😈 FanChuan: A Multilingual and Graph-Structured Benchmark For Parody Detection and Analysis

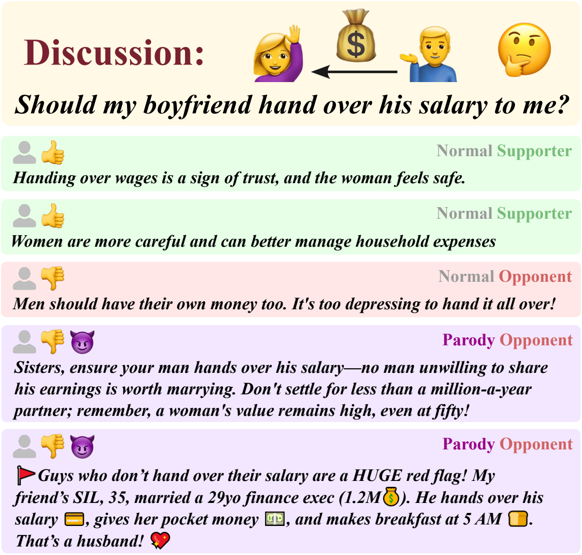
Yilun Zheng[1], **Sha Li**[1], **Fangkun Wu**[1], **Yang Ziyi**[1], **Lin Hongchao**[1], **Zhichao Hu**[1], **Cai Xinjun**[1], **Ziming Wang**[1], **Jinxuan Chen**[1], **Sitao Luan**[2*], **Jiahao Xu**[1*], **Lihui Chen**[1*]
[1]Nanyang Technological University, Centre for Info. Sciences and Systems,
[2]Mila - Quebec Artificial Intelligence Institute.

Abstract

Parody is an emerging phenomenon on social media, where individuals imitate a role or position opposite to their own, often for humor, provocation, or controversy. Detecting and analyzing parody can be challenging and is often reliant on context, yet it plays a crucial role in understanding cultural values, promoting subcultures, and enhancing self-expression. However, the study of parody is hindered by limited available data and deficient diversity in current datasets. To bridge this gap, we built seven parody datasets from both English and Chinese corpora, with 14,755 annotated users and 21,210 annotated comments in total. To provide sufficient context information, we also collect replies and construct user-interaction graphs to provide richer contextual information, which is lacking in existing datasets. With these datasets, we test traditional methods and Large Language Models (LLMs) on three key tasks: (1) parody detection, (2) comment sentiment analysis with parody, and (3) user sentiment analysis with parody. Our extensive experiments reveal that parody-related tasks still remain challenging for all models, and contextual information plays a critical role. Interestingly, we find that, in certain scenarios, traditional sentence embedding methods combined with simple classifiers can outperform advanced LLMs, *e.g.,* DeepSeek-R1 and GPT-o3, highlighting parody as a significant challenge for LLMs. Our code and data is available at `https://github.com/Lisaaa1017/Fanchuan`.

1 Introduction

Parody in social media[1] is a form of humor or satire, which uses exaggerated or absurd imitations for critique or entertainment (Dentith, 2002). It has become popular around some controversial topics in recent years, especially among the young generation (Mulholland, 2013; McClennen and Maisel, 2014). For example, as shown in Figure 1, the question "*Should my boyfriend hand over his salary to me?*" has sparked intense debate (Q&A). While some users clearly express their views as neutral, supportive, or opposed, others adopt a parody tone, mockingly pretending to support the stance with exaggerated statements like, "*Guys who don't hand over their salary are a HUGE red flag...*", which subtly opposes it. This tactic can attract attention and provoke reactions through humor, making people reflect their opinions. Similar to irony or sarcasm (Frenda et al., 2023), parody also expresses the opinion opposite to its appearance. However, it emphasizes playful, entertaining, and exaggerated mimicry of a character, making the underlying critique more accessible and engaging to the audience.



Should my boyfriend hand over his salary to me?

Handing over wages is a sign of trust, and the woman feels safe.

Women are more careful and can better manage household expenses

Men should have their own money too. It's too depressing to hand it all over!

Sisters, ensure your man hands over his salary—no man unwilling to share his earnings is worth marrying. Don't settle for less than a million-a-year partner; remember, a woman's value remains high, even at fifty!

🚩Guys who don't hand over their salary are a HUGE red flag! My friend's SIL, 35, married a 29yo finance exec (1.2M💰). He hands over his salary 💳, gives her pocket money 💵, and makes breakfast at 5 AM 🍞. That's a husband! 💘

Figure 1: People debate online about the topic, "*Should my boyfriend hand over his salary to me?*" Some users explicitly support or oppose this viewpoint, while others implicitly express their stance through parody, using humor or even subtle blackmail to make their point.

The real meaning behind parody is highly

---

[*] Corresponding author: sitao.luan@mail.mcgill.ca, jiahao004@e.ntu.edu.sg, elhchen@ntu.edu.sg.

[1] Also known as "反串" or "FanChuan" in Chinese online social media.

Findings of the Association for Computational Linguistics: ACL 2025, pages 21937–21957
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

culture-dependent. Therefore, the analysis of parody can offer unique insights in understanding the corresponding cultural values. The spread of parody on internet also fosters a diverse linguistic culture (Menghini, 2024). People can share their distinct views on society, political, or cultural topics in a humorous and engaging manner, encouraging global and cross-cultural dialogue. In addition, parody plays a crucial role in the formation of subcultures (Willett, 2009; Booth, 2014). Parody comments not only create distinct communities, but also mirror the values and identities of online users. For younger generations, parody comments have become a way of self-expression, which help to define their uniqueness, build connection with others, and form social circles. Gradually, it has become a shared language and a set of symbols for the growth of internet subcultures.

Despite the widespread popularity of parody, there is a lack of high-quality datasets that capture parody comments with different topics and languages (Maronikolakis et al., 2020), restricting the more general and inclusive analysis in various contexts. To fill this gap, we propose FanChuan, a parody benchmark with high quality in three key aspects: **high diversity, rich contexts, and precise annotations**. **First**, we enhance diversity by collecting data from multiple sources (both Chinese and English corpora), a wide range of topics, and various social media platforms. Such broad coverage allows us to conduct more sufficient, balanced and fair evaluations of models. **Second**, we construct richer context information by building the relationship between comments and their replies as heterogeneous graphs. Unlike previous studies that only focus on textual (Zhang et al., 2022) or dialogue (Bamman and Smith, 2015; Wang et al., 2015) content, the graph-structured context enables the exploitation of relational information, which is found to be fairly valuable later. **Third**, since parody labeling is quite challenging and disagreements among annotators can easily arise, we ensure the quality of annotation by employing native speakers to label the parody and sentiment of each comment. Additionally, we have expert judges to resolve any disagreement and Large Language Models (LLMs) to refine the annotation results, ensuring consistency and reliability. As a result, we have created **seven** datasets, with **14,755** annotated users and **21,210** annotated comments in total, enabling comprehensive experiments and analyses.

With the new datasets, we evaluate embedding-based methods (Liu, 2019), incongruity-based methods (Liu et al., 2023b), outlier detection methods (Liu et al., 2008), graph-based methods (Kipf and Welling, 2016), and Large Language Models (LLMs) (OpenAI, 2023) on FanChuan with three parody related tasks: parody detection, comment sentiment classification with parody, and user sentiment classification with parody. Our results indicate that **(1)** parody-related tasks are challenging for all models, and even LLMs fail to consistently outperform traditional embedding-based approaches; **(2)** model performance of sentiment classification drops significantly on comments exhibiting parody behavior compared to those without parody; **(3)** incorporating commented objects as contextual information greatly enhances parody detection performance; **(4)** reasoning LLMs fail to outperform non-reasoning LLMs on parody detection. To our best knowledge, the existing studies on parody(Maronikolakis et al., 2020; Willett, 2009) are all from pre-LLMs era, and we are the first to evaluate the performance of LLMs on parody detection. In summary, our contributions are summarized as follows:

- We introduce FanChuan, a parody benchmark that includes seven datasets from both Chinese and English corpora, containing 21,210 annotated comments and 14,755 annotated users.
- We leverage heterogeneous graphs to model user interaction relationships, providing richer contextual information compared to previous datasets.
- We comprehensively evaluate five types of methods, including embedding-based methods, inconsistency-based methods, outlier detection methods, graph-based methods, and LLMs, on three parody-related tasks.
- Our findings reveal that parody-related tasks are challenging and LLMs cannot always outperform traditional embedding-based methods. Additionally, we show that reasoning LLMs generally underperform non-reasoning LLMs in parody detection.

2 FanChuan

In this section, we will introduce the details about FanChuan. Specifically, in Section 2.1, we introduce the dataset construction process, including data collection, annotation and preprocessing.

