MuSaG: A Multimodal German Sarcasm Dataset with Full-Modal Annotations

Aaron Scott, Maike Züfle, Jan Niehues
Karlsruhe Institute of Technology, Germany
aaron.scott@student.kit.edu,
{maike.zuefle, jan.niehues}@kit.edu

Abstract

Sarcasm is a complex form of figurative language in which the intended meaning contradicts the literal one. Its prevalence in social media and popular culture poses persistent challenges for natural language understanding, sentiment analysis, and content moderation. With the emergence of multimodal large language models, sarcasm detection extends beyond text and requires integrating cues from audio and vision. We present MuSaG, the first German multimodal sarcasm detection dataset, consisting of 33 minutes of manually selected and human-annotated statements from German television shows. Each instance provides aligned text, audio, and video modalities, annotated separately by humans, enabling evaluation in unimodal and multimodal settings. We benchmark nine open-source and commercial models, spanning text, audio, vision, and multimodal architectures, and compare their performance to human annotations. Our results show that while humans rely heavily on audio in conversational settings, models perform best on text. This highlights a gap in current multimodal models and motivates the use of MuSaG for developing models better suited to realistic scenarios. We release MuSaG publicly to support future research on multimodal sarcasm detection and human–model alignment.

Keywords: Sarcasm Detection, Multimodality, German Dataset

1. Introduction

Sarcasm represents a complex form of figurative language, often conveying meanings that contradict their literal content. The Cambridge Dictionary defines sarcasm as *the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way*[2]. As such, sarcasm is widespread in user-generated content on social media platforms such as X, Facebook, and Reddit, as well as in popular culture, including sitcoms and movies, where it serves as a key vehicle for humor and mockery (Maynard and Greenwood, 2014).

Detecting sarcasm is essential for applications such as sentiment analysis (Joshi et al., 2017), hate speech detection (Frenda, 2018), and content moderation (Liu et al., 2025), since sarcasm can invert the perceived polarity of a statement. With the growing integration of language models into conversational systems, reliable sarcasm detection becomes increasingly important to ensure appropriate and contextually aware responses. Moreover, with the advent of multimodal large language models (Abouelenin et al., 2025; Comanici et al., 2025; Xu et al., 2025), sarcasm detection extends beyond text, requiring understanding across audio and visual modalities as well.

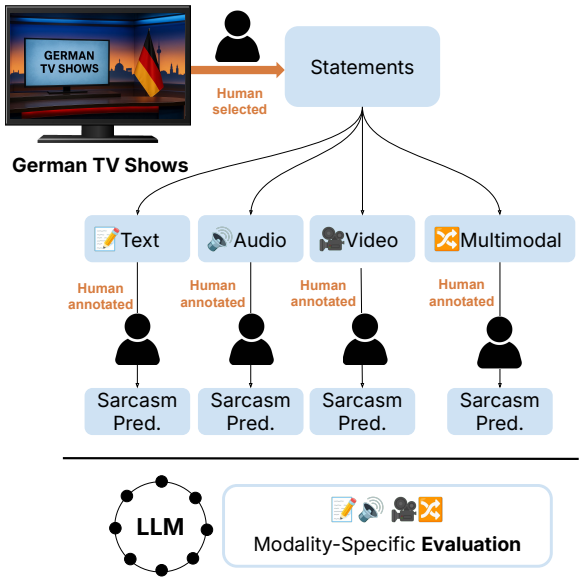

Figure 1: MuSaG, our human annotated German multimodal sarcasm detection dataset.

In text, sarcasm is often indicated by punctuation, hyperbole, or lexical incongruity (Tsur et al., 2010; Davidov et al., 2010). In spoken language, prosodic features such as tone, pitch, or emphasis serve as important auditory cues (Tepperman et al., 2006; Castro et al., 2019), while visual expressions, such as smirks or eye-rolls, can also clearly signal sarcastic intent. Accurate sarcasm detection therefore requires integrating cues across modalities and rec-

[1]MuSaG is available at https://huggingface.co/datasets/sc0ttypee/MuSaG

[2]https://dictionary.cambridge.org/dictionary/english/sarcasm

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 372–392
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Title	Genre	Lang.	# Srcs	Manual select	Human annot.	Single-Mod. annot.	Agreem. avail.	Text	Audio	Vision
Cai et al. (2019)*	Social Media	en	1	×	×	×	×	✓	×	
Schifanella et al. (2016)	Social Media	en	3	×	✓	✓	✓	✓	×	
Yue et al. (2024)	Social Media	en/zh	2	×	✓	✓	✓	✓	×	
Sangwan et al. (2020)	Social Media	en	1	×	(✓)	✓	✓	✓	×	
Alnajjar and Hämäläinen (2021)	TV-shows	es	2	✓	✓	×	×	✓	✓	
Castro et al. (2019)	TV-shows	en	4	×	✓	×	✓	✓	✓	
Zhang et al. (2023)	TV-shows	zh	18	n.s.	✓	×	✓	✓	✓	
Bedi et al. (2023)	TV-shows	hi/en	1	n.s.	✓	×	✓	✓	✓	
Ray et al. (2022)	TV-shows	en	5	n.s.	✓	×	✓	✓	✓	
MuSaG (ours)	TV-shows	de	4	✓	✓	✓	✓	✓	✓	

* Qin et al. (2023) later manually correct the labels in a derivative of this dataset.

Table 1: Comparison of available multimodal sarcasm datasets. Papers for which the criterion is fulfilled only for a subset of the data are marked with (✓), and criteria that are not specified are marked as n.s.

ognizing inconsistencies between them (Pan et al., 2020; Sangwan et al., 2020).

Despite progress in multimodal learning, sarcasm detection remains a challenging task for computational systems (Farabi et al., 2024), as successful interpretation depends on subtle contextual, linguistic, and paralinguistic information. A key limitation is that most existing multimodal sarcasm datasets are in English (Farabi et al., 2024), although sarcasm is a pervasive, multilingual phenomenon. Moreover, existing resources rarely support modality-specific evaluation.

To address this gap, we introduce MuSaG, a German multimodal sarcasm detection dataset comprising 33 minutes of human-annotated statements from German television shows. Each statement has been manually selected rather than relying on automatically tagged data (Schifanella et al., 2016; Cai et al., 2019; Castro et al., 2019; Sangwan et al., 2020). Each instance includes aligned text, audio, and video modalities, all separately annotated by humans, enabling evaluation in multimodal and unimodal settings (text-only, audio-only, vision-only, and their combinations). This is visualized in Fig. 1.

We benchmark nine open-source and commercial models, three text-based, one audio-based, two vision-based, and three multimodal systems, to examine their ability to detect sarcasm and compare their predictions with human annotations. We find that audio provides the strongest unimodal cues for humans, followed by text and then video. In contrast, models perform best on text, indicating that current multimodal systems still struggle to effectively integrate non-textual information. This highlights a gap between text-based model performance and real conversational sarcasm. Furthermore, we analyze the effect of adding broader conversational context and observe that it does not improve performance but instead degrades the models' effectiveness, thereby limiting its usefulness in real-world scenarios.

Our main contributions are as follows:

1. We release the first open, human-annotated German multimodal sarcasm dataset with modality-specific annotations.[1]
2. We evaluate nine state-of-the-art unimodal and multimodal models, both commercial and open-source, across modality configurations.
3. We show that in contrast to humans, current multimodal models fail to leverage audio and visual cues, instead relying primarily on text.

