot	4.2433 ± 0.0855	**3.0600 ± 0.1090**	3.8200 ± 0.0811	2.2733 ± 0.1234	2.8700 ± 0.1431
	Ours	**4.2933 ± 0.1090**	2.9967 ± 0.1330	**4.2667 ± 0.0565**	**2.1910 ± 0.0723**	**2.9233 ± 0.0703**
	Few-shot + Ours	3.6667 ± 0.1523	2.8433 ± 0.0917	3.3700 ± 0.1293	3.0100 ± 0.2084	2.8200 ± 0.0545
GPT-4o	Zero-shot	4.1500 ± 0.0214	2.9333 ± 0.0834	3.4500 ± 0.1112	2.7598 ± 0.0128	2.6000 ± 0.0121
	Few-shot	4.2833 ± 0.1392	3.2167 ± 0.0323	4.0000 ± 0.1437	**2.4350 ± 0.0548**	**2.9032 ± 0.1034**
	Ours	**4.3167 ± 0.2275**	**3.3667 ± 0.0288**	**4.2833 ± 0.0233**	2.5062 ± 0.0832	2.6833 ± 0.0947
	Few-shot + Ours	3.9833 ± 0.1210	3.1000 ± 0.0955	3.5667 ± 0.0935	2.9899 ± 0.0754	2.8000 ± 0.0838

Table 2: Comparison of different models and different methods across LLM metrics over 5 runs. Scores are reported as mean ± standard deviation. Best performing scores are highlighted in bold.

and transformation for production of sarcastic or non-sarcastic style.

In the first hop, the language model extracts the implicit and explicit emotions and sentiment from each utterance in the dialogue. Explicit emotions and sentiment reflect the surface level state of the conversation while implicit ones are inferred from the linguistic cues, tone and context of the conversation (Chauhan et al., 2020). Disparities between the explicit and implicit emotions and between sentiments of the utterances lead to an incongruity which indicates possibility of sarcasm (Joshi et al., 2017). Utilizing chain-of-thought for reasoning, the model then deduces the underlying rationale and constructs the dialogue context.

This contextual representation of the dialogue is then leveraged to generate a sarcastic or a non-sarcastic conditioned utterance for the dialogue in the second hop (Lee et al., 2025). This hop preserves the speaker's original tone, emotional state and conversational dynamics using chain-of-thought, using the contextual presence or absence of incongruity to accordingly produce the specific style. Figure 1 illustrates the working of our methodology using an example from the MUStARD dataset.

5 Experimentation

5.1 Dataset

We use the publicly available MUStARD dataset (Castro et al., 2019) for our experimentation. It is a multi-modal dataset comprising of 690 audiovisual utterances and dialogue contexts with an even number of annotated sarcastic and non-sarcastic labels. Although, we only use the textual data which is appropriate to our methodology. Initially developed for sarcasm detection, we utilize the 690 dialogue long dataset for sarcasm generation and removal. Each dialogue has an utterance with a label for sarcasm. To achieve complete bi-directionality, we generate both sarcastic and non-sarcastic utterances for each dialogue irrespective of its label.

Model	Type	Zero-shot	Few-shot	Ours	Few-shot+Ours
Gemma-1B	Sarcastic	258.2 ± 5.8	81.8 ± 2.5	$\mathbf{61.4 \pm 8.2}$	71.8 ± 5.4
	Non-Sarcastic	194.0 ± 9.1	97.2 ± 3.6	$\mathbf{53.8 \pm 6.1}$	61.8 ± 2.4
Gemma-4B	Sarcastic	23.4 ± 1.1	31.4 ± 0.5	$\mathbf{15.2 \pm 3.5}$	20.0 ± 1.0
	Non-Sarcastic	24.6 ± 0.9	34.8 ± 0.8	$\mathbf{18.0 \pm 5.4}$	35.4 ± 4.7
LLaMA-1B	Sarcastic	228.4 ± 24.1	82.6 ± 16.4	$\mathbf{61.8 \pm 48.2}$	341.6 ± 78.4
	Non-Sarcastic	122.0 ± 18.3	$\mathbf{74.2 \pm 19.4}$	85.2 ± 27.8	251.6 ± 63.1
LLaMA-3B	Sarcastic	13.6 ± 1.7	9.8 ± 1.3	$\mathbf{9.6 \pm 1.9}$	13.4 ± 3.8
	Non-Sarcastic	10.8 ± 1.5	8.8 ± 1.3	$\mathbf{8.3 \pm 1.9}$	15.4 ± 4.3
Qwen-1B	Sarcastic	10.6 ± 0.5	16.8 ± 0.8	$\mathbf{8.2 \pm 2.3}$	13.8 ± 3.9
	Non-Sarcastic	12.8 ± 0.4	18.0 ± 0.7	$\mathbf{11.0 \pm 2.0}$	13.2 ± 4.4
Qwen-4B	Sarcastic	10.8 ± 1.3	2.2 ± 0.4	$\mathbf{1.2 \pm 0.8}$	1.4 ± 0.9
	Non-Sarcastic	6.8 ± 1.3	1.8 ± 0.9	$\mathbf{0.2 \pm 0.4}$	1.6 ± 1.3

Table 3: Failures (mean $\pm$ std) per model across settings (Zero-shot, Few-shot, Ours, Few-shot+Ours) over 5 runs. Rows report sarcastic and non-sarcastic utterances separately. Best performing scores are highlighted in bold.

5.2 Setup

All tests were run on 2 NVIDIA Tesla T4 GPUs. We report the inference time and memory usage of models in Appendix. We used 4-bit quantization via the Unsloth framework (Han and team, 2023), significantly reducing memory and computation needs, allowing for scalable experimentation. We also use zero-shot and few-shot as baselines along with an ablation of few-shot in our methodology to compare results and efficacy of our strategy.

5.3 Models

We focus primarily on open-weight smaller language models (SLMs) because they can be efficiently deployed on local, on-premises GPUs, enabling cost-effective fine-tuning on configurable sarcastic styles. We use LlaMa 3.2's 1B and 3B variants (Van Der Maaten et al., 2024) , Gemma3's 1B and 4B variants (Kamath and team, 2025), Qwen-3 1.7B and 4B variants(Yang and Qwen Team, 2025) and GPT-4o as a state-of-the-art (SOTA) baseline and LLM-as-a-judge due to its strong reasoning abilities and intelligence (Hurst and Team, 2024).

5.4 Metrics

To evaluate the effectiveness of the proposed multi-hop inference strategy for sarcasm understanding, we employed a combination of automated, LLM and human evaluated metrics.

5.4.1 Automated Metrics

We employed a suite of automatic metrics with sarcasm classification for detecting sarcasm in generated utterances, BLEU-4 for lexical overlap with the reference utterance (Papineni et al., 2002), semantic similarity with the dialogue using Sentence-BERT [1] (Reimers and Gurevych, 2019) and DialogRPT Updown [2] as a dialog-level appropriateness and relevance measure for generated utterances (Gao et al., 2020). We used GPT-4o as a classifier due to its ability to capture context-sensitive pragmatic cues. For every utterance, the dialogue was embedded as context for sarcasm detection.

5.4.2 LLM Metrics

We employed GPT-4o as our LLM-as-a-judge (Gu et al., 2025) to assess the quality of the utterances generated through our multi-hop inference framework on the dimensions mentioned in Table 4. The temperature was set to zero to ensure deterministic and reproducible judgments across all generated outputs.