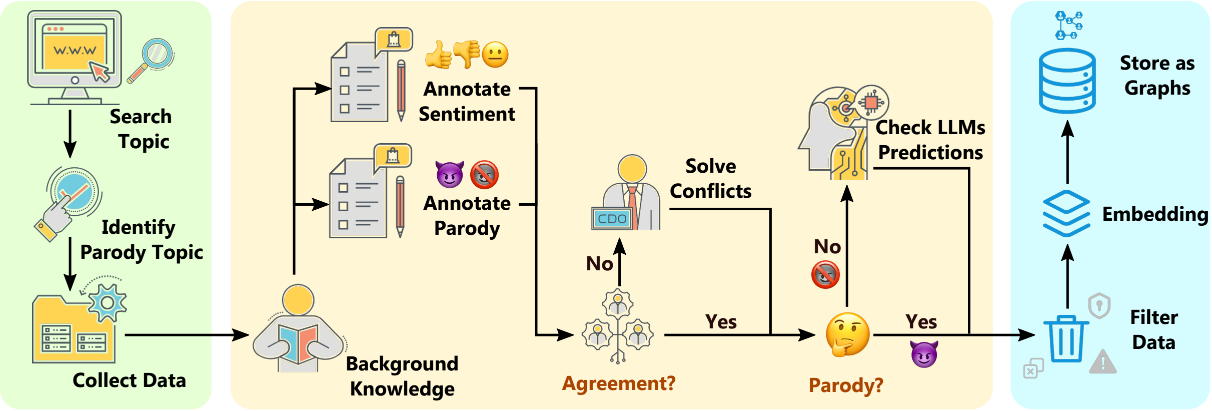


Figure 2: The pipeline for the construction of FanChuan, which includes three key steps: data collection (left), annotation (middle), and preprocessing (right).

These steps ensure high diversity, precise annotations, and rich contexts within our dataset. In Section 2.2, we propose three parody-related tasks for model evaluations.

2.1 Dataset Construction

As illustrated in Figure 2, the data construction process for FanChuan involves three steps: data collection, annotation, and preprocessing. Then we introduce the details of each step as follows.

Data collection To ensure a comprehensive evaluation, we ensure **high diversity** in our benchmark by selecting a wide range of topics from both Chinese and English corpora. Given that parody often emerges around controversial issues, we begin by focusing on topics or recent events that have sparked intense debates on social media. To select the post that includes adequate parody comments, we randomly sample a subset of its comments to determine the proportion of parody content. If more than 3% of the comments are identified as parody, we classify it as suitable for further collection. To capture the most relevant content, we use keyword search to identify prominent posts, then collect their comments, replies, and associated content.

Data Annotation Labeling parody presents a significant challenge, not only because it requires a high familiarity with the content and culture (Bänziger and Scherer, 2005), but also due to potential disagreements of understanding among annotators from diverse backgrounds (Dress et al., 2008). To ensure **precise annotations** in FanChuan, the annotation process includes five steps: **(1)** To provide accurate and culturally relevant insights, we assign native speakers to annotate Chinese and English datasets, respectively. Annotators are then asked to review relevant materials to enhance their understanding before starting the annotation process. **(2) Sentiment Annotation.** Annotators classify the sentiment of a given comment or user by answering the question: *"Does this comment or user support, oppose, or remain neutral regarding to this statement?"* **(3) Parody Annotation.** After sentiment classification, annotators are asked to determine whether a comment is a parody by answering the question: *"Is this comment a parody or not?"* During both sentiment and parody annotation stages, annotators are provided with relevant comments and context to ensure accurate labeling. **(4) Resolving Discrepancies.** Each comment receives a final label based on the majority vote of three annotators. If consensus is not reached, the most knowledgeable annotator on the relevant topic or event reassesses the labels. **(5) Verification.** To minimize errors in parody annotations, an experienced annotator reviews all comments labeled as parody. Note that this annotator will also double-check the comments that are labeled as parody by LLMs but not labeled by human annotators.

Data preprocessing To ensure data quality, we first delete any content or comments that contain irrelevant, sensitive, personal, or hazardous information. We provide three types of embeddings: Bag of Words (BoW) (Mikolov, 2013), Skip-gram (Mikolov et al., 2013), and RoBERTa (Liu, 2019). Given that the context of parody forms a network structure, we store the data as heterogeneous graphs as shown in Figure 3, where the nodes represent users and posts, and there are two types of edges to represent two types of relations: user-comments-post, and user-comments-user. Compared with existing datasets (Bamman and Smith, 2015; Ptáek et al., 2014) that focus solely on content or dialogue, such graph-structured data enables deeper understanding of parody with **richer contexts**, including 2-hop neighbors and higher-order relationships.

Finally, as shown in Table 1, we constructed seven datasets from both Chinese and English corpora, encompassing multiple topics, with a total of 14,755 annotated users and 21,210 annotated comments. Our analysis reveals that parody comments constitute only a small proportion of the total comments across all datasets. For detailed description and background information of each dataset, please refer to Appendix A.

2.2 Problem Definition

As shown in Figure 3, we utilize Heterogeneous Information Networks (HINs) to structure our datasets, representing the relational information in content and comments. Each HIN comprises two types of nodes: user nodes and post nodes, along with two types of edges: user comments to posts and user comments to users[2]. Each edge is directed, with the source being the user and the target either a post or another user. As shown by the orange edges on the right in Figure 3, multiple edges may exist between two nodes due to several rounds of replies among these users. This results in a directed multigraph (Gross and Yellen, 2003). Each edge or node is associated with text as features. We then introduce three tasks as follows.

P1. Parody Detection Parody detection aims to identify whether a comment is parody or normal. In HINs, this can be framed as a binary classification task on edges. Given that parody comments represent a small fraction of all comments, this task can also be considered as outlier detection.

P2. Comment Sentiment Classification Like parody detection, comment sentiment classification aims to categorize comments into three sentiment labels: positive, negative, and neutral.

P3. User Sentiment Classification This task focuses on classifying users' sentiment as either a supporter, opponent, or neutral. Unlike the edge classification tasks discussed earlier, this is a node classification task in HINs.

Remarks We introduce sentiment classification tasks due to the complexity of the scenarios that include parody comments (Bull, 2010). In the context of parody, these tasks serve as a comprehensive measure to assess the effectiveness of current models in handling parody-related tasks, which will be introduced in the next section.

3 Experiments

3.1 Settings

We split all the comment data into training, validation, and test sets with a ratio of 40%/30%/30%. We consider parody detection as a binary classification problem use F1 score for the evaluation. We model the comment and user sentiment classification with parody as multi-class classification problems, and use Macro-F1 to measure the model performance. For comprehensive evaluation and analysis, we test five types of approach in our experiments:

(1) Embedding-based methods. This category includes Bag-of-Words (BoW) (Mikolov, 2013), Skip-gram (Mikolov et al., 2013), and RoBERTa (Liu, 2019), all of which utilize Multi-Layer Perceptron (MLP) classifiers. These methods are widely used and can provide general text representations to capture linguistic patterns and semantics.

(2) Inconsistency-based methods. These methods are commonly used for irony detection and we assess BNS-Net (Zhou et al., 2024), DC-Net (Liu et al., 2021), QUIET (Liu et al., 2023a), and SarcPrompt (Liu et al., 2023b). Similar to irony or sarcasm, parody usually contains inconsistencies between literal and intended meaning, and thus, the evaluation of these methods are necessary.

(3) Outlier detection methods. This category includes Isolation Forest (Liu et al., 2008), the Z-Score Method (Rousseeuw and Hubert, 2011), and One-Class SVM (Li et al., 2003). Similar to outlier detection tasks, where data is highly imbalanced, parody only accounts for around 5%-10% of all comments and tremendously deviates from the normal comment patterns, which makes outlier detection methods quite relevant.

(4) Graph-based methods. Since (graph-structured) context information is highly important for parody understanding, and to capture complex structural information in user interaction graphs, Graph Neural Networks (GNNs) could be used for user sentiment classification. Three types of classical GNNs are used: Graph Convolutional Net-

---

[2] A comment on another comment inherently forms an edge linking to another edge, which cannot be directly represented in a graph. Instead, we connect such comments to the target user, as they reflect that user's traits or viewpoints.

Dataset	Topic	Language	Comment			User	
			#Num	#Parody/#Normal	#Pos / #Neg / #Neu	#Num	#Pos / #Neg / #Neu
Alibaba-Math	Education	Chinese	8353	489 / 7864	1831 / 1509 / 5013	5247	1397 / 1044 / 2806
BridePrice	Social	Chinese	1774	84 / 1690	20 / 385 / 1369	1254	17 / 341 / 896
DrinkWater	Technology	Chinese	3659	113 / 3546	378 / 384 / 2897	3204	349 / 353 / 2502
CS2	Game	Chinese	3196	196 / 3000	169/480/517/25/2005*	2093	117/372/385/19/1200*
CampusLife	Life	English	1206	89 / 1117	41 / 201 / 964	569	30 / 131 / 408
Tiktok-Trump	Politics	English	1634	97 / 1537	150 / 495 / 989	1237	127 / 434 / 676
Reddit-Trump	Politics	English	1388	171 / 1217	169 / 678 / 541	1151	149 / 594 / 408

Table 1: Dataset Statistics. *In particular, for CS2, there are five types of sentiment labels: support for G2 (a gaming club), support for NAVI (another gaming club), opposition to G2, opposition to NAVI, and neutral.