2. Related Work

We briefly look into unimodal datasets, before then discuss multimodal datasets.

Unimodal Sarcasm Detection The importance of sarcasm detection was recognized early, leading to the development of several text-based benchmarks (Tsur et al., 2010; Davidov et al., 2010; González-Ibáñez et al., 2011; Wallace et al., 2014, among others). These datasets were primarily constructed from social media platforms such as Twitter

or from product reviews, focusing on lexical and syntactic cues for sarcasm.

Audio-based sarcasm detection has also been explored, with datasets leveraging prosodic and intonational features directly from speech (Tepperman et al., 2006) or indirectly through transcribed TV dialogues (Joshi et al., 2016).

Multimodal Sarcasm Detection Table 1 provides an overview of publicly available multimodal sarcasm detection datasets and compares them with our proposed MuSaG corpus. Earlier resources primarily focus on social media content, combining text with accompanying images or metadata (Schifanella et al., 2016; Cai et al., 2019; Sangwan et al., 2020; Yue et al., 2024). More recent datasets based on television material (Castro et al., 2019; Ray et al., 2022; Bedi et al., 2023; Zhang et al., 2023) introduce aligned audio–visual components, yet often lack fine-grained modality separation or manual selection of source clips.

Most existing datasets are in English, with only three multilingual exceptions: English–Chinese (Yue et al., 2024), Hindi–English (Bedi et al., 2023), and Spanish (Alnajjar and Hämäläinen, 2021). Among these, only Alnajjar and Hämäläinen (2021) does not rely on automatically collected data. While several datasets include human annotations, only a subset, limited to text-image datasets, provides modality-specific labels (Schifanella et al., 2016; Sangwan et al., 2020; Yue et al., 2024).

To date, no dataset provides full multimodal coverage with modality-specific annotations, an essential requirement for analyzing how multimodal conversational models interpret sarcasm. In contrast, MuSaG offers manually curated, human-annotated German data with independent annotations across all modalities.

3. Dataset

We present MuSaG, a manually curated German multimodal sarcasm dataset enabling analysis across text, audio, and video modalities. This section details the dataset's collection, processing and human annotation, as well as dataset statistics. The dataset will be released on HuggingFace upon paper acceptance.

3.1. Data Collection

We selected four German TV shows known for their explicitly sarcastic style and officially described as such by their producers: *Reschke Fernsehn*[1], *heute show*[2], *Die Carolin Kebekus Show*[3], and *extra 3*[4]. All shows are part of official German public broadcasting productions and provide publicly accessible video recordings. We include videos released after April 2024.

From these sources, we manually collected a balanced set of candidate statements to ensure coverage across speaker gender and potential sarcastic content. In an initial selection phase, we identified short segments that appeared likely sarcastic or clearly non-sarcastic, capturing a range of expressions from overt to subtle. Final sarcasm labels were determined through a subsequent human annotation process as detailed in Section 3.3.

3.2. Data Processing

To prepare the collected segments for multimodal analysis, we downloaded the videos and split the audio and video streams into separate files. Audio files were sampled at 44.1 kHz with a bitrate of 320 kbps, while video files were downsampled to 426×240 pixels at 15 frames per second. The authors manually verified that this resolution and frame rate preserved key visual cues, including facial expressions and gestures.

To provide corresponding textual data, the audio was automatically transcribed using OpenAI Whisper 3 (Radford et al., 2022), and the transcripts were subsequently post-edited by a human annotator with native German proficiency and expertise in multimodal analysis.

3.3. Human Annotation

Multimodal Annotation. To label the dataset, we designed a human annotation process involving 12 participants with strong German language proficiency—11 native speakers and one highly proficient non-native speaker. Each annotator was assigned a subset of the dataset and asked to assign a label ('sarc' or 'non-sarc') based on the audiovisual representation of each statement, reflecting real conversational conditions. Annotators could also leave comments in case of technical issues. Each statement was annotated by three annotators, and the final label was determined using a majority vote. The inter-annotator agreement, measured using Fleiss' Kappa, is 0.623, indicating substantial agreement (Landis and Koch, 1977).

[1] https://www.ardmediathek.de/sendung/reschke-fernsehen/Y3JpZDovL2Rhc2Vyc3RlLm5kci5kZS80ODY3

[2] https://www.zdf.de/shows/heute-show-104

[3] https://www.ardmediathek.de/sendung/die-carolin-kebekus-show/Y3JpZDovL2Rhc2Vyc3RlLmRlL2RpZS1jYXJvbGluLWtlYmVrdXMtc2hvdw

[4] https://www.ndr.de/fernsehen/sendungen/extra_3

		# Statements	# Female Speakers	# Male Speakers	Total Video min	Avg. Video sec	Avg. Words in Transcript
MuSaG	Sarcastic	120	53	57	19.12	9.56 ± 4.68	24.15 ± 11.32
	Non-Sarcastic	94	53	50	13.55	8.65 ± 2.89	21.18 ± 7.78
	Total	214	107	107	32.68	9.16 ± 4.02	22.85 ± 10.03
MuSaG-FullAgree	Sarcastic	96	47	49	14.93	9.33 ± 4.80	23.65 ± 11.48
	Non-Sarcastic	59	38	21	8.75	8.89 ± 2.71	22.25 ± 7.87
	Total	155	85	70	23.67	9.16 ± 4.13	23.12 ± 10.28

Table 2: Dataset statistics for our German multimodal dataset MuSaG and the variant MuSaG-FullAgree, that has full annotator agreement.

Detailed instructions provided to the annotators are included in Fig. 2 in Appendix A (in German), with an English translation in Fig. 3.

Single Modality Annotation. In addition to multimodal labels, we obtained human annotations for isolated modalities as part of the initial annotation task. To avoid bias, each annotator classified a statement in only one modality. Annotators used the same processed data that were later provided to LLMs for classification, enabling direct comparison between human and model performance.

To ensure comparability, annotators were instructed to read, watch, or listen to each statement only once. The inter-annotator agreement for single-modality annotations is slightly lower at 0.594. Detailed instructions to the annotators are provided in Appendix A.

3.4. Dataset Statistics

Our final dataset contains 214 statements, of which 120 are sarcastic and 94 are non-sarcastic. Speaker gender is perfectly balanced, with 107 statements each from female and male speakers. On average, statements contain 22.85 words and are spoken over 9.16 seconds.

The dataset includes three modalities: audio, video, and transcript, with transcripts manually reviewed and corrected. In addition, we release the individual annotations from all human annotators, enabling detailed analysis of agreement and variability. Among the statements, 155 of 214 have full agreement across annotators; we refer to this subset as MuSaG-FullAgree, representing entries with unanimous labels for the audio-video modality.

Comprehensive statistics for both the full dataset and MuSaG-FullAgree can be found in Table 2.

4. Analysis

We now benchmark a range of state-of-the-art open-source and commercial models, both unimodal and multimodal, on the MuSaG dataset. Our analysis examines how well models detect sarcasm across different input modalities and how closely their predictions align with human judgments.

4.1. Experiment Setting

We assess model performance using precision, recall, and F1-score across multiple modality configurations to measure their ability to detect sarcasm from text, audio, and visual information.

For unimodal evaluation, text, audio, and video models are tested on their respective input types (text-only, audio-only, and video-only). To explore potential cross-modal benefits, we additionally evaluate unimodal models with the inclusion of textual input, i.e., audio-text and video-text configurations. Multimodal models are evaluated on single modalities as well as on all available combinations, including text–audio, text–video, and the most realistic, human-like condition: audio–video.