6 Human Evaluation

We recruited 5 annotators on a volunteer basis from the general public to evaluate a total of 60 cases from the multimodal MUStARD dataset (Castro et al., 2019) in the survey. The annotators were chosen from a pool of volunteers with a minimum of a 4-year bachelor's degree from a program taught strictly in English, ensuring they were proficient

[1]Huggingface: sentence-transformers/all-MiniLM-L6-v2
[2]Huggingface: microsoft:DialogRPT-updown

in the language. Our aim with conducting this human survey was to measure multiple linguistic and stylistic dimensions, giving a deeper insight of how humans comprehend machine-generated sarcastic responses. All evaluations, including human evaluation, are conducted using only the dialogue transcripts and context provided in the dataset, without incorporating visual or audio signals.

6.1 Experimental Setup

The survey presents the participants with 60 distinct, randomly selected cases from the MUStARD dataset (Castro et al., 2019). Each case featured the following:

1. Dialogue context with respective speakers.
2. Four generated candidate utterances, each produced by one of the distinct prompting strategies up for comparison:
    - Using **zero-shot methodology**.
    - Using **few-shot methodology**.
    - Using **our novel methodology**.
    - Using **our novel methodology with few-shot**.

To further reduce order effects and anchoring bias, both the case order and the sequence in which candidate utterances appeared were randomized for every participant. For each set, the annotators were asked to perform two tasks:

1. **Comparative Ranking:** Participants were to rank each utterance from best (1) to worst (4) based on their overall subjective preference.
2. **Likert Scale Rating:** Participants were to rate each of the four generated utterances on a 5-point Likert scale (where 1 = poor quality and 5 = excellent quality) according to the four criteria detailed in Table 4.

Criterion	Description
Sarcasticness	How well does each utterance convey sarcasm?
Creativity	How well does the utterance avoid formulaic or repetitive patterns? How stylistically flexible is it?
Contextual Appropriateness	How fitting is the utterance to the provided dialogue context?
Meaning Preservation	How well does each generated utterance preserve the meaning of the original reference?

Table 4: Evaluation criteria for assessing the quality of generated utterances.

6.2 Justifying Evaluation Criteria

Implicit Display Theory (IDT) (Utsumi, 2005) distinguishes sarcasm from non-sarcasm, and portrays it as a dynamic communicative act with cognitive preconditions such as shared context, emotional intelligence, and the ability to navigate the incongruity between literal and intended meaning. A model could, in theory, detect sarcasm with high accuracy yet fail completely at generating an appropriate sarcastic utterance. In fact, people tend to prefer non-sarcastic responses over incoherent, overly specific sarcastic responses (Oprea et al., 2022). Thus, the evaluation of sarcasm generation must mirror the complexity and nuances of human judgment. Drawing inspiration from above, we rely on human-centric evaluation criteria as automated metrics are often blind to the very pragmatic and contextual nuances. Significance of each criterion is detailed below:

1. **Sarcasticness:** From a *theoretical* point of view, it is a direct application of IDT's concept of the "degree of ironicalness". This criterion measures how effectively an utterance conveys implicit irony. From an *empirical* standpoint, it measures if the model has successfully employed cognitive criteria like pragmatic insincerity and emotional markers. It is also the primary measure of style transfer accuracy.

2. **Creativity:** This criterion measures stylistic expression of the generated utterances. Sarcasm was typically preferred by users only when it was also considered "funny" (Oprea et al., 2022). Creativity includes 'humor' and 'originality', proving to be very valuable. It evaluates the quality of style transfer, assessing if the generated sarcasm is not just recognizable but also potentially preferable to a literal alternative.

3. **Contextual Appropriateness:** This metric directly assesses whether the model has correctly identified a valid context for sarcasm. An utterance cannot be sarcastic in the absence of "ironic environment" (Utsumi, 2005), and inappropriateness in general leads to negative reception of machine-generated sarcasm (Oprea et al., 2021). Measure of contextual incongruity is crucial for evaluating the model's pragmatic and social intelligence.

4. **Meaning Preservation:** It is a cornerstone of any text style transfer task, becoming more nuanced in the specific case of sarcasm. Sarcasm often works by inverting the literal meaning or valence of a statement. This metric ensures that the stylistic transformation does not generate an off-topic utterance that discards the original meaning. Particularly critical for evaluating our bidirectional methodology, it is used to confirm that stylistic neutralization retains semantics.

The evaluation framework required for our task cannot be limited to measuring classification accuracy as a binary evaluation, as it is fundamentally misaligned with the nature of the phenomenon it seeks to measure. Hence, we use 5-point Likert scale to capture the nuances of sarcasm. It is perfect to evaluate 'Sarcasticness' as it explicitly asks the evaluator to place the generated utterance on a continuum, judging not just *if* it is sarcastic, but *how* sarcastic it is. This allows for a much more fine-grained assessment of stylistic success. Ranking the generations according to reader's preference forces a comparative judgment, acknowledging that even among several "sarcastic" outputs, some will simply be better than others. Results are discussed in 7.4.

7 Results

The results of our human survey are summarized in Table 5; the results of automated evaluation metrics are presented in Table 1; Table 2 shows the results of LLM evaluation metrics.

7.1 Need for semantically-aware metrics

It is worthwhile to note that even though BLEU is a widely applied metric for style transfer and generation tasks, it does not lead to any significant trends in our task. In fact, the relatively low BLEU scores observed in our experiments can be attributed to the inherent limitations of lexical overlap metrics in capturing sarcasm and pragmatic nuances. Further, semantic similarity using transformers also fails to capture the shifts in context, style and expression in sarcasm generation. While DialogRPT provides a suitable metric for assessing dialogue quality, it also does not account for subtle changes in pragmatic nuance and sarcastic intent (Gao et al., 2020).

7.2 Limitations in using few shot for multi-hop reasoning

Our methodology shows consistent increases in human, LLM and automated metrics over the baselines and its few-shot counterpart. Incorporating few-shot examples into our strategy showed some improvement over baselines in automated metrics but perform sub-optimally in case of LLM and human evaluated metrics, which again calls for metrics that can capture more than surface-level cues. This is likely because few-shot prompting introduces fixed exemplar biases that may constrain the small language model's reasoning pathways, limiting its ability to explore alternative interpretations to leverage the dialogue context. Further, we also observed formulaic patterns in sarcastic generations like "Oh, absolutely!" or "Oh, really?" and non-sarcastic generations like "That's a bummer" or "I'm sorry to hear that". We theorize these generations were likely due to model's limited reasoning capabilities as utterances became more creative as model-size increased.

7.3 Punts and Failures

SLMs are known to have limited reasoning which leads to failures like punts, text degeneration, text repetition, etc. A primary example is a punt, which is a response where the model explicitly avoids or refuses to fulfill the prompt (e.g., "I'm sorry, I cannot help with that"). Other failures include text degeneration, text repetition and so on. We analyzed our generations for these failures along with our task specific failures such as wrong speaker name and empty generation ('<your generated line>'). We enumerate these errors in Table 3. Our methodology demonstrates an improvement in reasoning over the other methods by giving fewer punts. We do not include GPT-4o in the table as it did not lead to failures.