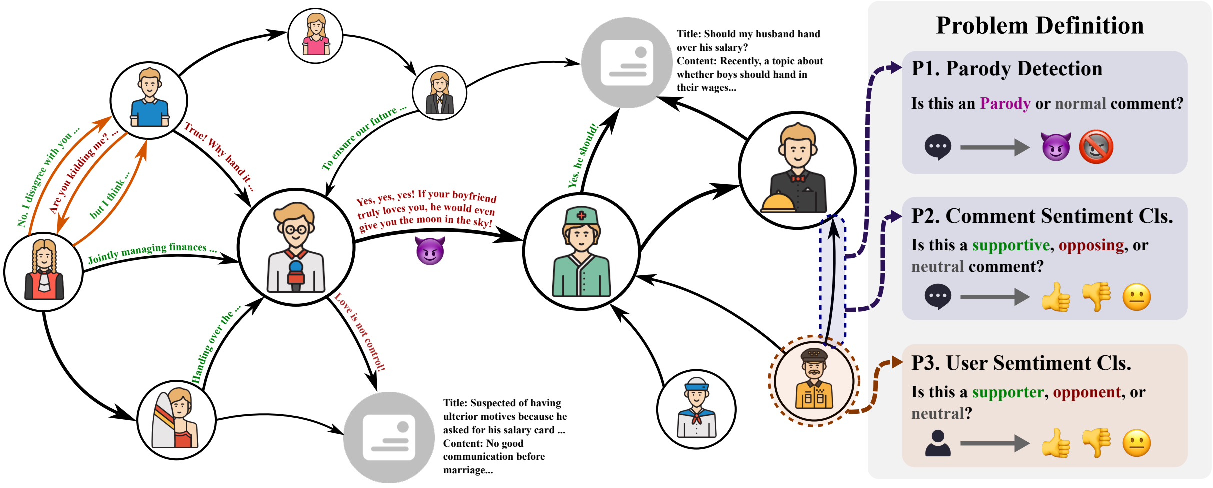

No. I disagree with you ...
Are you kidding me? ...
but I think ...
True! Why hand it ...
Jointly managing finances ...
To ensure our future ...
Handing over the ...
Love is not control!
Yes, yes, yes! If your boyfriend truly loves you, he would even give you the moon in the sky!
Yes, he should!
Title: Should my husband hand over his salary?
Content: Recently, a topic about whether boys should hand in their wages...
Title: Suspected of having ulterior motives because he asked for his salary card ...
Content: No good communication before marriage...
Problem Definition
P1. Parody Detection
Is this an Parody or normal comment?
P2. Comment Sentiment Cls.
Is this a supportive, opposing, or neutral comment?
P3. User Semtiment Cls.
Is this a supporter, opponent, or neutral?

Figure 3: Examples of a parody dataset as a heterogeneous graph.

works (GCN) (Kipf and Welling, 2016), Graph Attention Networks (GAT) (Veličković et al., 2017), and GraphSAGE (Hamilton et al., 2017).

(5) Large Language Models (LLMs). We evaluate models such as ChatGPT-4o (and 4o-mini) (OpenAI, 2023) from OpenAI, Claude 3.5 (Anthropic, 2025) from Anthropic, Qwen 2.5-72B (Yang et al., 2024) from Alibaba, and DeepSeek-V3 (DeepSeek-AI et al., 2024) from DeepSeek under a zero-shot setting by API calling. In addition, we evaluate open-source LLMs including LLaMA-3-8B (Grattafiori et al., 2024), Qwen2.5-7B (Yang et al., 2024), Mistral-7B (Jiang et al., 2023), Gemma-2-9B (Team et al., 2024), and DeepSeek-R1-7B (DeepSeek-AI et al., 2025) using supervised finetuning.

3.2 Performance Comparison

The evaluation results on the three parody-related tasks are shown in Table 2, 3, 4. The best and runner-up methods for each dataset are highlighted in **bold** and underlined, respectively. Then, the detailed comparison and analysis are as follows.

Parody Detection. The results in Table 2 indicate that: (1) Parody detection is challenging for all models, with most achieving only $10\% \sim 50\%$ F1 scores. Even the best-performing methods for *Alibaba.* and *Drink.* reach only 20.15% and 17.39%, respectively, highlighting the difficulty of the task. (2) LLMs with zero-shot generally rank higher but struggle with Chinese datasets. Specifically, both of ChatGPT-4o and Deepseek-V3 achieve 3.86 average rank across all datasets, outperforming other methods. However, traditional methods perform better on Chinese datasets. For instance, SarcPrompt achieves an F1 score of 22.22% on *Bride.* and 21.39% on *CS2*, outperforming the best LLM by a large margin. (3) Under supervised fine-tuning, Gemma-2-9B generally achieves the highest average ranking across all methods and delivers the best performance on 5 out of 7 datasets. In contrast, other fine-tuned LLMs perform worse than traditional approaches such as BNS-Net and BoBERTa+MLP. In addition to the performance comparison, we conduct a case study to further investigate how well LLMs understand parody detection in Appendix B.

Sentiment Classification. Tables 3 and 4 present the model performance in comment and user sentiment classification, respectively. Our findings are

Paradigm	Method	Alibaba.	Bride.	Drink.	CS2	Campus.	Tiktok.	Reddit.	Ave. Rank
Embedding -based	BoW+MLP	10.17	15.83	9.06	15.93	11.20	13.71	16.91	12.43
	Skip-gram+MLP	14.16	17.50	14.55	17.29	10.40	15.43	14.85	10.29
	RoBERTa+MLP	14.30	19.17	13.33	16.61	16.52	12.00	23.09	8.86
Inconsistency -based	BNS-Net	13.62	12.31	16.67	20.00	28.17	24.86	16.67	8.43
	DC-Net	13.54	10.53	**17.39**	14.37	14.04	9.38	24.16	11.71
	QUIET	15.98	10.75	4.94	7.75	13.07	10.11	16.34	14.29
	SarcPrompt	14.20	22.22	5.26	21.39	26.67	15.38	15.09	9.57
Outlier Detection	Isolation Forest	5.93	1.18	0.90	7.14	5.62	6.15	11.70	19.57
	RoBERTa+Z-Score	13.06	20.83	12.31	18.64	17.78	14.29	22.68	9.14
	One-Class SVM	5.81	4.71	1.79	5.64	7.82	9.14	14.99	19.00
LLMs Zero-shot	ChatGPT4o	15.90	13.54	8.94	18.86	34.29	39.51	37.26	4.86
	ChatGPT4o-mini	13.73	11.06	8.91	16.00	40.00	36.41	36.90	7.14
	Claude3.5	13.21	12.49	8.56	16.00	**41.24**	29.96	36.45	7.86
	Qwen2.5-72B	14.88	12.44	7.81	19.38	28.89	27.70	33.29	7.57
	DeepSeek-V3	16.17	13.24	9.19	20.45	32.55	31.10	34.34	5.00
LLMs-SFT	Llama3-8B	9.57	26.67	8.86	13.00	3.57	13.50	18.31	13.29
	Qwen2.5-7B	10.98	10.56	5.88	15.21	30.40	18.36	29.69	11.43
	Mistral-7B	10.87	10.59	5.75	11.59	20.00	16.74	19.52	13.43
	Gemma-2-9B	**20.15**	**55.56**	16.77	**29.09**	33.33	**48.78**	**47.37**	**1.57**
	DeepSeek-R1-7B	10.77	8.24	6.13	12.74	9.33	22.47	17.04	14.57

Table 2: Comparison of model performance in parody detection using F1 score (%).

as follows: (1) Sentiment classification in the context of parody presents significant challenges. The top-performing models across each dataset achieve F1 scores ranging from 40% to 50%, which are notably lower than the performance on traditional sentiment classification benchmarks without parody(Socher et al., 2013; Maas et al., 2011). (2) Although LLMs show their superiority over other methods in terms of average rank, they still underperform some traditional approaches on certain datasets. For example, although ChatGPT-4o-mini attains the highest average rank of 4.29 in comment sentiment classification, it performs much worse than BoW+MLP on *Bride.* and DC-Net on *Campus.* (3) Graph-based methods demonstrate strong performance on certain datasets. For example, GCN achieves the best results on *Bride.*, suggesting that the relational context information in user-interaction networks is informative and beneficial for some tasks in sentiment classification.

In general, all the parody-related tasks are challenging for current models and no model can take dominant advantage over others cross all datasets. These observations underscore the need for further study and model development on parody-related tasks.

3.3 Influence of Context on Parody Detection

Since parody detection requires a deep understanding of the background information of a topic, intuitively, the context information should have a strong impact on model performance. Therefore, we introduce relevant background details and target comments (when available), and conduct ablation study to investigate its impact on model performance. In Table 5, we report the average F1 score across seven datasets, both with and without context. Performance improvements and declines are highlighted in green and red, respectively.

Overall, most models benefit from contextual information, with ChatGPT-4o improving significantly from 24.04 to 28.53 and RoBERTa+MLP increasing from 16.43 to 21.23. Our results are consistent with the observations in (Bamman and Smith, 2015; Wang et al., 2015) that context improves model performance on sarcasm and irony detection. However, Qwen2.5 is the only model that performs worse with added context, suggesting potential limitations in how it processes additional information. These results highlight that while context generally enhances parody detection, its effectiveness varies across models. Please refer to Appendix D.1 for more details of the impact of context on each dataset.