Each model is instructed to classify a statement as either sarcastic or non-sarcastic. We use two prompting strategies, both in English:

1. Generic prompt: *"Decide based on the input whether the given utterance is sarcastic or not sarcastic. Answer only with 'sarc' or 'nonsarc'."*

2. Modality-specific prompt: The basic prompt with an addition to describe the input format (speech, video, or transcript). For example for the transcript, the addition is *"You will be given one short sentence at a time, containing the transcript of a spoken statement."*

For each model, we report results corresponding to the prompting strategy that yielded the best performance. Full prompt templates and model-specific settings are provided in Appendix B.

4.2. Extended Context

To investigate the effect of additional context on classification performance, we include up to 15 seconds of preceding content for each statement. This duration was chosen based on our dataset statistics (Table 2), which show that individual statements

Modality	Model	Precision	Recall	F1
N/A	random baseline	52.15	52.15	52.15
text	Qwen2-7B-Instruct	76.69	64.82	62.83
	Qwen2.5-7B-Instruct	75.11	71.18	71.33
	Qwen3-8B	83.32	83.24	**83.28**
	Phi-4-multimodal-instruct	65.79	65.62	65.68
	Qwen2.5-Omni-7B	79.01	64.59	62.14
	Gemini-2.5-flash	**85.34**	**83.75**	81.71
audio	Qwen2-Audio-7B- Instruct	58.38	57.65	55.18
	Phi-4-multimodal-instruct	55.17	53.76	48.28
	Qwen2.5-Omni-7B	**80.63**	67.78	66.45
	Gemini-2.5-flash	79.01	**71.67**	**66.95**
video	Qwen2-VL-7B-Instruct	56.66	56.76	56.43
	Qwen2.5-VL-7B-Instruct	60.23	52.87	39.65
	Phi-4-multimodal-instruct	21.96	50.00	30.52
	Qwen2.5-Omni-7B	**61.23**	59.29	55.48
	Gemini-2.5-flash	60.59	**60.74**	**60.53**
text-audio	Qwen2-Audio-7B- Instruct	61.50	60.50	58.01
	Phi-4-multimodal-instruct	46.89	49.33	34.41
	Qwen2.5-Omni-7B	79.29	65.12	62.88
	Gemini-2.5-flash	**87.47**	**87.87**	**86.91**
text-video	Qwen2-VL-7B-Instruct	66.00	61.30	59.95
	Qwen2.5-VL-7B-Instruct	74.92	75.15	74.99
	Phi-4-multimodal-instruct	63.79	52.75	37.33
	Qwen2.5-Omni-7B	79.84	66.19	64.33
	Gemini-2.5-flash	**84.09**	**84.49**	**83.63**
audio-video	Phi-4-multimodal-instruct	62.55	52.27	37.03
	Qwen2.5-Omni-7B	**76.45**	66.00	64.55
	Gemini-2.5-flash	74.87	**74.92**	**74.89**
text-audio-video	Phi-4-multimodal-instruct	62.34	59.26	54.17
	Qwen2.5-Omni-7B	81.37	72.80	72.79
	Gemini-2.5-flash	**83.42**	**83.34**	**83.38**

Table 3: Results on our newly proposed MuSaG dataset for different modalities. We report the macro average over the sacrcastic and non-sarcastic class. For each modality, we report modality-specific models (, ,) and multimodal models ().

are on average around 10 seconds long, meaning 15 seconds reliably captures at least one additional preceding utterance. Example utterances with extended context are shown in Table 10 in Appendix D. Prompts are constructed to provide this extended context to the model, with the target statement explicitly indicated in the prompt using its transcript. Full prompts for the extended context condition are provided in Appendix B.2.

4.2.1. Models

We benchmark nine different models on MuSaG, comprising eight open-source and one commercial model, Gemini (Comanici et al., 2025). The selection includes three text-based LLMs, one audio model, two vision models, and three fully multimodal LLMs, as detailed below:

- **Text-based LLMs:** Qwen3-8B (Yang et al., 2025), Qwen2.5-7B-Instruct (Qwen et al., 2025), Qwen2-7B-Instruct (Yang et al., 2024).
- **Modality-specific Models:** Qwen2-Audio-7B-Instruct (Chu et al., 2024); Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Qwen2-VL-7B-Instruct (Wang et al., 2024).
- **Multimodal LLMs:** Phi-4-Multimodal-Instruct (Abouelenin et al., 2025), Qwen2.5-

Modality	Model	Transcript	True Label	Disagreement
Audio	Gemini-2.5-flash	*„Ist jetzt vielleicht auf Anhieb nicht so schnell zu erkennen, es sind 83 Prozent Wirtschaft in einem Ausschuss. Und wenn ich das richtig überblicke, sind das deutlich mehr als 50 Prozent."*	non-sarc	H: sarc M: non-sarc
		„Ewige Chemikalien, man findet sie überall, im Regen, in menschlichem Blut, selbst in der Arktis wurde PFAS jetzt schon nachgewiesen, in der Arktis."	non-sarc	H: non-sarc M: sarc
Video	Gemini-2.5-flash	*„In der Regel. Was heißt denn das? Immer? Manchmal? Nur wenn die Sonne in Konjunktion mit Jupiter steht?"*	sarc	H: sarc M: non-sarc
		„Also, gerade kleine und mittelständische Unternehmen leiden wohl besonders unter der Bürokratie."	non-sarc	H: non-sarc M: sarc
Text	Qwen3-8B	*„Söder hat aktuell ein bisschen Zoff mit dem CDU-Mann Daniel Günther. Günther sagte, Söder solle nicht die ganze Zeit von Schwarz-Grün reden. Söder hat seriös und ohne persönlich zu werden geantwortet."*	sarc	H: sarc M: non-sarc
		„Why? Deutschlands Behörden, ist kein Witz, nutzen aktuell um die 10.000 verschiedene Softwarelösungen."	non-sarc	H: non-sarc M: sarc

Table 4: Examples of human–model disagreements per modality. For each modality, we show two cases: one where humans rated an utterance as sarcastic while the model did not (H: sarc, M: non-sarc), and one where the model predicted sarcasm while humans did not (H: non-sarc, M: sarc). H = human annotators, M = model prediction.

Omni-7B (Xu et al., 2025), and Gemini-2.5-Flash (Comanici et al., 2025).

4.3. Results

In the following, we report the results on MuSaG, for single and multimodal scenarios.

4.3.1. Unimodal Performance

We first evaluate each model using a single modality to understand how well sarcasm can be detected from transcript, audio, or video alone in Table 3.

Text modality. Among the text-only LLMs, Qwen3-8B achieves the best results with an 83.28 F1, outperforming smaller and older versions. Multimodal models evaluated on text alone generally perform slightly worse than dedicated text LLMs.

Audio modality. For audio-only input, unimodal audio LLMs show modest performance, with Qwen2-Audio-7B-Instruct achieving 55.18 F1. Interestingly, the multimodal models Qwen2.5-Omni-7B and Gemini-2.5-flash outperform the audio-specific model: Gemini-2.5-flash reaches an F1-score of 66.95, but only Qwen2.5-Omni-7B is able to leverage prosodic features to detect sarcasm and improve over its text-only performance.