7.4 Result Analysis

Across all four bidirectional style transfer tasks, our method was consistently preferred by human annotators over the zero-shot and few-shot baselines. The inter-annotator agreement was calculated using Krippendorff's alpha which yielded a score of 0.4536. This depicts a moderate level of agreement which seems reasonable due to the highly subjective and nuanced nature of the task, where individual interpretations tend to vary. Our method achieved the best performance in Sarcasm Genera-

Category	Method	Context	Creativity	Meaning	Rank	Sarcasticness.
S→ S	ZS	3.45	3.23	3.32	2.79	3.57
	FS	3.59	3.19	3.13	2.65	3.44
	Ours	3.68	**3.68**	**3.38**	**2.50**	3.70
	Ours+FS	**3.80**	3.64	3.29	2.63	**3.76**
S→ NS	ZS	4.11	2.78	3.57	2.75	**3.15**
	FS	3.83	2.41	3.01	2.60	2.88
	Ours	**4.24**	**2.98**	**3.91**	**2.17**	2.48
	Ours+FS	4.07	2.85	3.68	2.48	2.84
NS→ S	ZS	3.73	3.39	3.01	2.73	2.96
	FS	3.71	3.61	2.80	2.60	3.34
	Ours	**4.01**	**3.81**	**3.51**	**2.21**	3.56
	Ours+FS	3.96	3.69	3.28	2.41	**3.60**
NS→ NS	ZS	3.71	2.55	3.25	2.48	**1.84**
	FS	3.79	2.61	3.17	2.45	1.76
	Ours	**4.17**	**2.77**	**3.97**	**2.11**	1.65
	Ours+FS	3.87	2.61	3.47	2.16	1.79

Table 5: Direction-wise Human Evaluation Metrics. Sarcastic (S), Non-sarcastic (NS). Best performing scores are highlighted in bold.

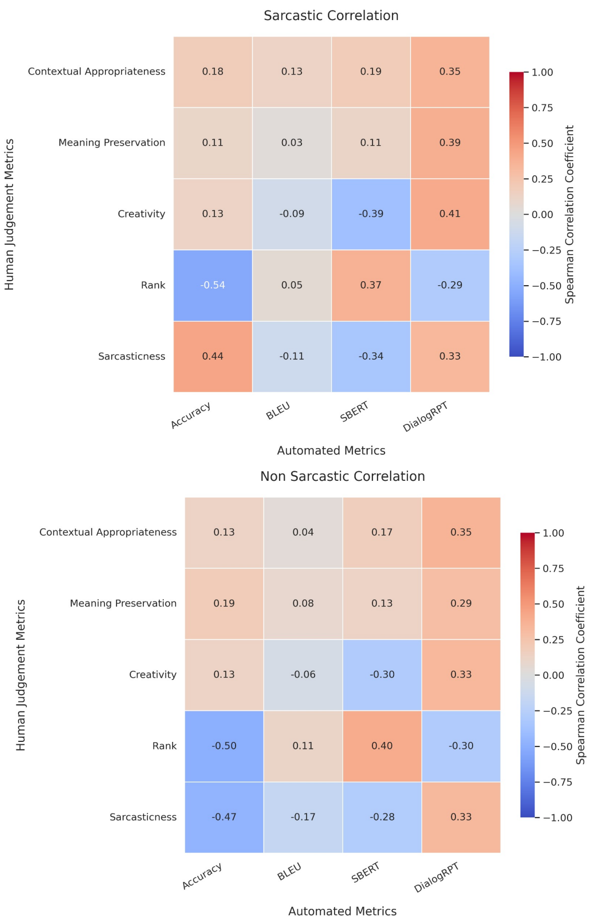

Figure 2: Spearman correlation between human and automated metrics across sarcastic and non-sarcastic cases.

tion, demonstrating that cognitive reasoning of the first hop enables the model to generate sarcastic utterances that are not only stylistically accurate but also fit naturally in the context. Our method also excelled in Sarcasm Removal and style maintenance tasks.

Our methodology shows consistent improvements over baselines in LLM evaluations and automated metrics as well. While Qwen Models show higher BLEU scores, they also show relatively lower Creativity scores in LLM evaluations. Tables in the Appendix display direction-wise LLM metrics and direction-wise automated metrics. Sarcasm Removal reports lower Creativity and Meaning Preservation scores indicating a loss in creativity and change of meaning, when going from sarcastic to non-sarcastic. However, Sarcasm Generation shows improvements in creativity while inducing sarcasm in utterances over their counterparts. Figures 2 show correlation of human and automated metrics for our task. Accuracy shows a positive and negative correlation with Sarcasticness for sarcastic and non-sarcastic generation respectively. Further, DialogRPT proves to be a good metric for nuanced communication analysis of human metrics. SBERT has a poor correlation demonstrating that higher semantic similarity with the original context doesn't lead to better generations.

8 Conclusion and Future Work

We have performed bidirectional transformation to approach the novel task of understanding of sarcasm as a style using a multi-hop CoT-based framework, helping SLMs generate utterances of specific styles with authenticity while maintaining their contextual relevance. By including a hop to first understand the context and perform reasoning to gain insight into its stylistic nature, in accordance with the Theory of Mind; we were able to generate new utterances in the next hop that preserved the original intent while being expressed creatively to suit the target style. Our experimentation was performed on the textual data of the MUStARD dataset with models taken from across three SLM families, as well as GPT-4o, a SOTA LLM model. Along with automated metrics, we employed human assessment and LLM-as-a-judge for evaluating these generations. We supplemented the results of our methodology with experimentation using other methods such as zero-shot and few-shot. The insights gained highlight the effectiveness of our strategy which approaches sarcasm inspired by principles of human cognition. In the future, we would like to improve the ability of small language models to perform reasoning for sarcasm using Implicit Display Theory (IDT) (Utsumi, 2005) and reinforcement learning, making use of the multimodal features of the MUStARD dataset, as well as employing newer datasets such as SE-MUStARD (Chauhan et al., 2020) with sentiment and emotion annotations.

Limitations

Our human evaluation process involved only 5 human annotators. While this added a valuable source to verify our generations, the paucity of our annotators limits the degree of diversity in insights that could have helped observe trends of human preference for various directions of style transfer. Furthermore, since we have only used the textual data presented within the MUStARD dataset, we were limited to experiment with 690 dialogue cases. We were also limited in the design of our prompts, since our experimentation did not involve the multimodal features of MUStARD which add further context to each dialogue case. We also found automated metrics such as BLEU and SBERT to show inconsistent alignment with human judgments of sarcasm, with only DialogRPT demonstrating robust correspondence, thus highlighting the scarcity of automated metrics for evaluating stylistic generations.

Ethics Statement

This work only uses public domain datasets and does not use any personal data. We appointed all of our human evaluators on volunteer-basis. Our system is intended solely for informational and research purposes.

Acknowledgments

We would like to express our sincere appreciation to Unsloth for providing an efficient and accessible framework that enabled low-resource conditioning and generation from large language models. Their contributions were instrumental in scaling our experiments across different model sizes while maintaining computational feasibility. We would also like to acknowledge the importance of MUStARD dataset in helping make the task of cognitively understanding sarcasm approachable. Furthermore, we also extend our gratitude to the human annotators who participated in our survey.

References

John D. Campbell. 2012. Investigating components of sarcastic context. *Electronic Thesis and Dissertation Repository*.

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.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

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

Simona Frenda, Alessandra Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

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

Debanjan Ghosh, Alexander R. Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792.

Montgomery Gole, Williams-Paul Nwadiugwu, and Andriy Miranskyy. 2024. On sarcasm detection with openai gpt-based models. In *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*, pages 1–6.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Han and Unsloth team. 2023. Unsloth.

Hurst and OpenAI Team. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5).

Faria Binte Kader, Nafisa Hossain Nujat, Tasmia Binte Sogir, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2023. "when words fail, emojis prevail": A novel architecture for generating sarcastic sentences with emoji using valence reversal and semantic incongruity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 334–351, Toronto, Canada. Association for Computational Linguistics.

Kamath and Gemma team. 2025. Gemma 3: Technical Report. *arXiv preprint arXiv:2503.19786*. Published Mar 12 2025.

Joshua Lee, Wyatt Fong, Alexander Le, Sur Shah, Kevin Han, and Kevin Zhu. 2025. Pragmatic metacognitive prompting improves LLM performance on sarcasm detection. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 63–70, Online. Association for Computational Linguistics.

Edwin Lunando and Ayu Purwarianti. 2013. Indonesian social media sentiment analysis with sarcasm detection. In *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 195–198.