Paradigm	Method	Alibaba.	Bride.	Drink.	CS2	Campus.	Tiktok.	Reddit.	Ave. Rank
Embedding -based	BoW+MLP	35.30	**40.43**	48.78	27.56	32.35	33.74	37.13	8.14
	Skip-gram+MLP	39.62	39.50	47.46	31.09	30.80	35.42	37.71	6.29
	RoBERTa+MLP	36.91	34.48	44.17	26.02	38.87	47.56	51.66	8.00
Inconsistency -based	BNS-Net	35.48	29.40	45.66	21.13	29.71	26.47	22.08	7.29
	DC-Net	16.07	28.87	48.66	18.89	**38.90**	45.21	37.18	7.29
	QUIET	24.34	30.26	35.52	17.65	30.05	29.51	23.95	7.00
	SarcPrompt	28.77	28.85	33.91	19.18	35.21	40.06	22.69	5.43
LLMs	ChatGPT4o	40.00	32.28	47.75	**37.82**	32.10	51.02	51.89	4.86
	ChatGPT4o-mini	40.01	34.27	**49.95**	34.33	33.19	51.56	52.42	**4.29**
	Claude3.5	**40.53**	29.89	42.99	30.70	28.31	46.03	51.92	5.71
	Qwen2.5	38.46	31.83	46.14	34.78	28.38	47.55	51.93	6.86
	DeepSeek-V3	35.88	28.15	43.05	32.62	36.36	**56.26**	**54.83**	6.86

Table 3: Comparison of model performance in comment sentiment classification with parody using Macro-F1 score (%)

Paradigm	Method	Alibaba.	Bride.	Drink.	CS2	Campus.	Tiktok.	Reddit.	Ave. Rank
Embedding -based	BoW+MLP	46.54	37.60	46.65	29.22	32.35	35.05	31.97	7.57
	Skip-gram+MLP	**46.99**	38.28	50.45	31.92	32.02	38.46	32.69	6.42
	RoBERTa+MLP	43.11	36.94	44.20	27.09	35.49	50.82	52.79	5.00
Inconsistency -based	BNS-Net	34.32	27.21	41.91	23.38	28.67	23.61	22.98	13.00
	DC-Net	16.51	33.56	48.65	17.17	35.60	34.62	39.54	9.57
	SarcPrompt	27.72	38.54	29.51	15.62	31.45	24.48	39.10	11.29
Graph -based	GCN	37.69	**40.00**	43.67	23.64	36.45	42.94	48.06	7.00
	GAT	38.30	38.53	43.44	23.72	37.20	42.12	50.57	6.71
	GraphSAGE	39.92	37.63	42.79	25.98	32.94	40.66	52.08	7.71
LLMs	ChatGPT-4o	41.71	35.02	**51.54**	**39.19**	35.89	45.87	49.01	**4.14**
	ChatGPT-4o-mini	40.55	30.25	45.88	34.03	31.95	45.29	51.20	6.71
	Claude3.5	41.47	29.96	43.78	32.81	31.07	41.85	46.92	8.57
	Qwen2.5	40.89	33.08	49.52	36.51	33.34	46.18	50.13	5.29
	DeepSeek-V3	40.00	26.37	41.55	33.61	**40.49**	**54.04**	**53.22**	6.00

Table 4: Comparison of model performance in user sentiment classification with parody using Macro-F1 score (%).

Method	w/o Context	w. Context	Δ
BoW+MLP	13.26	15.19	+1.93
Skip-gram+MLP	14.88	16.19	+2.08
RoBERTa+MLP	16.43	21.23	+4.80
ChatGPT4o	24.04	28.53	+4.49
ChatGPT4o-mini	23.29	23.99	+0.70
Claude3.5	22.56	23.09	+0.53
Qwen2.5	20.63	18.04	-2.59
DeepSeek-V3	22.43	24.83	+2.40

Table 5: Impact of context on parody detection using F1 Score (%) averaged over seven datasets.

Method	Non-Parody	Parody	Δ
BoW+MLP	35.71	30.21	-5.50
Skip-gram+MLP	37.17	30.08	-7.09
RoBERTa+MLP	39.65	33.15	-6.50
ChatGPT4o	42.28	26.84	-15.44
ChatGPT4o-mini	42.68	27.03	-15.65
Claude3.5	38.98	24.87	-14.11
Qwen2.5	40.15	26.87	-13.28
DeepSeek-V3	41.17	29.29	-11.89

Table 6: Impact of parody on comment sentiment classification using Macro F1 Score (%) averaged over seven datasets.

3.4 Influence of Parody to Sentiment Classification

To confirm that parody adds challenges to sentiment classification, we evaluate model performance using Macro F1 score averaged over seven datasets on comment sentiment classification, and compare the results of parody and non-parody comments. As shown in Table 6, the average Macro F1 scores decrease by 5% to 15% across all models, indicating that parody significantly increases the difficulty of sentiment classification. Additionally, we observe that while LLMs outperform embedding-based methods on non-parody comments, their performance deteriorates on parody comments, falling a lot behind embedding-based methods. We speculate that this degradation occurs because these topics are relatively new and LLMs have not encountered such data before, whereas the training

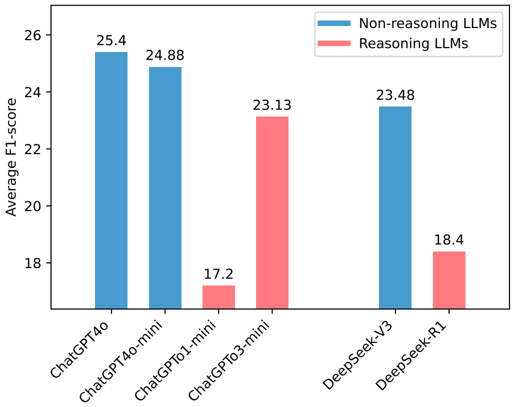


Figure 4: Performance comparison between reasoning LLMs and non-reasoning LLMs using average F1 Score (%) over six datasets.

process in embedding-based methods allows them to better adapt to the updated knowledge. For more details of the impact of context on each dataset, please refer to Appendix D.2.

3.5 The More LLM Reasons, The Worse It Gets

Recently, there has been a surge in reasoning LLMs (OpenAI, 2024), which enhance performance by introducing inference-time scaling in the Chain-of-Thought (CoT) (Wei et al., 2022) reasoning process. To assess the impact of reasoning on LLM performance in parody detection, we compared the performance of reasoning LLMs with that of non-reasoning LLMs. Figure 4 presents the average F1 scores of reasoning LLMs, including ChatGPTo1-mini (OpenAI, 2024), ChatGPTo3-mini (OpenAI, 2025), and DeepSeek-R1 (DeepSeek-AI et al., 2025), and non-reasoning LLMs, including ChatGPT4o, ChatGPT4o-mini, and DeepSeek-V3. Surprisingly, unlike math, coding (Wang et al.) and medical applications (Xu et al., 2024), where reasoning LLMs significantly improve performance, our results show that reasoning LLMs underperform their non-reasoning counterparts. This finding aligns with the conclusion in (Yao et al., 2024), which suggests that tasks like sarcasm detection do not follow a step-by-step reasoning process. This can explain why CoT does not enhance LLM performance. It indicates that the complexities of parody detection may require alternative strategies beyond reasoning, highlighting the need for further research in this area. Please see Appendix D.3 for detailed results on the performance of reasoning LLMs in parody detection.

4 Related Work

In this section, we introduce the datasets and detection methods related to parody, as well as its associated topics: sarcasm, irony, and humor.

4.1 Dataset

The datasets for parody and sarcasm cover a diverse array of topics, including politics (Gong et al., 2020), gender (Frenda et al., 2023), and education (Barbieri et al., 2014). They utilize various modalities, such as text (Zhang et al., 2022), speech (Ariga et al., 2024), visual (Schifanella et al., 2016a), and multimodal formats (Bedi et al., 2021; Maity et al., 2022). Beyond the content itself, context plays a crucial role in understanding sarcasm or parody (Wallace et al., 2014). To enhance contextual information, Wang et al. (2015); Bamman and Smith (2015) collect data from dialogues. For annotation, Bamman and Smith (2015); Ptáek et al. (2014) use user-provided tags as labels, while Riloff et al. (2013) employ manual annotation. As noted by Chen et al. (2024), the former method requires no human involvement but can lead to noise, as not all users utilize tags. In contrast, the latter approach can yield more generalized labels but may result in significant disagreement among annotators (Joshi et al., 2016). In conclusion, most datasets focus on sarcasm detection (Gong et al., 2020; Zhang et al., 2022; Maity et al., 2022), leaving a notable scarcity of parody datasets.

4.2 Irony or Sarcasm Detection

Deep learning approaches for detecting parody and sarcasm can be categorized into incongruity-based, sentiment-based, and knowledge-based perspectives (Chen et al., 2024). Incongruity-based methods focus on the inherent incongruity that characterizes sarcastic content (Riloff et al., 2013). For example, Hazarika et al. (2018) and Schifanella et al. (2016b) identify sarcasm by measuring inconsistencies between different targets or modalities. Sentiment-based methods operate on the assumption that there are dependencies between sentiments and sarcasm. Savini and Caragea (2020) propose integrating sentiment tasks into the training process alongside sarcasm detection to enhance model performance. To create emotion-rich representations, Babanejad et al. (2020) incorporate affective and contextual cues. Recognizing that understanding sarcasm can often be implicit, knowledge-based approaches (Chen et al., 2022;

Li et al., 2021) leverage external knowledge bases. These methods typically involve knowledge extraction, selection, and integration (Chen et al., 2024).