Video modality. Sarcasm detection from video alone is particularly challenging. Vision-only models achieve moderate performance (56.43 F1), while multimodal models show mixed results. Gemini-2.5-flash again performs best (60.53 F1), whereas Phi-4-Multimodal-Instruct performs worse than chance.

4.3.2. Multimodal Performance

We next examine model performance when multiple modalities are available, including combinations of text, audio, and video.

Text–audio and text–video. Combining transcripts with audio or video improves performance for most models. The commercial Gemini 2.5 Flash model achieves the highest scores for both text–audio (86.91 F1) and text–video (83.63 F1) inputs, showing a clear benefit from multimodal integration compared to single-modality settings. This improvement, however, is not consistent across all

Modality	Model	Precision		Recall		F1		κ
		FullAgr.	ΔStand.	FullAgr.	ΔStand.	FullAgr.	ΔStand.	
N/A	random baseline	49.34		49.31		49.19		–
text	Qwen3-8B	87.55	-4.23	88.01	-4.77	87.76	-4.48	–
	human	85.88	-3.67	86.55	-4.07	86.14	-3.84	53.13
audio	Gemini-2.5-flash	76.82	+2.19	73.44	-1.77	66.83	+0.12	–
	human	**88.35**	-4.72	**87.62**	-3.90	**87.93**	-4.21	68.01
video	Gemini-2.5-flash	59.34	+1.25	59.87	+ 0.87	59.1	+1.43	–
	human	69.12	-3.52	68.24	-4.13	68.55	-4.46	31.20
audio-video	Gemini-2.5-flash	79.44	-4.57	79.8	-4.88	79.61	-4.72	
	human	100.00	–	100.00	–	100.00	–	100

Table 5: Results on MuSaG-FullAgree, the subset with full human agreement. We compare the best-performing models for each modality according to Table 3 against their corresponding results on MuSaG and human single-modality annotations. Human audio–video annotations are treated as the gold standard, reflecting how people naturally integrate multimodal cues in communication. Δ*Stand.* indicates the difference MuSaG-FullAgree - MuSaG.

models. For instance, Qwen2.5-Omni-7B benefits from combining text with video, but not with audio.

Audio-video. When only audio and video are available, performance decreases relative to transcript-inclusive configurations, but still exceeds that of single-modality (audio-only or video-only) setups. Gemini-2.5-Flash achieves the best F1-score (74.89). This not only confirms that the transcript remains the most informative modality, but also highlights that in conditions resembling real human communication, where information is conveyed through speech and visual cues, commercial models still outperform open-source alternatives.

Text-audio-video. The full multimodal condition, including transcript, audio, and video, provides strong, but surprisingly not the best performance. Gemini-2.5-Flash achieves an F1-score of 83.38, slightly lower than its transcript–audio performance (F1-score of 86.91). We hypothesize that the addition of video can sometimes introduce noise or distract from the most informative cues. In contrast, Qwen2.5-Omni-7B benefits from combining all three modalities, achieving an F1 of 72.79. These results suggest that transcript and audio carry the majority of sarcasm-relevant information, while video cues may only provide marginal gains or, for some models, slightly reduce performance.

Best performing open models. Among all evaluated systems, Gemini 2.5 Flash achieves the highest overall performance across all modalities except text-only input. Focusing on open-source models, Qwen2.5-Omni-7B consistently outperforms all other open models, showing strong results especially on audio-only input. Adding video or text to audio does not further improve its performance, only the full multimodal configuration achieves the best results overall.

4.4. Comparison with Human Classification

We now assess how model predictions align with human classifications, using MuSaG-FullAgree as a gold standard. This subset contains only examples with full annotator agreement, reflecting how people naturally perceive sarcasm in multimodal communication. We evaluate the best-performing model per modality, Qwen3-8B for text, and Gemini-2.5-flash for audio, video, and audio-video, against human classifications based on the corresponding single modality (transcript, audio, or video only).

Representative Examples. Disagreements between humans and models arise in both directions: models sometimes predict sarcasm where humans do not, and vice versa. Representative examples of such disagreements across all modalities are shown in Table 4.

Humans vs. Models. Table 5 compares human and model performance on MuSaG-FullAgree across all modalities, where humans and models were asked to classify sarcasm based solely on a single modality (text, audio, or video).

Humans derive the most reliable cues from audio (87.93 F1), suggesting that prosodic features such as tone and intonation provide strong indicators of sarcasm. In contrast, multimodal models are not yet able to leverage these cues effectively: humans outperform models by substantial margins, nearly 21 F1 points for audio and 10 F1 points for video. For video-only input, human performance drops to 68.6 F1, indicating that visual cues alone are

Modality	Model	Precision		Recall		F1	
		Context	ΔStand.	Context.	ΔStand.	Context	ΔStand.
N/A	random baseline	52.15	–	52.15	–	52.15	–
text	Qwen3-8B	45.61	+37.71	45.77	+37.47	45.57	+37.71
audio	Gemini-2.5-flash	55.31	+23.7	54.99	+16.68	53.01	+13.94
video	Gemini-2.5-flash	53.48	+7.11	52.70	+8.04	51.10	+9.43
audio-video	Gemini-2.5-flash	58.35	+16.52	55.15	+19.77	52.2	+22.69

Table 6: Results on MuSaG with 15 seconds of extended context (*Context*) in comparison to statements without context (*Stand.*) for the best performing models for each modality according to Table 3. Δ*Stand.* indicates the difference $Stand. - Context$. None of the models benefits from additional context.

often insufficient for sarcasm detection, with model performance similarly decreasing to 58.5 F1.

Generally, most models perform better on MuSaG-FullAgree than on the full dataset (indicated in blue in Table 5), reflecting that examples with perfect human agreement are likely less ambiguous and thus easier to interpret.

Results on MuSaG-FullAgree for all models, not only the best performing ones, can be found in Table 9 in Appendix C.

4.5. Extending the Context

Lastly, we investigate whether providing models with additional temporal context improves sarcasm detection performance, for the best model for each modality according to Table 3. Specifically, we extend the input by 15 seconds surrounding the target utterance, while explicitly including the sentence transcript to be classified in the model prompt. We report these results in Table 6.

Surprisingly, this makes the task significantly harder for all models, and performance drops to chance. We hypothesize that the models, even when explicitly provided with the target utterance, may struggle to attribute their decision to the correct segment within the extended input. The added temporal context likely introduces distracting or conflicting cues, making it harder for the model to focus on the relevant part of the provided signal.

Qualitative examples illustrating these three outcomes, 1) cases where context helped, 2) hurt, or 3) had no effect on classification, are shown in Table 10 in Appendix D. Notably, the *Degraded* example in Table 10 suggests that surrounding utterances can introduce misleading cues: the phrase *„Das klingt ja erstmal so, als wäre jetzt alles in trockenen Tüchern"* adopts a tone that superficially resembles sarcasm, apparently causing three out of four models to misclassify the subsequent non-sarcastic target sentence.

This finding has important implications for real-world deployment. In natural conversations, utterances never occur in isolation, they are embedded in a broader discourse with preceding and following context. If models are already misled by a mere 15 seconds of surrounding signal, their reliability in practical applications such as opinion mining, content moderation, or social media analysis may be substantially lower than isolated-utterance benchmarks suggest. Addressing this sensitivity to context should therefore be a priority for future work on sarcasm detection in the wild.