Skye McDonald. 1999. Exploring the process of inference generation in sarcasm: A review of normal and clinical studies. *Brain and Language*, 68(3):486–506.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6144–6154, Hong Kong, China. 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.

Silviu Oprea, Steven Wilson, and Walid Magdy. 2021. Chandler: An explainable sarcastic response generator. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 339–349, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Should a chatbot be sarcastic? understanding user preferences towards sarcasm generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7686–7700, Dublin, Ireland. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Penny Pexman and Kara Olineck. 2002. Does sarcasm always sting? investigating the impact of ironic insults and ironic compliments. *Discourse Processes*, 33:199–218.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Simone Shamay-Tsoory, Rachel Tomer, and Judith Aharon-Peretz. 2005. The neuroanatomical basis of understanding sarcasm and its relationship to social cognition. *Neuropsychology*, 19:288–300.

Mayukh Sharma, Ilanthenral Kandasamy, and Vasantha W B. 2022. R2D2 at SemEval-2022 task 6: Are language models sarcastic enough? finetuning pre-trained language models to identify sarcasm. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1018–1024, Seattle, United States. Association for Computational Linguistics.

Gaurav Srivastava, Shuxiang Cao, and Xuan Wang. 2025. ThinkSLM: Towards reasoning in small language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32600–32650, Suzhou, China. Association for Computational Linguistics.

Akira Utsumi. 2005. Implicit display theory of verbal irony : Towards a computational model of irony.

Laurens Van Der Maaten and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yang and Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2024. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *Preprint*, arXiv:2408.11319.

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 Appendix

A.1 Implementation Details

A.1.1 Model Cards

We have used six open-source small language models, with all converted by Unsloth's 4-bit quantization. These are the links to the official model cards for each model:

- unsloth/LlaMa-3.2-1B-Instruct-unsloth-bnb-523-4bit
- unsloth/LlaMa-3.2-3B-Instruct-unsloth-bnb-525-4bit
- unsloth/Qwen3-1.7B-unsloth-bnb-4bit
- unsloth/Qwen3-4B-unsloth-bnb-4bit
- unsloth/gemma-3-1b-it-unsloth-bnb-4bit
- unsloth/gemma-3-4b-it-unsloth-bnb-4bit

A.1.2 Inference Settings

Models were loaded with a maximum sequence length of 1024 tokens. Temperature was set to 0.7 for 5 validation runs for all models. For evaluation, since greedy decoding is not supported by OpenAI API, so we try using deterministic outputs by setting `temperature=0`.

A.1.3 Instruction Template

We follow a structured, instruction-based multihop prompting strategy to guide the model in generating new utterances of particular styles. In each prompt, the dialogue and utterance is specified followed by clear directions to help it understand the context in the first hop and then generate suitable utterances in the second hop. Additionally, we also provide the prompts used for few-shot and zero-shot strategies.

For the first hop which serves the purpose of contextual understanding, the model is prompted as follows:

```
Read the following dialogue.
Dialogue:
{dialogue}
{utterance}
For each line in the dialogue
as well as the utterance, do the following:
1. Identify the speaker.
2. Identify the **explicit sentiment**
(positive, neutral, negative) expressed
directly in what is said.
3. Identify the **implicit sentiment**
(positive, neutral, negative) inferred
from tone, choice of words, or context.
4. Name the **explicit emotion**
(anger, excited, fear, sad, surprised,
frustrated, happy, neutral, disgust).
5. Name the **implicit emotion**
(anger, excited, fear, sad, surprised,
frustrated, happy, neutral, disgust).
6. Briefly explain the reasoning for both
explicit and implicit sentiment/emotion.
7. Estimate the context based off of
the dialogue, identified implicit
and explicit sentiment, and emotion.
Format for each line:
Speaker: <name>
Explicit Sentiment:
<positive/neutral/negative/mixed>
Implicit Sentiment:
<positive/neutral/negative/mixed>
Explicit Emotion:
<emotion>
Implicit Emotion:
<emotion>
Reasoning:
<brief explanation>
Context: <context>
```

For the second hop which serves the purpose of generating utterances, the model is prompted as follows:

```
You are given a dialogue
and a detailed analysis of
explicit and implicit sentiment/emotion
for each line.
Previous sentiment analysis:
{hop1_output}
Original utterance:
{utterance}
Task:
Generate ONE new utterance that:
- Fits naturally after the dialogue.
- Is written in a {mode} style
that is not harmful.
- **Preserves the core meaning and intent**
of the original utterance.
- Matches the original speaker's tone,
emotional state, and relationship dynamics.
- Maintains the estimated
context of the conversation.

{example}

Format:
New utterance: {speaker_name}: <your
generated line>
```

For few-shot strategy, we used the following prompt:

```
Read the following dialogue.
Dialogue:
{dialogue}
Utterance:
{utterance}
Task:
Generate ONE new utterance that:
- Fits naturally after the dialogue.
- Is written in a {mode} style that is
not harmful.
- Preserves the core meaning and
intent of the original utterance.
- Matches the original speaker's tone,
emotional state, and relationship dynamics.
- Maintains the estimated context
of the conversation.
Example 1:
Input:
Dialogue:
Output:
{example_output_1}
Example 2:
Input:
Dialogue:
Output:
{example_output_2}
Format:
New utterance: {speaker_name}:
<your generated line>
"""
```

We evaluate the generated utterances using LLM by giving it the following prompt:

```
"You are evaluating multiple candidate
utterances for a dialogue.\n\n"
"Your evaluation must follow these steps:
\n\n"
"Step 1: Rank each utterance.\n"
"Rank the four generated utterances
in order from 1 (best) to 4 (worst) "
"based on their overall impression of
quality and effectiveness.\n\n"
"Step 2: Sarcasticness.\n"
"This dimension measures how well
each utterance conveys sarcasm. "
"Rate on a scale from 1 (not sarcastic
at all) to 5 (highly sarcastic). "
"\n\n"
"Step 3: Creativity.\n"
"Creativity assesses the originality
and inventiveness of the utterance. "
"A score of 1 means the utterance
is very plain or formulaic, "
"while 5 indicates a highly novel
and imaginative expression. "
"\n\n"
"Step 4: Contextual Appropriateness.
\n"
"This measures how well the utterance
fits within the dialogue context. "
"Rate on a scale from 1
(very inappropriate or off-topic)
to 5 (very natural and
contextually fitting).
\n\n"
"Step 5: Meaning Preservation
vs Reference.\n"
"How well does each generated utterance
preserve the meaning of the original
reference utterance? "
"Rate from 1 (completely different)
to 5 (very faithful).\n\n"
"Return your answer as a JSON object
mapping each 'utterance i'
```

```
to an object with:\n"
"{rank:int, sarcasticness:int,
creativity:int,
context:int, meaning:int}.\n\n"
"Example:\n"
"{\n"
"  \"utterance 1\": {\"rank\":2,
\"sarcasticness\":4, \"creativity\":3,
\"context\":5, \"meaning\":4},\n"
"  \"utterance 2\": {\"rank\":1,
\"sarcasticness\":2, \"creativity\":2,
\"context\":3, \"meaning\":3}\n"
"}\n\n"
f"Dialogue: {dialogue}\n"
f"Reference utterance (Label:
{label}): \"{reference}\"\n"
```

A.2 Examples

We have provided some examples of utterances generated from bidirectional style transfer according to our methodology. We cover six models across three SLM families that we conducted our experimentation on along with a SOTA LLM model, as listed below:

1. **LLaMA family:** LLaMA3.2 1B, LLaMA3.2 3B
2. **Qwen family:** Qwen3 1.7B, Qwen3 4B
3. **Gemma family:** Gemma3 1B, Gemma3 4B
4. **LLM model:** GPT-4o