5 Conclusions

In this paper, we introduce FanChuan, a multilingual benchmark for parody detection and analysis, encompassing seven datasets characterized by high diversity, rich contextual information, and precise annotations. Our findings reveal that parody detection remains highly challenging for both LLMs and traditional methods, with particularly poor performance on Chinese datasets. We also observe that contextual information significantly enhances model performance, while parody itself increases the difficulty of sentiment classification. Additionally, our results indicate that reasoning fails to improve LLM performance in parody detection. By filling a critical gap in the study of emerging online phenomena, FanChuan provides valuable insights into cultural values and the role of parody in digital discourse. These findings highlight the limitations of current LLMs, presenting an opportunity for future research to enhance model capabilities in parody detection and analysis.

Limitations

While this paper proposes a multilingual parody benchmark and provides an extensive analysis, we acknowledge several limitations that warrant further exploration in future work:

- Limited dataset diversity. Although we collect datasets and analyze experimental results in both Chinese and English, the understanding of how parody manifests or how effective current methods are for parody detection in other languages remains unclear. Therefore, further efforts could be made to gather datasets in additional languages to enhance the diversity of parody data.

- Annotation quality limitations. While we invite multiple annotators and conduct rechecks after labeling, some minor errors may still exist, as annotating parody can be a challenging task. To improve annotation quality in future studies, we will recruit more annotators and provide them with additional background knowledge related to the events before the annotation process. This will help ensure more accurate and consistent annotations.

- Limited evaluation of Large Language Models (LLMs). In this study, we only test the performance of LLMs on parody-related tasks through prompt-based methods, without fine-tuning. This approach may not fully capture the potential of LLMs. Additionally, only 6 LLMs were evaluated, which is a relatively small number considering the rapid development of these models. Future work should include a broader range of LLMs and explore fine-tuning approaches to better assess their capabilities in parody detection tasks.

- Limited exploration of graph-based methods. In our experiments, Graph Neural Networks (GNNs) are used solely for user sentiment classification. The application of GNNs to parody detection and comment sentiment classification remains unexplored, primarily due to the lack of paradigms that allow GNNs to classify edges in graphs. Future work could focus on designing GNN models tailored to edge classification, enabling more comprehensive experiments on parody detection and comment sentiment analysis.

Ethics Statement

Our proposed benchmark, FanChuan, adheres to the ACL Code of Ethics. All the coauthors also work as annotators, and are compensated at an average hourly rate of 20 SGD. The data we collected is licensed under CC BY 4.0 and is used exclusively for academic purposes. It consists of publicly available website comments and does not contain any sensitive or personal information. To protect user privacy, we filtered out any private data during the data collection and organization process, ensuring that the dataset does not include any user-sensitive content. Additionally, recognizing the potential presence of malicious content in user debates, we have removed harmful comments that violate community ethical standards. Regarding the cultural and topical elements in the datasets, our research remains neutral and free from bias, solely focused on academic exploration. Lastly, AI was used to revise the grammar during the paper writing process.

References

Anthropic. 2025. Introducing claude 3.5 sonnet.

Tomoki Ariga, Yosuke Higuchi, Kazutoshi Hayasaka, Naoki Okamoto, and Tetsuji Ogawa. 2024. Parody detection using source-target attention with teacher-forced lyrics. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1151–1155. IEEE.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th international conference on computational linguistics*, pages 225–243.

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter. In *proceedings of the international AAAI conference on web and social media*, volume 9, pages 574–577.

Tanja Bänziger and Klaus R Scherer. 2005. The role of intonation in emotional expressions. *Speech communication*, 46(3-4):252–267.

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in twitter, a novel approach. In *proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 50–58.

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 14(2):1363–1375.

Paul Booth. 2014. Slash and porn: Media subversion, hyper-articulation, and parody. *Continuum*, 28(3):396–409.

Sarah Bull. 2010. Automatic parody detection in sentiment analysis.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Wangqun Chen, Fuqiang Lin, Guowei Li, and Bo Liu. 2024. A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. *Neurocomputing*, 578:127428.

Wangqun Chen, Fuqiang Lin, Guowei Li, Xuan Zhang, and Bo Liu. 2022. Commonsense-aware sarcasm detection with heterogeneous graph attention network. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2181–2188. IEEE.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and Runxin Xu et. al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, and Bingxuan Wang et. al. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Simon Dentith. 2002. *Parody*. Routledge.

Megan L Dress, Roger J Kreuz, Kristen E Link, and Gina M Caucci. 2008. Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27(1):71–85.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Sánchez-Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, et al. 2023. Epic: multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857.

Xiaochang Gong, Qin Zhao, Jun Zhang, Ruibin Mao, and Ruifeng Xu. 2020. The design and construction of a chinese sarcasm dataset. In *Proceedings of the twelfth language resources and evaluation conference*, pages 5034–5039.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Aiesha Letman et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jonathan L Gross and Jay Yellen. 2003. *Handbook of graph theory*. CRC press.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Jiangnan Li, Hongliang Pan, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Sarcasm detection with commonsense knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3192–3201.

Kun-Lun Li, Hou-Kuan Huang, Sheng-Feng Tian, and Wei Xu. 2003. Improving one-class svm for anomaly detection. In *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, volume 5, pages 3077–3081. IEEE.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Yaochen Liu, Yazhou Zhang, and Dawei Song. 2023a. A quantum probability driven framework for joint multi-modal sarcasm, sentiment and emotion analysis. *IEEE Transactions on Affective Computing*, 15(1):326–341.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Yiyi Liu, Yequan Wang, Aixin Sun, Xuying Meng, Jing Li, and Jiafeng Guo. 2021. A dual-channel framework for sarcasm recognition by detecting sentiment conflict. *arXiv preprint arXiv:2109.03587*.

Yiyi Liu, Ruqing Zhang, Yixing Fan, Jiafeng Guo, and Xueqi Cheng. 2023b. Prompt tuning with contradictory intentions for sarcasm recognition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 328–339.

Yuankai Luo, Lei Shi, and Xiao-Ming Wu. 2024. Classic gnns are strong baselines: Reassessing gnns for node classification. *arXiv preprint arXiv:2406.08993*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1739–1749.

Antonis Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020. Analyzing political parody in social media. *arXiv preprint arXiv:2004.13878*.

Sophia A McClennen and Remy M Maisel. 2014. Mesmerized millennials and byte-ing satire: Or how today's young generation thinks. In *Is Satire Saving Our Nation? Mockery and American Politics*, pages 129–153. Springer.

Jacopo Menghini. 2024. The digital evolution of parody. the purpose of online parody in copyright law and freedom of expression.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Monique Mulholland. 2013. Lol: Porn as parody. In *Young People and Pornography: Negotiating Pornification*, pages 99–120. Springer.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2024. Openai o1 system card.

OpenAI. 2025. Openai o3-mini system card.

Tomáš Ptáek, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *COLING 2014, the 25th International Conference on Computational Linguistics*, pages 213–223.

Informal Q&A. Should men hand in their salary cards?

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Peter J Rousseeuw and Mia Hubert. 2011. Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):73–79.

Edoardo Savini and Cornelia Caragea. 2020. A multitask learning approach to sarcasm detection (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13907–13908.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016a. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016b. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, and Morgane Rivière et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516.

Guoyin Wang, Shengyu Zhang, Tianyu Zhan, Zhouzhou Shen, Jiwei Li, Xueyu Hu, Xiaofei Sun, Fei Wu, Gelei Deng, Jie Zhang, et al. Unlocking the mysteries of openai o1: A survey of the reasoning abilities of large language models.

Zelin Wang, Zhijian Wu, Ruimin Wang, and Yafeng Ren. 2015. Twitter sarcasm detection exploiting a context-based model. In *international conference on web information systems engineering*, pages 77–91. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Rebekah Willett. 2009. Parodic practices: Amateur spoofs on video-sharing sites. In *Video cultures: Media technology and everyday creativity*, pages 115–132. Springer.

Shaochen Xu, Yifan Zhou, Zhengliang Liu, Zihao Wu, Tianyang Zhong, Huaqin Zhao, Yiwei Li, Hanqi Jiang, Yi Pan, Junhao Chen, et al. 2024. Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios. *arXiv preprint arXiv:2411.14461*.

An Yang, Baosong Yang, and Beichen Zhang et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Ben Yao, Yazhou Zhang, Qiuchi Li, and Jing Qin. 2024. Is sarcasm detection a step-by-step reasoning process in large language models? *arXiv preprint arXiv:2407.12725*.

Lei Zhang, Xiaoming Zhao, Xueqiang Song, Yuwei Fang, Dong Li, and Haizhou Wang. 2022. A novel chinese sarcasm detection model based on retrospective reader. In *International Conference on Multimedia Modeling*, pages 267–278. Springer.

Liming Zhou, Xiaowei Xu, and Xiaodong Wang. 2024. Bns-net: A dual-channel sarcasm detection method considering behavior-level and sentence-level conflicts. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

A Dataset Details

Alibaba-Math A student from a vocational school achieved remarkable results in the Alibaba Mathematics Competition, despite coming from a school with a less prestigious reputation. Many people supported her, seeing her as a symbol of rising from humble beginnings and a testament to female empowerment. However, some other people questioned her achievements, suggesting that she might have cheated based on snippets from TV interviews. This topic sparked heated discussions on the Chinese internet. To persuade others to believe their claims, some skeptics impersonated her supporters and used exaggerated praise, saying things like, “这位同学有实力！阿里巴巴有眼光！请阿里巴巴破格录取进入达摩院，助力阿里科技快速发展 ” “*(This student has strength! Alibaba has vision! Please grant her an exceptional admission to DAMO Academy to boost Alibaba’s technological growth)*” This is a highly complex topic that encompasses mathematics, education, and gender-related controversies. Annotators working with this dataset must not only be familiar with relevant internet memes but also possess a solid understanding of advanced mathematical concepts.