5. Conclusion

We introduced MuSaG, the first manually curated German multimodal sarcasm dataset with independent annotations for text, audio, and video modalities. The dataset includes both full statements and modality-specific labels, enabling fine-grained analysis of multimodal sarcasm understanding. Moreover, we also release MuSaG-FullAgree, a subset with full annotator agreement (annotated in audio–video), which can serve as a gold standard for how humans perceive sarcasm and can be used to evaluate how well humans and models perform when only partial modalities are available. MuSaG will be released publicly to support research on multimodal language models.

Our benchmarking experiments show that humans rely primarily on audio cues, followed by text and then video, indicating that the strongest signals for sarcasm lie in prosody and intonation. In contrast, models perform strongest on text, revealing that they fail to fully exploit audio cues and are not yet capable of genuine multimodal understanding. While commercial models generally outperform open-source models, all architectures struggle to integrate non-textual information effectively.

These findings underscore the challenges of building systems for nuanced sarcasm detection, highlighting that current multimodal models are still unable to effectively leverage non-textual cues, emphasizing the value of MuSaG as a benchmark for developing and evaluating truly multimodal models.

6. Ethical Considerations

All annotators participated voluntarily and will be acknowledged by name upon paper acceptance. The dataset only includes publicly available content, and we release links to the original videos rather than the video files themselves. Researchers should note that sarcasm detection can reflect cultural and subjective biases.

7. Acknowledgements

Part of this work received support from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People).

8. Bibliographical References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. ArXiv:2503.01743 [cs].

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. ArXiv:2502.13923 [cs].

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio Technical Report. ArXiv:2407.10759 [eess].

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, Mu Cai, Mohammed Badawi, Corey Fry, Ale Hartman, Daniel Zheng, Eric Jia, James Keeling, Annie Louis, Ying Chen, Efren Robles, Wei-Chih Hung, Howard Zhou, Nikita Saxena, Sonam Goenka, Olivia Ma, Zach Fisher, Mor Hazan Taege, Emily Graves, David Steiner, Yujia Li, Sarah Nguyen, Rahul Sukthankar, Joe Stanton, Ali Eslami, Gloria Shen, Berkin Akin, Alexey Guseynov, Yiqian Zhou, Jean-Baptiste Alayrac, Armand Joulin, Efrat Farkash, Ashish Thapliyal, Stephen Roller, Noam Shazeer, Todor Davchev, Terry Koo, Hannah Forbes-Pollard, Kartik Audhkhasi, Greg Farquhar, Adi Mayrav Gilady, Maggie Song, John Aslanides, Piermaria Mendolicchio, Alicia Parrish, John Blitzer, Pramod Gupta, Xiaoen Ju, Xiaochen Yang, Puranjay Datta, Andrea Tacchetti, Sanket Vaibhav Mehta, Gregory Dibb, Shubham Gupta, Federico Piccinini, Raia Hadsell, Sujee Rajayogam, Jiepu Jiang, Patrick Griffin, Patrik Sundberg, Jamie Hayes, Alexey Frolov, Tian Xie,

Adam Zhang, Kingshuk Dasgupta, Uday Kalra, Lior Shani, Klaus Macherey, Tzu-Kuo Huang, Liam MacDermed, Karthik Duddu, Paulo Zacchello, Zi Yang, Jessica Lo, Kai Hui, Matej Kastelic, Derek Gasaway, Qijun Tan, Summer Yue, Pablo Barrio, John Wieting, Weel Yang, Andrew Nystrom, Solomon Demmessie, Anselm Levskaya, Fabio Viola, Chetan Tekur, Greg Billock, George Necula, Mandar Joshi, Rylan Schaeffer, Swachhand Lokhande, Christina Sorokin, Pradeep Shenoy, Mia Chen, Mark Collier, Hongji Li, Taylor Bos, Nevan Wichers, Sun Jae Lee, Angéline Pouget, and Santhosh Thangaraj. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. ADS Bibcode: 2025arXiv250706261C.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8020–8028. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Simona Frenda. 2018. The role of sarcasm in hate speech. a multilingual perspective. In *Proceedings of the Doctoral Symposium of the XXXIV-International Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, volume Vol-2251.

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

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174. Publisher: International Biometric Society.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian J. McAuley, Wei Ai, and Furong Huang. 2025. Large language models and causal inference in collaboration: A comprehensive survey. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7668–7684. Association for Computational Linguistics.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4238–4243, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. ArXiv:2412.15115 [cs].

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. ArXiv:2409.12191 [cs].

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-Omni Technical Report. ArXiv:2503.20215 [cs].

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu,

Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. ArXiv:2505.09388 [cs].

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. ArXiv:2407.10671 [cs].

9. Language Resource References

Alnajjar, Khalid and Hämäläinen, Mika. 2021. *¡Qué maravilla! Multimodal Sarcasm Detection in Spanish: a Dataset and a Baseline*. Association for Computational Linguistics. PID https://zenodo.org/records/4701383.

Bedi, Manjot and Kumar, Shivani and Akhtar, Md Shad and Chakraborty, Tanmoy. 2023. *Multi-Modal Sarcasm Detection and Humor Classification in Code-Mixed Conversations*. IEEE Transactions on Affective Computing. PID https://github.com/LCS2-IIITD/MSH-COMICS.

Cai, Yitao and Cai, Huiyu and Wan, Xiaojun. 2019. *Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model*. Association for Computational Linguistics. PID https://github.com/headacheboy/data-of-multimodal-sarcasm-detection.

Castro, Santiago and Hazarika, Devamanyu and Pérez-Rosas, Verónica and Zimmermann, Roger and Mihalcea, Rada and Poria, Soujanya. 2019. *Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper)*. Association for Computational Linguistics. PID https://github.com/soujanyaporia/MUStARD.

Davidov, Dmitry and Tsur, Oren and Rappoport, Ari. 2010. *Semi-Supervised Recognition of Sarcasm in Twitter and Amazon*. Association for Computational Linguistics.

González-Ibáñez, Roberto and Muresan, Smaranda and Wacholder, Nina. 2011. *Identifying Sarcasm in Twitter: A Closer Look*. Association for Computational Linguistics.

Joshi, Aditya and Tripathi, Vaibhav and Bhattacharyya, Pushpak and Carman, Mark J. 2016. *Harnessing Sequence Labeling for Sarcasm Detection in Dialogue from TV Series 'Friends'*. Association for Computational Linguistics.

Qin, Libo and Huang, Shijue and Chen, Qiguang and Cai, Chenran and Zhang, Yudi and Liang, Bin and Che, Wanxiang and Xu, Ruifeng. 2023. *MMSD2.0: Towards a Reliable Multimodal Sarcasm Detection System*. Association for Computational Linguistics. PID https://github.com/JoeYing1019/MMSD2.0.

Ray, Anupama and Mishra, Shubham and Nunna, Apoorva and Bhattacharyya, Pushpak. 2022. *A Multimodal Corpus for Emotion Recognition in Sarcasm*. European Language Resources Association.

Sangwan, Suyash and Akhtar, Md Shad and Behera, Pranati and Ekbal, Asif. 2020. *I didn't mean what I wrote! Exploring Multimodality for Sarcasm Detection*. 2020 International Joint Conference on Neural Networks (IJCNN). PID http://www.iitp.ac.in/ai-nlp-ml/resources.htm. ISSN: 2161-4407.