In the following tables, we provide examples of generation performing style maintenance and style transfer performed on the utterance by each model. Table 10 shows generation performed over a sarcastic reference utterance, while Table 11 shows generation performed over a non-sarcastic reference utterance. The reference dialogue and utterances, both taken from the MUStARD dataset, are presented below:

1. **Sarcastic reference**:
   - PERSON: Leonard. Come, join us.
   - LEONARD: Hey, Dave.
     And Penny, what a surprise.
   - PENNY: Dave was just showing me around the university. This place is unbelievable!
   - LEONARD: I know, I've been offering to show you around for a year and a half. You always said you had yoga.
   - LEONARD: *Maybe I heard you wrong. A lot of words sound like "yoga."* (Reference utterance)
2. **Non-sarcastic reference**:
   - LEONARD: You'll never guess who they got to replace you at work.
   - SHELDON: Okay, I know what you're doing.
   - LEONARD: Really?
   - SHELDON: Yes, you're using chocolates as positive reinforcement for what you consider correct behaviour.
   - LEONARD: *Chocolate? - No, I don't want any chocolate!* (Reference utterance)

Model	Method	NS → NS				
		R	S	Cr	Cx	M
Gemma 3 1B	1	2.99 ± 0.32	1.49 ± 0.12	2.08 ± 0.14	2.99 ± 0.32	2.56 ± 0.34
	2	2.45 ± 0.17	1.75 ± 0.12	2.11 ± 0.09	3.87 ± 0.21	3.41 ± 0.09
	3	**2.31 ± 0.27**	**1.39 ± 0.26**	**2.48 ± 0.14**	**3.87 ± 0.30**	**3.61 ± 0.14**
	4	2.39 ± 0.25	1.53 ± 0.29	2.16 ± 0.14	3.63 ± 0.21	3.46 ± 0.19
Gemma 3 4B	1	2.56 ± 0.23	1.49 ± 0.06	2.41 ± 0.03	4.11 ± 0.12	3.75 ± 0.15
	2	2.37 ± 0.21	1.55 ± 0.14	2.53 ± 0.23	4.36 ± 0.14	3.95 ± 0.18
	3	**2.32 ± 0.10**	1.47 ± 0.16	**2.57 ± 0.23**	**4.44 ± 0.15**	**3.96 ± 0.21**
	4	2.75 ± 0.32	**1.42 ± 0.31**	2.19 ± 0.21	3.76 ± 0.40	3.64 ± 0.32
LlaMa 3.2 1B	1	2.80 ± 0.26	1.48 ± 0.29	2.19 ± 0.17	3.25 ± 0.20	2.85 ± 0.18
	2	2.43 ± 0.06	1.39 ± 0.17	2.12 ± 0.12	3.81 ± 0.14	3.44 ± 0.22
	3	**2.23 ± 0.12**	**1.29 ± 0.27**	**2.79 ± 0.17**	**3.91 ± 0.23**	**3.45 ± 0.24**
	4	2.55 ± 0.33	2.19 ± 0.14	2.48 ± 0.31	3.01 ± 0.44	2.75 ± 0.41
LlaMa 3.2 3B	1	2.30 ± 0.24	1.85 ± 0.18	2.49 ± 0.15	4.24 ± 0.17	3.77 ± 0.17
	2	2.24 ± 0.12	1.84 ± 0.41	2.60 ± 0.21	**4.35 ± 0.28**	3.91 ± 0.18
	3	**2.11 ± 0.14**	**1.61 ± 0.15**	**2.73 ± 0.20**	3.81 ± 0.24	**3.97 ± 0.21**
	4	2.18 ± 0.22	1.71 ± 0.34	2.25 ± 0.35	3.52 ± 0.31	3.81 ± 0.20
Qwen 3 1.7B	1	2.63 ± 0.13	1.79 ± 0.12	2.13 ± 0.08	**4.32 ± 0.06**	3.91 ± 0.13
	2	2.31 ± 0.10	1.57 ± 0.16	**2.28 ± 0.20**	4.20 ± 0.12	3.91 ± 0.19
	3	**2.19 ± 0.13**	**1.40 ± 0.15**	1.87 ± 0.17	4.28 ± 0.18	**4.12 ± 0.25**
	4	2.45 ± 0.22	1.68 ± 0.10	2.08 ± 0.18	3.69 ± 0.25	3.53 ± 0.37
Qwen 3 4B	1	2.60 ± 0.23	1.34 ± 0.12	2.13 ± 0.12	3.97 ± 0.25	3.71 ± 0.19
	2	2.35 ± 0.12	1.33 ± 0.09	2.39 ± 0.06	4.23 ± 0.06	3.99 ± 0.17
	3	**2.17 ± 0.28**	**1.19 ± 0.18**	2.43 ± 0.13	**4.44 ± 0.28**	**4.39 ± 0.28**
	4	2.84 ± 0.22	1.31 ± 0.26	**2.65 ± 0.26**	3.88 ± 0.32	3.79 ± 0.25
GPT-4o	1	3.20 ± 0.32	1.67 ± 0.11	2.33 ± 0.12	4.00 ± 0.18	3.53 ± 0.11
	2	2.53 ± 0.26	1.53 ± 0.14	2.80 ± 0.11	4.27 ± 0.09	3.80 ± 0.21
	3	**2.11 ± 0.17**	**1.30 ± 0.12**	**2.93 ± 0.19**	**4.67 ± 0.03**	**4.97 ± 0.09**
	4	2.13 ± 0.11	1.43 ± 0.31	2.73 ± 0.08	4.33 ± 0.12	4.87 ± 0.18

Table 6: LLM Evaluation Metrics for Models with Non-Sarcastic Source Text (NS → NS). Categories are NS (Non-Sarcastic) and S (Sarcastic). Parameters are R (Rank), S (Sarcasticness), Cr (Creativity), Cx (Context), and M (Meaning). Methods are 1 (zero-shot), 2 (few-shot), 3 (ours) and 4 (ours + few-shot).