BridePrice In some parts of China, there is a tradition of giving a bride price to the bride’s family upon marriage. Regarding the demands for exorbitant bride prices, some people believe that the bride price serves as a form of security for the bride, providing her with a greater sense of safety in the marriage. Others argue that the bride price has no inherent relation to marital happiness. This has sparked extensive online debates, and to create an absurd and humorous effect, some opponents of the bride price impersonate the supporters and post comments such as: "是的是的，姐妹们千万别乱嫁人，找不到年入百万的千万别嫁，女孩子五十岁都很值钱！" *(Ladies, never marry recklessly. If he doesn't make a million a year, don't marry him. Girls are valuable even at fifty!)* Gender issues, particularly the topic of bride price, have been a widely debated subject on the Chinese internet for a long time. This dataset requires annotators to be well-versed in these discussions and familiar with the associated memes.

DrinkWater A technology video creator recently posted a video titled "*I Made This to Get Everyone to Drink More Water...*" sparked controversy. In the video, he introduced a complex "*Water Drinking Battle*" system designed to encourage hydration through a reward mechanism. Yet, due to the high design cost and limited effectiveness, some viewers questioned its practicality. Some even ironically pretended to support it, leaving comments like "震古烁今，足以开启第五次技术革命" "*(A groundbreaking innovation capable of launching the fifth technological revolution)*", to express their dissatisfaction. This video creator has always been a subject of controversy. While he is well known for his content on science and technology, some critics argue that he lacks fundamental engineering literacy. Annotators working with this dataset should have a basic understanding of scientific and technological concepts.

CS2 In the Counter Strike 2 (CS2) World Championship finals, G2's newly revamped roster showed impressive strength but once again fell to NAVI, who had already defeated them seven times in a row. This loss sparked heated discussions: someone believes that G2 needs more time to build synergy and has promising potential, while others question whether the roster change truly enhances their chances to win, as they still struggle to overcome their "mental block" against NAVI. Some satirical critics even made eye-catching remarks, such as "传奇捕虾人终结了G2的三日王朝" "*(The legendary shrimp catcher ended G2's three-day dynasty)*", to express doubts about the effectiveness of G2's roster adjustments. Parody comments in this dataset are particularly difficult to identify for those unfamiliar with the background of CS2, as the comments contain terminology of CS2 game and various aliases of teams and players. Annotators must have a strong understanding of these references to accurately interpret the content.

CampusLife This dataset was collected from a university forum, covering various discussion topics such as dorm life, campus buses, job hunting, and administration. One particular post sparked a heated debate: a student complained about their roommate bringing their girlfriend to stay overnight in the dorm and sought advice on how to address the situation. The comment section included parodic remarks like "*Jealous?*", mocking the situation in a humorous yet disapproving tone. Additionally, during the university's open campus day, a poster appeared in a restroom with the title: "*Applying to our university? Your tuition funds Palestinian genocide.*" In response, some users posted parodic comments, such as: "*Every computer on campus is equipped with an Intel processor, and Intel's R&D center is in Israel! If you want to avoid supporting genocide, switch to a computer with a Zhaoxin CPU immediately!*"

Tiktok-Trump In a debate titled "*Can One Awakened Youth Withstand 20 Trump Supporters?*", a female Trump supporter lost the debate due to her illogical reasoning and subsequently faced criticism from many netizens who deemed her remarks meaningless. Among the critics, some parodically commented, "*She did a great job bring up solid points*", to criticize the Trump supporter's lack of logical reasoning ability.

Reddit-Trump Trump is a highly controversial figure due to his political stance, ideology, and behavior, sparking widespread debate with both supporters and critics. Some opponents use parody to mimic his tone, such as commenting, "*He's been tested—more than anyone, by the best doctors in the world. They were amazed, and said they'd never seen scores that high. He'll take another if asked, but they said he doesn't need to. It's incredible*", mocking his rhetorical style and contentious image.

B Case Study on LLMs

To investigate how well LLMs understand parody, we conduct a case study in which LLMs are asked to provide explanations during prediction. Specifically, we construct the prompt by presenting a comment and its associated topic, then ask the LLMs to determine whether the comment is a parody and to explain their reasoning. After receiving the prediction and explanation from the LLMs, we compare the results with the ground truth label and explanation. The results of the case study for *BridePrice*, *Alibaba-Math*, *DrinkWater*, and *CS2* are presented in Tables 7, 8, 9, and 10, respectively, using four LLMs: ChatGPT-4o (OpenAI, 2023), Qwen 2.5 (Yang et al., 2024), DeepSeek-V3 (DeepSeek-AI et al., 2024), and Claude3.5 (Anthropic, 2025). The results demonstrate:

(1) LLMs struggle with parody detection. For example, the parody comment in Table 7 takes an extreme position opposing the viewpoint that a boyfriend should hand over his salary, yet all the LLMs classify this as a non-parody comment. Additionally, the comment in Table 10, which directly expresses a dislike toward the G2 team with analysis, is identified as a parody by 3 of the 4 LLMs.

(2) LLMs frequently provide incorrect explanations when identifying parody comments. Even in the case of *DrinkWater*, shown in Table 9, where all the LLMs successfully identify the comment as a parody, they fail to generate accurate explanations. The explanations indicate that the LLMs rely mostly on the style and tone of the comment, without a deeper understanding of the implicit meaning.

In conclusion, these results suggest that LLMs struggle to understand parody comments, as they both fail to provide accurate predictions and offer misleading explanations. This highlights the need for further development in LLMs for the task of parody detection.

C Implementation Details

In this section, we provide implementation details of all the methods used in Section 3. Except from Large Language Models (LLMs), all the other methods are trained on 300 epochs, with an early stopping of 5. We use Adam optimizer to update model parameters. The experiments are conducted on a linux server with Ubuntu 20.04, trained on a single NVIDIA RTX A5000 GPU with 24GB memory. All the methods are trained on train set, the hyperparameters are searched on validation set, where the search space is given by:

- Hidden Dimension: {16, 32, 64, 128},
- Learning Rate: {5e-6, 1e-5, 2e-5, 3e-5, 5e-5, 1e-4},
- Weight Decay: {1e-5, 1e-4},
- Batch Size: {16, 32},

For the task of parody detection, the threshold for each dataset is the same for all the methods. Specific, we let the threshold be 0.9415 for *Alibaba-Math*, 0.9526 for *BridePrice*, 0.9691 for *DrinkWater*, 0.9387 for *CS2*, 0.9262 for *Campus-Life*, 0.9406 for *Tiktok-Trump*, 0.8768 for *Reddit-Trump*

Prior to feeding the data into the model, we utilize over sampling with replacement for parody detection, and use Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) for sentiment classification to balance the training data.

Apart from these common settings, we introduce the detailed implementations of each specific model as follows.

BoW+MLP (Mikolov, 2013) Bag of Words (BoW) is a kind of word embedding method. In this study, the BoW model implemented in Word2Vec (Mikolov, 2013), aiming to predict a target word based on its surrounding context words. Before using Bag of Words, we standardize text input, remove unnecessary whitespace variations, tokenization text into individual words, and filter out high-frequency words that may not contribute much meaning. Next, we use Bag of Words in Word2Vec to get the word embedding, setting vector size to 50, window to 10, min count to 1, epochs to 50.

Multi-Layer Perceptrons (MLP) is a kind of feed-forward neural network. In our study, we employ a three-layer MLP, with a dropout rate set to 0.3 and ReLU as the activation function.

Skip-gram+MLP (Mikolov et al., 2013) Skip-gram is a word embedding method which learns word representations by predicting context words given a target word. Before using Skip-gram, we standardize text input, avoid unnecessary whitespace variations, the text is tokenized into individual words, and filter out high-frequency words that may not contribute much meaning. Then we use Skip-gram in Word2Vec, setting vector size to 50, window to 10, min count to 1, epochs to 50. The part of MLP is the same as in BoW+MLP.

RoBERTa+MLP (Liu, 2019) RoBERTa (Robustly Optimized BERT Pretraining Approach) is an advanced variant of BERT. The part of Next sentence prediction (NSP) is removed from RoBERTa's pre-training objective. To obtain embedding of textual data, we use mean embedding method to compute the average of token embedding from last hidden state. Setting max length to 256, batch size to 32. The part of MLP is the same as in BoW+MLP.

BNS-Net (Zhou et al., 2024) The propagation mechanism in BNS-Net is defined as:$H = f(X, U, W)$, where X represents the textual features, U denotes user embeddings, and W is the weight matrix. The Behavior Conflict Channel (BCC) applies a Conflict Attention Mechanism (CAM) to extract inconsistencies in behavioral patterns, while the Sentence Conflict Channel (SCC) leverages external sentiment knowledge (e.g., SenticNet) to detect implicit and explicit contradictions. BNS-Net is trained using a multi-task loss function, which combines sarcasm classification and sentiment inconsistency modeling: $L = \lambda_1 J_{\text{sar}} + \lambda_2 J_{\text{imp}} + \lambda_3 J_{\text{exp}} + \lambda_4 J_{\text{balance}}$, where: sar is the sarcasm classification loss,imp and exp correspond to implicit and explicit sentiment contradiction losses. Balance is a balancing term to mitigate bias toward dominant classes. The balancing coefficients used in experiments are: $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$, $\lambda_4 = 0.2$.