Schifanella, Rossano and de Juan, Paloma and Tetreault, Joel and Cao, LiangLiang. 2016. *Detecting Sarcasm in Multimodal Social Platforms*. Association for Computing Machinery, MM '16.

Tepperman, Joseph and Traum, David and Narayanan, Shrikanth. 2006. *yeah right: sarcasm recognition for spoken dialogue systems*. ISCA.

Tsur, Oren and Davidov, Dmitry and Rappoport, Ari. 2010. *ICWSM — A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews*. Proceedings of the International AAAI Conference on Web and Social Media.

Wallace, Byron C. and Choe, Do Kook and Kertz, Laura and Charniak, Eugene. 2014. *Humans Require Context to Infer Ironic Intent (so Computers Probably do, too)*. Association for Computational Linguistics. PID https://github.com/bwallace/ACL-2014-irony.

Yue, Tan and Shi, Xuzhao and Mao, Rui and Hu, Zonghai and Cambria, Erik. 2024. *SarcNet: A Multilingual Multimodal Sarcasm Detection Dataset*. ELRA and ICCL. PID https://github.com/yuetanbupt/SarcNet.

Zhang, Yazhou and Yu, Yang and Guo, Qing and Wang, Benyou and Zhao, Dongming and Uprety, Sagar and Song, Dawei and Li, Qiuchi and Qin, Jing. 2023. *CMMA: benchmarking multi-affection detection in chinese multi-modal conversations*. Curran Associates Inc., NIPS '23. PID https://github.com/annoymity2022/Chinese-Dataset.

A. Human Annotation Instructions

Detailed instructions for the human annotations can be found in Fig. 2. For reference, we also provide the English translations in Fig. 3, these have not been used in the annotation process.

B. Technical Details

B.1. General Prompts

We experiment with different prompts for sarcasm detection. The exact prompts are listed in Fig. 4, Fig. 5 and Fig. 6.

B.2. Extended context

For extended context, we modify the prompt to specify for which utterance the label should be provided, as in Fig. 7.

B.3. Inference Parameters

Table 7 lists the generation parameters for different configurations.

Config	max new tokens*	sampling	beams	temp
1	3	true	2	0.7
2	3	true	2	1.8
3	3	false	1	-

* For Gemini, we set max new tokens to 5.
* When enabling thinking, we set max new tokens to 2000.

Table 7: Different Generation Configurations.

We use the best performing of the three configurations for each model. This mapping is given in Table 8.

C. Results MuSaG-FullAgreement

We report results for all models on MuSaG-FullAgreement in Table 9. In Table 5 in the main paper, we also report human evaluation on different modalities.

D. Examples Predictions for Extended Context

Table 10 illustrates how extended context influences model predictions. We show three cases: examples where additional context *improved* classification (all models corrected their prediction), where it *degraded* classification (previously correct models were misled by the surrounding utterances), and where context had *no impact* (models predicted correctly regardless). In the *With context* rows, the target utterance is highlighted in **bold**.

Model	Modality	Transformers version	Generatrion	Prompt
Qwen3-8B		4.55.2	config 2	modality specific
Qwen2.5-7B-Instruct		4.55.2	config 3	modality specific
Qwen2-7B-Instruct		4.55.2	config 3	modality specific
Qwen2-Audio-7B-Instruct		4.55.2	config 3	general
		4.55.2	config 3	modality specific
Qwen2-VL-7B-Instruct		4.56.0.dev0	config 3	general
		4.56.0.dev0	config 1	general
Qwen2.5-VL-7B-Instruct		4.56.0.dev0	config 3	general
		4.56.0.dev0	config 3	general
Phi-4-multimodal-instruct		4.48.2	config 2	modality specific
		4.48.2	config 1	modality specific
		4.48.2	config 2	modality specific
		4.48.2	config 3	general
		4.48.2	config 2	modality specific
		4.48.2	config 3	general
		4.48.2	config 1	general
Qwen2.5-Omni-7B		4.52.3	config 2	modality specific
		4.52.3	config 1	modality specific
		4.52.3	config 3	modality specific
		4.52.3	config 1	modality specific
		4.52.3	config 2	modality specific
		4.52.3	config 1	modality specific
		4.52.3	config 2	modality specific
Gemini-2.5-flash		n.s.	config 3	modality specific
		n.s.	config 3	modality specific
		n.s.	config 3	modality specific
		n.s.	config 3	modality specific
		n.s.	config 3	modality specific
		n.s.	config 3	modality specific
		n.s.	config 3	modality specific

Table 8: The model configurations and prompts that returned the results presented in Table 3. Since we access Gemini-2.5-flash through the API, there is no transformer version specified.

Anleitung zur Umfrage

Diese Umfrage ist Teil einer Bachelorarbeit zum Thema Sarkasmus in deutschen Fernsehshows. Ich habe einzelne Aussagen herausgesucht, die jetzt als sarkastisch oder nicht sarkastisch eingestuft werden sollen.

Im Anhang befinden sich zwei Excel-Dateien, eine für die Klassifizierung und eine mit ein paar Fragen zur Person und zur Bearbeitung.

1. Formate der Aussagen

Die Aussagen liegen in vier verschiedenen Formaten von:

- Nur Text – maschinell erstellte Transkripte der Aussagen
- Nur Video – Video in reduzierter Qualität, ohne Ton
- Nur Audio – Tonspur der Aussage
- Video mit Audio – Video in guter Qualität mit Ton

Audio- und Videodateien sind über einen Google-Drive-Link in der Tabelle zugänglich und sollten sich durch einfaches Anklicken öffnen lassen. Falls das nicht funktioniert, bitte direkt Bescheid geben.

2. Vorgehensweise bei der Klassifizierung

- Jede Aussage soll nur ein einziges mal gelesen/gehört/geschaut werden.
- Direkt danach muss entschieden werden, ob die Aussage eher sarkastisch oder nicht
- sarkastisch ist.
- Pausen können beliebig eingelegt werden, die Bearbeitung muss nicht in einem Durchgang
- erfolgen.
- Am Ende müssen alle Aussagen klassifiziert sein – auch wenn Unsicherheit besteht. Bitte in diesem Fall die tendenziell passendere Option wählen und dies ggf. in der Kommentarspalte vermerken.

3. Klassifizierung in der Excel-Datei

Die Excel-Datei enthält vier Tabellenblätter (je eines pro Format) mit einer ähnlichen Anzahl an Aussagen.

In jeder Zeile befindet sich:

1. Die ID (dient nur der internen Zuordnung und kann ignoriert werden)
2. Die Aussage (als Text oder als Link zur Datei)
3. Spalte „classification“ – Auswahl über ein Dropdown-Menü:
    1. sarc = Aussage ist sarkastisch
    2. non-sarc = Aussage ist nicht sarkastisch
4. Kommentarspalte – hier können Anmerkungen, Auffälligkeiten oder Hinweise auf Unsicherheit notiert werden.

4. Persönliche Einschätzung

Zusätzlich gibt es eine kurze, freiwillige Befragung zu persönlichen Einschätzungen. In einer separaten Excel-Datei schicke ich einen kurzen Fragebogen, der einige persönliche Fragen und Fragen zur Bearbeitung enthält. Alle gesammelten Daten werden vollständig anonymisiert

5. Abschluss

Nach der Bearbeitung die Excel-Datei speichern und an mich zurücksenden. Bei Fragen oder technischen Problemen stehe ich jederzeit zur Verfügung.

Figure 2: Instructions for human annotators in German.