Model	Method	NS → S				
		R	S	Cr	Cx	M
Gemma 3 1B	1	2.67 ± 0.15	2.75 ± 0.20	2.52 ± 0.14	3.08 ± 0.28	2.61 ± 0.17
	2	2.37 ± 0.16	2.73 ± 0.13	2.59 ± 0.16	3.64 ± 0.15	3.78 ± 0.13
	3	**2.27 ± 0.18**	**3.80 ± 0.39**	**3.19 ± 0.22**	**3.75 ± 0.28**	**3.94 ± 0.08**
	4	2.99 ± 0.13	2.92 ± 0.22	2.45 ± 0.18	3.56 ± 0.40	3.80 ± 0.48
Gemma 3 4B	1	2.79 ± 0.14	3.85 ± 0.13	**3.51 ± 0.13**	4.61 ± 0.09	3.53 ± 0.07
	2	2.37 ± 0.25	2.68 ± 0.14	2.83 ± 0.14	**4.69 ± 0.13**	**3.87 ± 0.08**
	3	**2.25 ± 0.23**	**4.15 ± 0.30**	3.32 ± 0.23	4.05 ± 0.21	3.81 ± 0.27
	4	3.12 ± 0.17	2.99 ± 0.25	2.48 ± 0.17	3.89 ± 0.12	3.68 ± 0.35
LlaMa 3.2 1B	1	2.85 ± 0.28	2.35 ± 0.26	2.47 ± 0.26	3.05 ± 0.47	2.69 ± 0.46
	2	**2.25 ± 0.22**	**3.32 ± 0.25**	**3.03 ± 0.18**	**3.87 ± 0.34**	**3.40 ± 0.39**
	3	2.46 ± 0.16	2.71 ± 0.38	2.76 ± 0.25	3.24 ± 0.12	3.02 ± 0.21
	4	3.10 ± 0.30	2.20 ± 0.40	1.92 ± 0.29	2.44 ± 0.37	2.24 ± 0.23
LlaMa 3.2 3B	1	2.17 ± 0.29	3.56 ± 0.28	3.29 ± 0.23	4.19 ± 0.25	3.48 ± 0.19
	2	**1.92 ± 0.27**	3.45 ± 0.36	**3.32 ± 0.34**	**4.51 ± 0.26**	**4.25 ± 0.18**
	3	2.57 ± 0.23	**3.67 ± 0.32**	2.97 ± 0.31	3.93 ± 0.32	3.52 ± 0.06
	4	3.33 ± 0.11	2.61 ± 0.24	2.43 ± 0.08	3.43 ± 0.28	3.44 ± 0.26
Qwen 3 1.7B	1	6.67 ± 0.21	2.66 ± 0.22	2.71 ± 0.08	4.24 ± 0.13	4.05 ± 0.13
	2	2.40 ± 0.31	2.68 ± 0.26	**2.81 ± 0.23**	**4.32 ± 0.13**	3.83 ± 0.08
	3	**2.23 ± 0.15**	**2.68 ± 0.28**	2.77 ± 0.34	4.25 ± 0.20	**4.20 ± 0.16**
	4	3.25 ± 0.19	2.43 ± 0.24	2.35 ± 0.21	3.76 ± 0.09	3.71 ± 0.24
Qwen 3 4B	1	2.41 ± 0.04	3.08 ± 0.11	3.31 ± 0.14	4.20 ± 0.08	3.53 ± 0.13
	2	2.32 ± 0.30	**4.07 ± 0.33**	3.63 ± 0.17	4.23 ± 0.17	3.93 ± 0.17
	3	**2.16 ± 0.27**	4.01 ± 0.25	**3.91 ± 0.26**	**4.29 ± 0.31**	**4.13 ± 0.18**
	4	3.15 ± 0.41	3.76 ± 0.35	3.23 ± 0.36	3.55 ± 0.17	3.11 ± 0.23
GPT-4o	1	2.67 ± 0.12	3.77 ± 0.11	3.42 ± 0.17	4.47 ± 0.11	3.51 ± 0.06
	2	2.56 ± 0.17	3.83 ± 0.02	3.53 ± 0.19	**4.83 ± 0.07**	3.83 ± 0.11
	3	**2.47 ± 0.06**	**3.91 ± 0.09**	**3.56 ± 0.12**	3.87 ± 0.12	**3.92 ± 0.12**
	4	2.90 ± 0.07	3.87 ± 0.04	3.48 ± 0.13	3.67 ± 0.08	3.75 ± 0.09

Table 7: LLM Evaluation Metrics for Models with Non-Sarcastic Source Text (NS → S). Categories are NS (Non-Sarcastic) and S (Sarcastic). Parameters are R (Rank), S (Sarcasticness), Cr (Creativity), Cx (Context), and M (Meaning). Methods are 1 (zero-shot), 2 (few-shot), 3 (ours) and 4 (ours + few-shot).

Model	Method	S → NS				
		R	**S**	**Cr**	**Cx**	**M**
Gemma 3 1B	1	3.60 ± 0.23	$\mathbf{1.43 \pm 0.28}$	1.91 ± 0.16	2.81 ± 0.30	2.28 ± 0.23
	2	2.39 ± 0.14	2.64 ± 0.19	2.60 ± 0.17	3.97 ± 0.15	3.41 ± 0.17
	3	2.32 ± 0.17	2.07 ± 0.39	2.52 ± 0.32	3.85 ± 0.09	3.19 ± 0.08
	4	$\mathbf{1.63 \pm 0.29}$	2.88 ± 0.39	$\mathbf{2.66 \pm 0.45}$	$\mathbf{4.02 \pm 0.24}$	$\mathbf{3.41 \pm 0.22}$
Gemma 3 4B	1	3.15 ± 0.10	2.19 ± 0.23	2.51 ± 0.14	$\mathbf{3.96 \pm 0.14}$	3.21 ± 0.20
	2	2.66 ± 0.19	2.96 ± 0.30	2.85 ± 0.16	3.85 ± 0.22	3.89 ± 0.20
	3	$\mathbf{2.60 \pm 0.25}$	$\mathbf{2.16 \pm 0.42}$	2.98 ± 0.17	3.63 ± 0.22	$\mathbf{3.92 \pm 0.23}$
	4	2.78 ± 0.17	3.20 ± 0.36	$\mathbf{3.09 \pm 0.22}$	3.65 ± 0.19	3.35 ± 0.18
LlaMa 3.2 1B	1	3.44 ± 0.35	$\mathbf{2.32 \pm 0.35}$	2.11 ± 0.23	2.96 ± 0.32	2.44 ± 0.33
	2	2.73 ± 0.15	2.91 ± 0.29	2.17 ± 0.13	3.33 ± 0.27	2.83 ± 0.25
	3	2.06 ± 0.21	2.60 ± 0.56	$\mathbf{2.88 \pm 0.36}$	$\mathbf{3.82 \pm 0.28}$	$\mathbf{3.28 \pm 0.30}$
	4	$\mathbf{2.02 \pm 0.22}$	2.89 ± 0.52	2.81 ± 0.38	2.90 ± 0.38	2.60 ± 0.40
LlaMa 3.2 3B	1	3.08 ± 0.26	2.99 ± 0.18	2.47 ± 0.24	3.76 ± 0.24	3.17 ± 0.22
	2	2.41 ± 0.24	2.87 ± 0.29	2.71 ± 0.18	$\mathbf{4.05 \pm 0.09}$	3.55 ± 0.21
	3	2.39 ± 0.29	$\mathbf{2.53 \pm 0.38}$	2.98 ± 0.27	3.96 ± 0.30	$\mathbf{3.55 \pm 0.38}$
	4	$\mathbf{2.12 \pm 0.31}$	2.83 ± 0.12	$\mathbf{3.14 \pm 0.19}$	3.68 ± 0.32	3.39 ± 0.26
Qwen 3 1.7B	1	2.57 ± 0.27	2.91 ± 0.31	2.64 ± 0.17	3.93 ± 0.17	3.60 ± 0.17
	2	2.71 ± 0.15	2.27 ± 0.28	2.40 ± 0.21	3.55 ± 0.06	3.04 ± 0.21
	3	$\mathbf{1.77 \pm 0.33}$	$\mathbf{2.21 \pm 0.47}$	$\mathbf{3.20 \pm 0.29}$	$\mathbf{4.57 \pm 0.25}$	$\mathbf{4.39 \pm 0.30}$
	4	2.32 ± 0.14	3.20 ± 0.21	2.53 ± 0.31	3.79 ± 0.20	3.47 ± 0.36
Qwen 3 4B	1	2.55 ± 0.27	2.60 ± 0.18	2.76 ± 0.27	4.09 ± 0.19	3.71 ± 0.18
	2	2.73 ± 0.18	2.41 ± 0.22	2.59 ± 0.28	3.88 ± 0.22	3.35 ± 0.21
	3	$\mathbf{1.89 \pm 0.24}$	$\mathbf{1.53 \pm 0.34}$	$\mathbf{3.37 \pm 0.30}$	$\mathbf{4.44 \pm 0.22}$	$\mathbf{4.21 \pm 0.29}$
	4	2.51 ± 0.44	2.71 ± 0.37	2.77 ± 0.20	3.80 ± 0.30	3.48 ± 0.27
GPT-4o	1	3.40 ± 0.13	2.42 ± 0.08	2.27 ± 0.13	3.47 ± 0.05	2.73 ± 0.15
	2	2.67 ± 0.14	2.13 ± 0.14	2.93 ± 0.18	4.13 ± 0.11	3.40 ± 0.13
	3	2.12 ± 0.11	$\mathbf{1.93 \pm 0.13}$	3.07 ± 0.21	4.40 ± 0.08	3.73 ± 0.11
	4	$\mathbf{1.87 \pm 0.38}$	2.87 ± 0.15	$\mathbf{3.80 \pm 0.03}$	$\mathbf{4.47 \pm 0.02}$	$\mathbf{4.00 \pm 0.03}$

Table 8: LLM Evaluation (S → NS). Parameters: R (Rank), S (Sarcasticness), Cr (Creativity), Cx (Context), M (Meaning). Best performing scores are highlighted in bold.