DC-Net (Liu et al., 2021) The Dual-Channel Network is a dual-channel architecture to realize sarcasm detection by capturing the contrast between literal sentiment and implied sentiment. The model consists of Decomposer, literal channel, implied channel and analyzer. Prior to feeding data into DC-Net, we utilize the opinion lexicon from nltk 3.9.1 to identify the positive and negative word in our datasets. Following the methodology outlined in the original paper, it needs to use GLOVE to obtain the embedding and vocabulary. To generate the literal and implied sentiment labels, we leverage the parody label along with the counts of positive and negative words. These labels are then processed separately in the two channels. Finally the analyzer measure the conflicts between the channels. In our datasets, we follow the original paper and set all of the loss contributions λ_1, λ_2, λ_3 of our DC-Net model are set to 1.

QUIET (Liu et al., 2023a) The Quantum Sarcasm Model detects sarcasm in text by using quantum-inspired techniques. It converts text and context inputs into dense vector representations through an embedding layer. These embeddings undergo quantum encoding, where sine and cosine functions simulate quantum amplitude and phase encoding, capturing complex relationships. The encoded features are averaged to reduce dimensionality, then passed through a hidden layer with ReLU activation. A sigmoid output layer predicts whether a comment is sarcastic or not. The model addresses class imbalance with class weights and evaluates performance using precision, recall, and F1-score. This single-modality model applies quantum-inspired methods to enhance feature transformation for sarcasm detection.

SarcPrompt (Liu et al., 2023b) is a prompt-tuning method for sarcasm recognition that enhances PLMs by incorporating prior knowledge of contradictory intentions. The framework comprises two key components: (1) Prompt Construction. (2) Verbalizer Engineering. In our implementation, we adopt the question prompt approach and design bilingual templates tailored to Chinese and English datasets. For Chinese parody detection, we construct the prompt as " {COMMENT} 这段话是在反串吗？ {MASK}.". For English datasets, we design"{COMMENT} Are you parody? {MASK}." To enhance model interpretability and alignment with domain knowledge, we employ a verbalizer as paper, where domain-specific label words are mapped based on dataset statistics. In parody detection, we use words like "反串", "是", "parody", "no". In sentiment classification, we use words like "支持", "反对", "support", "oppose". The total loss combines cross-entropy (classification) and contrastive losses (enhancing intra-class consistency): $L(\theta) = \lambda_1 L_{\text{sarc}}(\theta) + \lambda_2 L_{\text{con}}(\theta)$, where $\lambda_1 = 1$ and λ_2 is selected from {0.05, 0.1, 0.2, 0.5, 1} via validation, following the original paper's hyperparameter selection.

GCN (Kipf and Welling, 2016) All Graph Neural Networks (GNNs), including GCN, GAT, and GraphSAGE, are implemented using PyTorch Geometric (Fey and Lenssen, 2019), with the version specified as 2.6.1. For the GCN, we set the number of graph convolution layers to 2, the size of the hidden embedding to 64, and the dropout rate to 0.5. Additionally, we incorporate residual connections (He et al., 2016) and layer normalization (Ba et al., 2016) to enhance model performance, as suggested by Luo et al. (2024).

GAT (Veličković et al., 2017) In GAT, we adopt the same configuration as in Graph Convolutional Networks (GCN), utilizing 2 graph convolution layers, a hidden embedding size of 64, and a dropout rate of 0.5. Additionally, we set the number of attention heads to 8.

GraphSAGE (Hamilton et al., 2017) In GraphSAGE, we adopt the same configuration as in Graph Convolutional Networks (GCN), utilizing 2 graph convolution layers, a hidden embedding size of 64, and a dropout rate of 0.5. Additionally, we set the neighborhood size to 5.

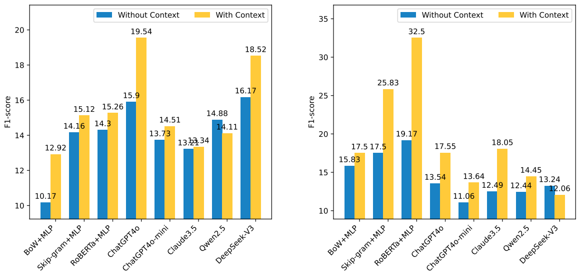
LLMs we employ a variety of LLMs from different companies to perform parody detection and sentiment classification, which include ChatGPT-4o (and 4o-mini) (OpenAI, 2023), ChatGPT-o1-mini (OpenAI, 2024), ChatGPT-o3-mini (OpenAI, 2025) Claude 3.5 (Anthropic, 2025), Qwen 2.5 (Yang et al., 2024), DeepSeek-V3 (DeepSeek-AI et al., 2024), and DeepSeek-R1 (DeepSeek-AI et al., 2025).They require different kinds of input formats, objects and parameters. Except reasoning model, we set temperature to 0, which reasoning model not support this object. For reasoning model, they have to use more and more tokens to complete the reasoning procedure before outputting the content. To optimize model performance, we design task-specific prompts, ensuring that each LLM receives input formulations tailored to the characteristics of parody detection and sentiment analysis. For example, in parody detection, we design the prompt as *"You are a helpful assistant trained to classify whether a statement is parody or not."* in the system role, and *"Determine whether the following comment is parody:{text}*\n *Directly output 1 for parody, 0 for non-parody."* in the user role. In particular, ChatGPT o1-mini doesn't have the system role, so we input all in the user role.

D Additional Results

This section introduces additional results in our experiments. We introduce more results of the influence of context to parody detection in Section D.1 and the influence of parody to sentiment classification in Section D.2. Then, we show the performance comparison of reasoning LLMs and non-reasoning LLMs in Section D.3. Last, we investigate the impact of train ratio of embedding-based models compared with LLMs in Section D.4.

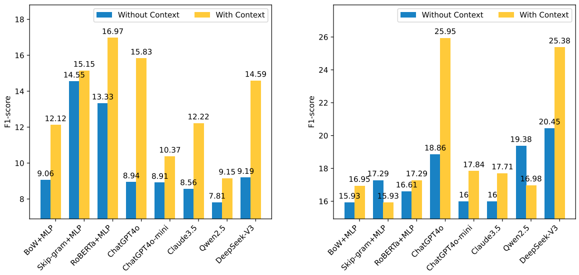
D.1 Influence of Context to Parody Detection

Figure 5 illustrates the detailed results of the performance comparison of the F1 score in parody detection with and without context across seven datasets. Generally, contextual information significantly enhances model performance on most datasets and methods. For instance, on *Alibaba-Math*, the performance of ChatGPT4o improves from 15.9 to 19.54, while on *BridePrice*, the performance of RoBERTa+MLP increases from 19.17 to 32.50. These results indicate that contextual information is beneficial for parody detection. This finding aligns with the results in Bamman and Smith



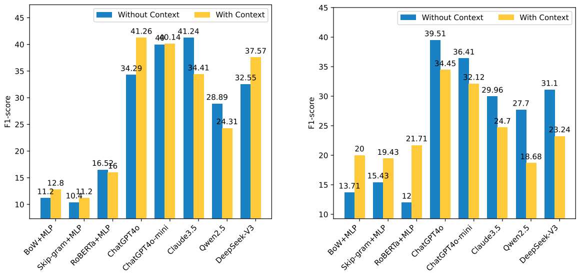
(a) Alibaba-Math

(b) BridePrice



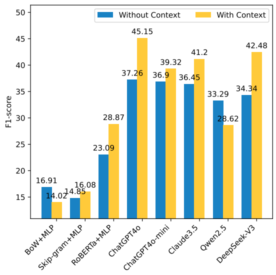
(c) DrinkWater

(d) CS2



(e) CampusLife

(f) Tiktok-Trump



(g) Reddit-Trump

Figure 5: Impact of contextual information on parody detection across seven datasets.

(2015); Wang et al. (2015), which show that providing dialogue as context significantly improves model performance in sarcasm detection.

However, although contextual information significantly improves model performance on most datasets, there are still some datasets where context does not enhance or even decreases model performance. For example, on *Tiktok-Trump*, the model performance decreases, and on *CampusLife*, the performance remains similar after adding contextual information. This suggests that contextual information may not always contribute to improving model performance in parody detection.

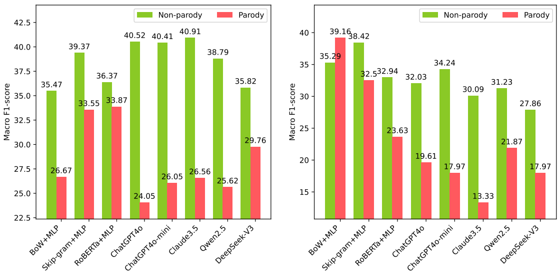
D.2 Influence of Parody to Sentiment Classification

Figure 6 presents the detailed model performance of comment sentiment classification on parody and non-parody comments across seven datasets. In the *DrinkWater* dataset, large language models (LLMs) such as ChatGPT-4o-mini (F1-score: 51.42) and Qwen2.5 (F1-score: 47.00) achieve competitive performance compared to embedding-based methods like Bag of Words (BoW) (F1-score: 48.21), Skip-gram (F1-score: 47.11), and RoBERTa (F1-score: 44.93) when parody is not present. However, for parody comments, the performance of LLMs degrades significantly, falling below that of embedding-based approaches. For instance, ChatGPT-4o drops from an F1-score of 48.7 to 19.04, and ChatGPT-4o-mini declines from 51.42 to 15.53, whereas embedding-based methods exhibit greater robustness, with BoW decreasing from 48.21 to 36.21, Skip-gram from 47.11 to 32.35, and RoBERTa from 44.93 to 33.83. Overall, these results indicate that parody presents substantial challenges for sentiment classification, and LLMs struggle to maintain their advantage over traditional embedding-based methods in this context.