Instructions for the survey

This survey is part of a bachelor's thesis on sarcasm in German television shows. I have selected individual statements that are now to be classified as sarcastic or non-sarcastic.

Attached are two Excel files, one for classification and one with a few questions about the person and the processing.

1. Formats of statements

The statements are available in four different formats:

- Text only – machine-generated transcripts of the statements
- Video only – reduced-quality video without sound
- Audio only – audio track of the statement
- Video with audio – high-quality video with sound

Audio and video files are accessible via a Google Drive link in the table and should open with a simple click. If this does not work, please let us know immediately.

2. Classification procedure

- Each statement should only be read/heard/viewed once.
- Immediately afterwards, you must decide whether the statement is sarcastic or not sarcastic.
- You can take breaks as needed; you do not have to complete the task in one sitting.
- At the end, all statements must be classified — even if you are unsure. In this case, please choose the option that seems more appropriate and note this in the comments column if necessary.

3. Classification in the Excel file

The Excel file contains four spreadsheets (one for each format) with a similar number of statements.

Each row contains:

1. The ID (used only for internal assignment and can be ignored)
2. The statement (as text or as a link to the file)
3. Column “classification” – selection via a drop-down menu:
    1. sarc = statement is sarcastic
    2. non-sarc = statement is not sarcastic
4. Comment column – comments, anomalies, or notes on uncertainty can be noted here.

4. Personal assessment

In addition, there is a short, voluntary survey on personal assessments. I will send a short questionnaire in a separate Excel file, which contains some personal questions and questions about the processing. All collected data will be completely anonymized.

5. Completion

After editing, save the Excel file and send it back to me. If you have any questions or technical problems, please do not hesitate to contact me.

Figure 3: Instructions for human annotators in English, for reference. During the annotation process, the German instructions were used.

General Prompt:

```
Decide based on the input, if the given utterance is sarcastic or not
sarcastic.
Answer ONLY with 'sarc' or 'non-sarc'.
```

Modality Specific Prompt - text:

```
Decide based on the input, if the given utterance is sarcastic or not
sarcastic.
Answer ONLY with 'sarc' or 'non-sarc'.
You will be given ONE short sentence at a time, containing the
transcript of a spoken statement.
Examples:
Input: Na das läuft ja mal wieder super!
Output: sarc
Input: Heute morgen war das Wetter eher schlecht.
Output: non-sarc
ANSWER ONLY WITH 'sarc' OR 'non-sarc'!
```

Modality Specific Prompt - audio:

```
Decide based on the input, if the given utterance is sarcastic or not
sarcastic.
Answer ONLY with 'sarc' or 'non-sarc'.
You will receive ONE short audio clip at a time with NO transcript or
video available.
Base your decision solely on the audio recording of their speech.
Use ONLY vocal cues such as tone, pitch, pacing, rhythm, stress,
intonation, and prosody to decide if the speakers intent is sarcastic.
Examples:
Audio example 1: Speaker uses exaggerated rising intonation and slow
pacing in a positive phrase that sounds insincere
Output: sarc
Audio example 2: Speaker uses normal pitch, steady pacing, and neutral
 tone in a factual statement
Output: non-sarc
ANSWER ONLY WITH 'sarc' OR 'non-sarc'!
```

Modality Specific Prompt - video:

```
Decide based on the input, if the given utterance is sarcastic or not
sarcastic.
Answer ONLY with 'sarc' or 'non-sarc'.
You do NOT have access to audio or transcript.
Base your  decision solely on the video of the speaker while speaking.
Focus exclusively on visual sarcasm cues such as:
- Facial expressions (e.g., smirks, raised eyebrows, eye rolls)
- Gestures and hand movements
- Body language and posture
Use these visual signals to decide if the speakers intent is sarcastic
.
Examples:
Input Video: Speaker rolls eyes and smirks while speaking
Output: sarc
Input Video: Speaker maintains neutral expression and relaxed posture
Output: non-sarc
ANSWER ONLY WITH 'sarc' OR 'non-sarc'!
```

Figure 4: Prompts for multimodal sarcasm detection, single-modality.

Modality Specific Prompt - audio-text:

```
Decide based on the input, if the given utterance is sarcastic or not
sarcastic.
Answer ONLY with 'sarc' or 'non-sarc'.
You are provided with two inputs:
1. A transcript of the spoken text (analyze phrasing, irony,
exaggeration, and contradiction)
2. An audio recording of the speech (analyze prosody, tone, pitch,
pacing, and intonation)
Use BOTH inputs together to decide if the speaker's intent is
sarcastic.
Examples:
Input Transcript: Na das Wetter ist ja mal wieder super!
Input Audio: Speaker uses slow pacing and exaggerated rising
intonation
Output: sarc
Input Transcript: Heute morgen war das Wetter eher schlecht.
Input Audio: Speaker uses neutral tone and steady pacing
Output: non-sarc
ANSWER ONLY WITH 'sarc' OR 'non-sarc'!
```

Modality Specific Prompt - video-text:

```
Decide based on the input, if the given utterance is sarcastic or not
sarcastic.
Answer ONLY with 'sarc' or 'non-sarc'.
You are provided with two kinds of input:
1. A transcript of the spoken text (what is said)
2. A video of the speaker while speaking (how it is said)
Analyze the transcript for linguistic cues such as irony,
contradiction, exaggeration, and phrasing.
Simultaneously analyze the video for visual cues including facial
expressions (e.g., smirks, raised eyebrows, eye rolls), gestures,
posture, and body language that typically signal sarcasm.
Use BOTH modalities together to make your judgment.
Examples:
Input Transcript: Na das Wetter ist ja mal wieder super!
Input Video: Speaker rolls eyes and smirks while saying the sentence
Output: sarc
Input Transcript: Heute morgen war das Wetter eher schlecht.
Input Video: Speaker maintains neutral facial expression and relaxed
posture
Output: non-sarc
ANSWER ONLY WITH 'sarc' OR 'non-sarc'!
```

Modality Specific Prompt - audio-video:

```
Decide based on the input, if the given utterance is sarcastic or not
sarcastic.
Answer ONLY with 'sarc' or 'non-sarc'.
You are provided with two inputs:
1. An audio recording of the speaker (analyze prosody, tone, pitch,
pacing, intonation)
2. A video recording of the speaker (analyze facial expressions such
as smirks, raised eyebrows, eye rolls, and body language)
Use BOTH audio and video cues together to judge if the speakers intent
 is sarcastic.
Examples:
Audio: The speaker uses exaggerated rising intonation and slow pacing
on a positive phrase
Video: The speaker smirks and raises eyebrows while speaking
Output: sarc
Audio: The speaker uses neutral tone and steady pacing
Video: The speaker maintains relaxed posture and neutral facial
expression
Output: non-sarc
ANSWER ONLY WITH 'sarc' OR 'non-sarc'!
```


Figure 5: Prompts for multimodal sarcasm detection, different modality combinations.