Model	Method	S → S				
		R	**S**	**Cr**	**Cx**	**M**
Gemma 3 1B	1	2.79 ± 0.14	2.41 ± 0.13	2.36 ± 0.17	2.94 ± 0.21	2.51 ± 0.22
	2	2.45 ± 0.28	3.21 ± 0.12	2.75 ± 0.18	3.97 ± 0.28	3.76 ± 0.13
	3	$\mathbf{1.82 \pm 0.15}$	$\mathbf{4.14 \pm 0.21}$	$\mathbf{3.40 \pm 0.25}$	$\mathbf{4.32 \pm 0.20}$	$\mathbf{3.98 \pm 0.25}$
	4	2.97 ± 0.18	2.85 ± 0.22	2.37 ± 0.13	3.33 ± 0.21	3.16 ± 0.14
Gemma 3 4B	1	2.18 ± 0.15	3.85 ± 0.25	$\mathbf{3.53 \pm 0.14}$	$\mathbf{4.68 \pm 0.09}$	3.99 ± 0.13
	2	2.26 ± 0.13	3.88 ± 0.20	3.31 ± 0.13	4.51 ± 0.15	$\mathbf{4.24 \pm 0.09}$
	3	$\mathbf{2.14 \pm 0.23}$	$\mathbf{3.91 \pm 0.38}$	3.48 ± 0.42	4.53 ± 0.12	4.03 ± 0.23
	4	3.23 ± 0.08	3.03 ± 0.14	2.65 ± 0.17	3.52 ± 0.13	3.24 ± 0.10
LlaMa 3.2 1B	1	2.57 ± 0.31	2.79 ± 0.36	2.57 ± 0.33	3.48 ± 0.36	2.89 ± 0.28
	2	2.67 ± 0.27	3.04 ± 0.32	2.59 ± 0.26	3.52 ± 0.31	3.32 ± 0.14
	3	$\mathbf{2.35 \pm 0.30}$	$\mathbf{3.32 \pm 0.38}$	$\mathbf{2.95 \pm 0.47}$	$\mathbf{3.55 \pm 0.36}$	$\mathbf{3.36 \pm 0.45}$
	4	3.22 ± 0.06	2.01 ± 0.22	1.81 ± 0.18	2.21 ± 0.22	2.07 ± 0.26
LlaMa 3.2 3B	1	2.77 ± 0.14	4.11 ± 0.10	3.53 ± 0.12	$\mathbf{4.72 \pm 0.17}$	3.95 ± 0.18
	2	2.43 ± 0.20	4.13 ± 0.37	3.57 ± 0.17	4.63 ± 0.16	$\mathbf{4.21 \pm 0.18}$
	3	$\mathbf{1.77 \pm 0.27}$	$\mathbf{4.23 \pm 0.35}$	$\mathbf{3.62 \pm 0.27}$	4.63 ± 0.23	4.17 ± 0.17
	4	2.98 ± 0.12	3.69 ± 0.24	2.33 ± 0.12	3.94 ± 0.19	3.01 ± 0.13
Qwen 3 1.7B	1	2.36 ± 0.08	3.25 ± 0.23	$\mathbf{3.29 \pm 0.23}$	4.17 ± 0.05	4.06 ± 0.09
	2	$\mathbf{2.32 \pm 0.14}$	3.74 ± 0.11	3.21 ± 0.21	4.21 ± 0.23	3.89 ± 0.14
	3	2.35 ± 0.15	$\mathbf{3.81 \pm 0.33}$	3.25 ± 0.28	$\mathbf{4.35 \pm 0.13}$	$\mathbf{4.16 \pm 0.11}$
	4	3.19 ± 0.17	3.20 ± 0.35	2.64 ± 0.51	3.65 ± 0.11	3.37 ± 0.25
Qwen 3 4B	1	2.83 ± 0.15	3.83 ± 0.05	3.61 ± 0.10	$\mathbf{4.79 \pm 0.07}$	4.11 ± 0.08
	2	$\mathbf{2.01 \pm 0.28}$	3.97 ± 0.14	3.64 ± 0.18	4.44 ± 0.17	4.01 ± 0.14
	3	2.23 ± 0.24	$\mathbf{4.05 \pm 0.15}$	$\mathbf{3.68 \pm 0.38}$	4.53 ± 0.36	$\mathbf{4.33 \pm 0.16}$
	4	3.55 ± 0.17	3.17 ± 0.33	2.72 ± 0.14	3.44 ± 0.08	3.11 ± 0.08
GPT-4o	1	2.83 ± 0.18	3.87 ± 0.09	3.60 ± 0.19	4.67 ± 0.09	3.93 ± 0.11
	2	2.62 ± 0.13	4.38 ± 0.05	4.20 ± 0.11	4.82 ± 0.02	4.27 ± 0.04
	3	$\mathbf{2.46 \pm 0.09}$	$\mathbf{4.56 \pm 0.03}$	$\mathbf{4.27 \pm 0.08}$	$\mathbf{4.88 \pm 0.04}$	$\mathbf{4.33 \pm 0.01}$
	4	3.12 ± 0.43	3.27 ± 0.19	3.17 ± 0.23	4.01 ± 0.07	3.90 ± 0.13

Table 9: LLM Evaluation (S → S). Methods are 1 (zero-shot), 2 (few-shot), 3 (ours) and 4 (ours + few-shot).

Model	Method	Non-Sarcastic → Sarcastic				Non-Sarcastic → Non-Sarcastic			
		BLEU	SBERT	DRPT	Acc.	BLEU	SBERT	DRPT	Acc.
Gemma 3 1B	1	0.019	0.455	0.616	0.902	0.020	0.424	0.607	0.880
	2	**0.074**	0.443	0.664	0.798	**0.065**	0.430	0.614	0.951
	3	0.040	**0.468**	0.669	**0.911**	0.038	**0.432**	**0.636**	0.968
	4	0.051	0.457	**0.670**	0.844	0.052	0.430	0.623	**0.978**
Gemma 3 4B	1	0.140	0.426	0.660	0.932	0.089	0.435	0.644	0.947
	2	**0.184**	0.427	0.661	0.908	**0.109**	0.433	0.636	0.956
	3	0.111	**0.432**	**0.670**	**0.952**	0.078	**0.444**	**0.647**	0.964
	4	0.144	0.429	0.665	0.923	0.091	0.438	0.638	**0.967**
LlaMa 3.2 1B	1	0.039	**0.510**	**0.643**	**0.626**	0.043	**0.487**	**0.656**	0.964
	2	**0.064**	0.417	0.612	0.409	**0.059**	0.411	0.634	0.986
	3	0.043	0.459	0.629	0.456	0.042	0.442	0.646	0.988
	4	0.048	0.424	0.594	0.194	0.056	0.415	0.619	**0.990**
LlaMa 3.2 3B	1	0.039	0.470	0.648	0.868	0.045	0.458	0.638	0.903
	2	0.082	0.471	0.657	0.837	0.070	0.460	0.637	0.947
	3	0.077	**0.496**	**0.670**	**0.920**	0.069	**0.462**	0.640	0.970
	4	**0.103**	0.477	0.668	0.905	**0.083**	0.461	**0.641**	**0.973**
Qwen 3 1.7B	1	**0.513**	**0.467**	0.661	0.722	**0.329**	**0.455**	0.637	0.759
	2	0.225	0.447	0.654	0.729	0.249	0.437	0.639	0.805
	3	0.321	0.437	**0.674**	**0.747**	0.280	0.432	**0.653**	0.819
	4	0.278	0.432	0.657	0.652	0.238	0.428	0.638	**0.841**
Qwen 3 4B	1	**0.177**	0.437	0.679	0.940	**0.170**	0.427	0.644	0.942
	2	0.073	0.450	0.673	0.959	0.084	0.439	0.642	0.965
	3	0.122	**0.469**	**0.695**	**0.986**	0.146	**0.456**	**0.655**	**0.976**
	4	0.084	0.468	0.685	0.979	0.102	0.442	0.645	0.974
GPT-4o	1	0.019	0.440	0.685	0.984	0.020	0.421	0.679	0.973
	2	0.019	0.430	0.686	0.995	**0.022**	0.412	0.683	0.993
	3	0.019	0.446	**0.712**	**0.998**	0.021	**0.431**	**0.685**	**0.997**
	4	**0.020**	**0.447**	0.701	0.996	**0.022**	0.429	0.684	0.995