Modality Specific Prompt -audio-video-text:

```
Decide based on the input, if the given utterance is sarcastic or not
sarcastic.
Answer ONLY with 'sarc' or 'non-sarc'.
You are provided with three inputs:
1. A transcript of the spoken text (analyze linguistic cues such as
irony, exaggeration, contradiction)
2. An audio recording of the speech (analyze prosody, tone, pitch,
pacing, and intonation)
3. A video of the speaker while speaking (analyze facial expressions
like smirks, raised eyebrows, eye rolls, as well as gestures and body
language)
Use all three modalities together to accurately judge if the speakers
intent is sarcastic.
Examples:
Transcript: Na das Wetter ist ja mal wieder super!
Audio: Speaker uses exaggerated rising intonation and slow pacing
Video: Speaker smirks and raises eyebrows while speaking
Output: sarc
Transcript: Heute morgen war das Wetter eher schlecht.
Audio: Speaker uses neutral tone and steady pacing
Video: Speaker maintains relaxed posture and neutral facial expression
Output: non-sarc
ANSWER ONLY WITH 'sarc' OR 'non-sarc'!
```

Figure 6: Prompts for multimodal sarcasm detection, including all modalities.

Modality Specific Prompt -audio-video-text

```
You are an expert at detecting sarcasm in text data.
Your task is to classify a TARGET STATEMENT based on the isolated text
 data of the statement.
To do this, you are provided with the text data of the TARGET
STATEMENT including up to 15 seconds of leading CONVERSTAION CONTEXT.
For identification of the statement to classify, the TARGET STATEMENT
is again cited in text form.
Use the CONVERSATIONAL CONTEXT only to interpret the target; do not
classify the context itself.
Examples:
Context: We've been stuck in traffic for an hour. Oh great, perfect
timing for a road trip.
Target: Oh great, perfect timing for a road trip.
Output: sarc
Context: I finished my report early today.
Target: I finished my report early today.
Output: non-sarc
Context: We forgot the keys. That was an absolutely brilliant idea.
Target: That was an absolutely brilliant idea.
Output: sarc
Context: The sun rises in the east.
Target: The sun rises in the east.
Output: non-sarc
Classify the TARGET STATEMENT using ONLY the provided text data!
Classify ONLY the TARGET STATEMENT as 'sarc' (sarcastic) or 'non-sarc
'.
ANSWER ONLY WITH 'sarc' OR 'non-sarc'!
```

Figure 7: Prompts for multimodal sarcasm detection, including all modalities.

Modality	Model	Precision	Recall	F1
N/A	random baseline	49.34	49.31	49.19
text	Qwen2-7B-Instruct	82.59	69.30	70.01
	Qwen2.5-7B-Instruct	78.2	72.95	74.03
	Qwen3-8B	**87.55**	**88.01**	**87.76**
	Phi-4-multimodal-instruct	69.13	68.72	68.9
	Qwen2.5-Omni-7B	79.3	63.37	62.31
	Gemini-2.5-flash	85.13	87.17	84.37
audio	Qwen2-Audio-7B- Instruct	59.31	58.97	54.79
	Phi-4-multimodal-instruct	55.54	54.03	45.68
	Qwen2.5-Omni-7B	**83.81**	68.12	**68.50**
	Gemini-2.5-flash	76.82	**73.44**	66.83
video	Qwen2-VL-7B-Instruct	59.96	60.46	58.50
	Qwen2.5-VL-7B-Instruct	**61.60**	53.51	37.48
	Phi-4-multimodal-instruct	19.03	50.00	27.57
	Qwen2.5-Omni-7B	61.42	**60.47**	55.30
	Gemini-2.5-flash	59.34	59.87	**59.10**
text-audio	Qwen2-Audio-7B- Instruct	62.5	62.42	58.71
	Phi-4-multimodal-instruct	46.74	49.24	31.77
	Qwen2.5-Omni-7B	83.44	67.28	67.41
	Gemini-2.5-flash	**91.55**	**93.75**	**92.05**
text-video	Qwen2-VL-7B-Instruct	69.70	64.61	64.88
	Qwen2.5-VL-7B-Instruct	76.77	77.39	77.02
	Phi-4-multimodal-instruct	62.59	52.31	34.04
	Qwen2.5-Omni-7B	80.31	65.06	64.62
	Gemini-2.5-flash	**88.11**	**89.19**	**88.54**
audio-video	Phi-4-multimodal-instruct	58.41	51.95	34.47
	Qwen2.5-Omni-7B	**82.15**	68.45	68.97
	Gemini-2.5-flash	79.44	**79.8**	**79.61**
text-audio-video	Phi-4-multimodal-instruct	61.41	59.56	52.98
	Qwen2.5-Omni-7B	84.68	73.53	74.95
	Gemini-2.5-flash	**91.79**	**91.79**	**91.79**

Table 9: Results on MuSaG-FullAgree. For each modality, we report the same modality-specific models (📝, 🔊, 🎥) and multimodal models (🔀).

Effect	Setting	Transcript	True Label	Model Preditions
Improved	Without context	*„Genau so und nur so gewinnt man die Herzen der Menschen."*	`sarc`	Text: `non-sarc` Audio: `non-sarc` Video: `non-sarc` AV: `non-sarc`
	With context	*„[. . .] Wahlprogramm nicht gelesen haben und dass sie das, was sie gerade sagen, auswendig gelernt haben. Genau das Gegenteil. Ja, offensichtlich. Ja, darf ich mal? Klar darfst du.* ***Genau so und nur so gewinnt man die Herzen der Menschen."***	`sarc`	Text: `sarc` Audio: `sarc` Video: `sarc` AV: `sarc`
Degraded	Without context	*„Bevor das Gesetz endgültig beschlossen wird, dürfen jetzt erstmal alle Fraktionen im Bundestag draufschauen und Änderungsvorschläge einbringen."*	`non-sarc`	Text: `non-sarc` Audio: `non-sarc` Video: `non-sarc` AV: `non-sarc`
	With context	*„Damit sich das ändert, will die Bundesregierung die Bonitätseinschätzung von Auskunftteilen nun stärker reglementieren. Das klingt ja erstmal so, als wäre jetzt alles in trockenen Tüchern, Schufa ausgedribbelt, aber weit gefehlt. Denn dieser Entwurf ist ja erst der Anfang.* ***Bevor das Gesetz endgültig beschlossen wird, dürfen jetzt erstmal alle Fraktionen im Bundestag draufschauen und Änderungsvorschläge einbringen."***	`non-sarc`	Text: `sarc` Audio: `non-sarc` Video: `sarc` AV: `sarc`
No impact	Without context	*„Haben Sie das auch gehört? Hat Herr Klingbeil gerade gesagt, deutsche Stärken sind moderne Bildung, bezahlbares Wohnen und digitale Infrastruktur? Er war aber schon in Deutschland, oder? Ja."*	`sarc`	Text: `sarc` Audio: `sarc` Video: `sarc` AV: `sarc`
	With context	*„[. . .] ob Lars Klingbeil das hier wirklich ernst meint. Wir setzen auf die Stärke unseres Landes, auf gute Arbeit, auf moderne Bildung, auf bezahlbares Wohnen, auf digitale Infrastruktur, auf klimafreundliche Energie.* ***Haben Sie das auch gehört? Hat Herr Klingbeil gerade gesagt, deutsche Stärken sind moderne Bildung, bezahlbares Wohnen und digitale Infrastruktur? Er weiß aber schon in Deutschland, oder? Ja."***	`sarc`	Text: `sarc` Audio: `sarc` Video: `sarc` AV: `sarc`

Table 10: Examples illustrating the effect of extended context on classification. In the *With context* rows, the target sentence is shown in **bold**. Model predictions are shown per modality: Text = Qwen3-8B, Audio = Gemini-2.5-flash (audio), Video = Gemini-2.5-flash (video), AV = Gemini-2.5-flash (audio-video).