Table 10: Automated metrics for transfers from a non-sarcastic source. The 'Method' column is abbreviated as: 1 (zero-shot), 2 (few-shot), 3 (ours) and 4 (ours + few-shot).

Model	Setting	Sarcastic → Sarcastic				Sarcastic → Non-Sarcastic			
		BLEU	SBERT	DRPT	Acc.	BLEU	SBERT	DRPT	Acc.
Gemma 3 1B	1	0.017	0.417	0.589	**0.985**	0.013	0.405	0.574	0.762
	2	**0.096**	0.412	0.655	0.785	**0.065**	0.407	0.601	0.912
	3	0.037	**0.420**	**0.669**	0.913	0.033	0.410	**0.630**	0.930
	4	0.053	0.418	0.666	0.837	0.052	**0.414**	0.611	**0.965**
Gemma 3 4B	1	0.132	**0.408**	0.656	0.953	0.075	0.425	0.625	0.845
	2	**0.198**	0.407	0.655	0.947	**0.109**	0.420	0.621	0.855
	3	0.110	0.404	0.663	**0.978**	0.059	0.425	**0.629**	**0.888**
	4	0.157	0.405	**0.665**	0.963	0.076	**0.426**	0.626	0.883
LlaMa 3.2 1B	1	0.057	0.427	0.610	**0.567**	0.050	0.421	**0.634**	0.902
	2	0.066	0.400	0.599	0.449	0.058	0.400	0.620	0.949
	3	0.052	**0.434**	**0.612**	0.476	0.045	**0.427**	0.631	0.973
	4	**0.067**	0.396	0.577	0.203	**0.061**	0.402	0.601	**0.982**
LlaMa 3.2 3B	1	0.045	0.431	0.665	0.904	0.049	0.415	**0.628**	0.808
	2	0.088	0.442	0.657	0.899	0.069	0.421	0.624	0.897
	3	0.087	**0.449**	**0.672**	**0.913**	0.066	**0.434**	0.625	0.920
	4	**0.111**	0.440	0.656	0.878	**0.080**	0.423	0.627	**0.930**
Qwen 3 1.7B	1	**0.601**	**0.435**	0.647	**0.944**	**0.356**	0.424	0.634	0.595
	2	0.297	0.426	0.638	0.882	0.313	0.420	0.628	0.726
	3	0.367	0.424	**0.659**	0.931	0.321	**0.427**	**0.636**	**0.768**
	4	0.349	0.421	0.642	0.812	0.299	0.415	0.631	0.705
Qwen 3 4B	1	**0.231**	**0.443**	0.660	0.982	**0.197**	**0.434**	0.629	0.866
	2	0.086	0.435	0.651	0.980	0.090	0.423	0.626	0.938
	3	0.157	0.430	**0.668**	0.976	0.162	0.429	**0.637**	0.942
	4	0.110	0.434	0.659	**0.988**	0.112	0.415	0.629	**0.945**
GPT-4o	1	0.017	**0.416**	0.667	0.981	0.019	0.413	0.655	0.962
	2	0.020	0.407	0.663	0.989	0.020	0.403	0.654	0.984
	3	0.020	0.396	**0.703**	**0.995**	0.020	**0.420**	**0.657**	0.988
	4	**0.021**	0.397	0.689	0.992	**0.021**	0.416	0.656	**0.990**

Table 11: Automated metrics for transfers from a sarcastic source. The 'Method' column is abbreviated as: 1 (zero-shot), 2 (few-shot), 3 (ours) and 4 (ours + few-shot).

Model	Sarcastic to Sarcastic	Sarcastic to Non-Sarcastic
Llama3.2 1B	Oh, great, just what I needed, another thing to add to my never-ending list of things I've been meaning to do for years.	I'm glad you're finally here, I've been looking forward to showing you around all morning.
Llama3.2 3B	Yeah, because nothing says "yoga" like a bunch of guys in their 30s doing downward-facing dog in the middle of a university building.	I'm glad you're here, though. I was just trying to figure out what you were saying.
Qwen3 1.7B	Maybe I heard you wrong. A lot of words sound like "yoga."	Maybe I heard you wrong. A lot of words sound like "yoga."
Qwen3 4B	Oh, I see. So it's not just the university that's unbelievable — it's the fact that I'm suddenly part of this conversation.	Maybe I heard you wrong. A lot of words sound like "yoga."
Gemma3 1B	Oh, really? You're really getting into the details of the campus layout, are you?	Don't worry, let's just take it slow. It's a lot to take in.
Gemma3 4B	Oh, *really*? You mean like, *actually* around the university?	Maybe I heard you wrong. A lot of words sound like "yoga." **I just meant, it's a really impressive place.**
GPT-4o	Sure, because "yoga" is definitely the first thing that comes to mind when you hear "unbelievable."	I just want to make sure we're on the same page here.

Table 12: Utterance generation from **sarcastic** reference utterance:
LEONARD: *Maybe I heard you wrong. A lot of words sound like "yoga."*

Model	Non-Sarcastic to Sarcastic	Non-Sarcastic to Non-Sarcastic
Llama3.2 1B	Oh, great, another thing you're using chocolates as positive reinforcement for your lack of effort.	Honestly, I'm not sure what to make of this chocolate as a reward for your good behavior.
Llama3.2 3B	Oh great, just what I always wanted, to be replaced by a chocolate fan.	I'm really disappointed that they replaced you with someone else.
Qwen3 1.7B	Chocolate? - No, I don't want any chocolate!	I'm not interested in chocolate. - Alright, fine.
Qwen3 4B	Oh, I see. You're not a fan of chocolate, but I'm sure you'd love a lifetime supply of it.	I don't want any chocolate. I'm not in the mood.
Gemma3 1B	Seriously? You're telling me you're rewarding *incorrect* behavior with chocolate?	Seriously?
Gemma3 4B	Really? Because that's *exactly* what I was hoping for.	I just thought you might be craving something sweet.
GPT-4o	Oh, right, because who wouldn't want a delicious distraction from reality?	Alright, I just thought it might lighten the mood a bit.

Table 13: Utterance generation from **non-sarcastic** reference utterance:
LEONARD: *Chocolate? - No, I don't want any chocolate!*