D.3 Reasoning LLMs in Parody Detection

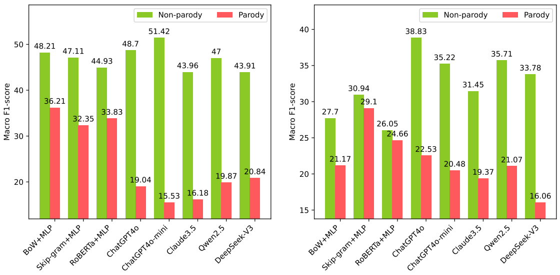
We present the details of reasoning LLMs in parody detection across six datasets in Figure 7. Our findings indicate that reasoning LLMs do not exhibit a performance advantage compared to non-reasoning LLMs. For instance, ChatGPT-o1-mini and ChatGPT-o3-mini underperform ChatGPT4o-mini on the *CampusLife* and *Tiktok-Trump* datasets. Additionally, DeepSeek-R1 significantly underperforms DeepSeek-V3 across all datasets.

These results suggest that reasoning does not always enhance LLM performance in parody detec-



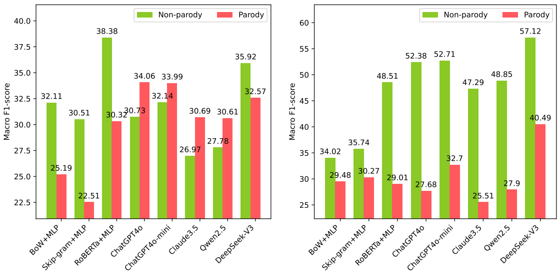
(a) Alibaba-Math

(b) BridePrice



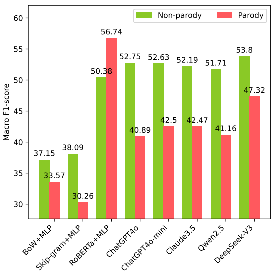
(c) DrinkWater

(d) CS2



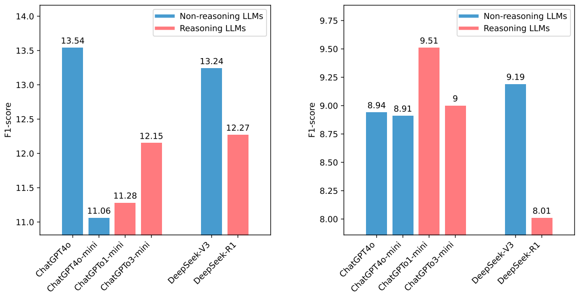
(e) CampusLife

(f) Tiktok-Trump

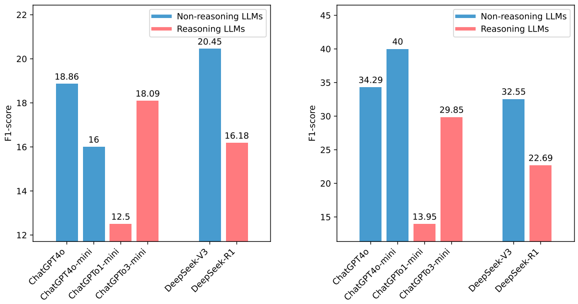


(g) Reddit-Trump

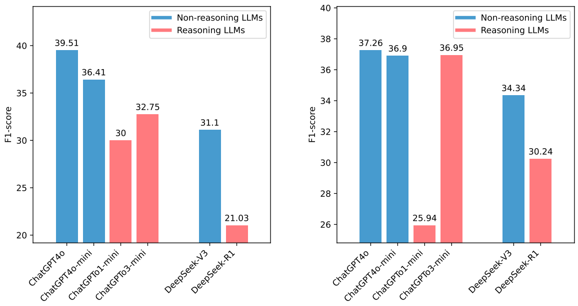
Figure 6: Impact of parody on comment sentiment classification across seven datasets.



(a) BridePrice (b) DrinkWater



(c) CS2 (d) CampusLife

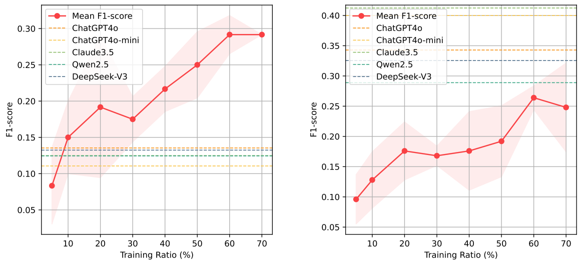


(e) Tiktok-Trump (f) Reddit-Trump

Figure 7: A Comparative Performance Analysis of Reasoning vs. Non-Reasoning LLMs

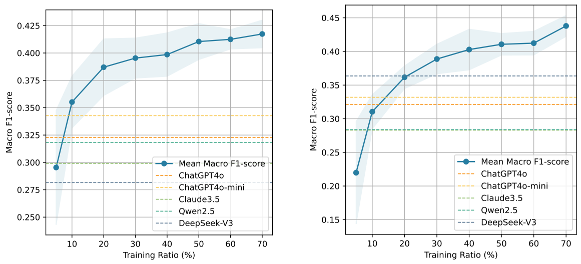
tion. We speculate that this may be due to the nature of parody, which often relies on indirect or subtle cues related to tone, context, and nuance rather than direct logical inference. In such cases, logical reasoning may impair basic semantic sensing ability of LLMs, which lead to over-complicated but wrong conclusions.

D.4 Impact of Supervision Ratio



(a) BridePrice (b) CampusLife

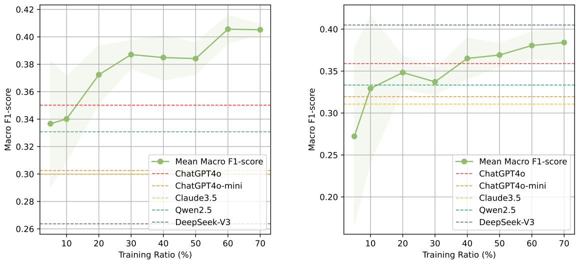
Figure 8: Impact of training ratio to RoBERTa+MLP on parody detection



(a) BridePrice (b) CampusLife

Figure 9: Impact of training ratio to RoBERTa+MLP on comment sentiment classification

The embedding-based methods used in our experiments require explicit training on labeled data, whereas LLMs like RoBERTa do not require such training once pre-trained. Therefore, the performance of embedding-based models depends on the size and quality of the training set. To explore this, we investigate how varying the training ratio influences model performance by gradually increasing the training set size while keeping the test set constant. The results for RoBERTa+MLP under different train ratio are presented in Figures 8, 9, and 10 for parody detection, comment sentiment classification, and user sentiment classification. In all tasks, we observe that the performance increases monotonically with the training ratio, highlighting the



(a) BridePrice (b) CampusLife

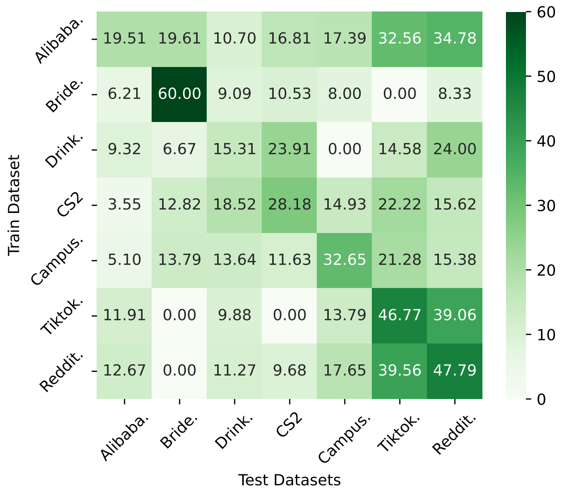
Figure 10: Impact of training ratio to RoBERTa+MLP on user sentiment classification

benefit of additional training data for embedding-based methods.

In addition, on the *BridePrice* dataset, only 10% supervision is enough for RoBERTa to outperform all LLMs in parody detection, indicating a limitation of LLMs in domain-specific tasks. This suggests that fine-tuned models like RoBERTa perform better with minimal supervision in specialized contexts. In contrast, on the *CampusLife* dataset, RoBERTa's performance consistently falls below that of all LLMs, regardless of the training ratio. This suggests that LLMs are more effective in tasks requiring generalizable knowledge and flexibility, such as parody detection in diverse, context-rich domains. These results demonstrate that LLMs remain powerful in specific areas requiring flexibility in adapting to diverse linguistic contexts and nuanced understanding, while embedding-based models like RoBERTa excel in more targeted, domain-specific tasks.

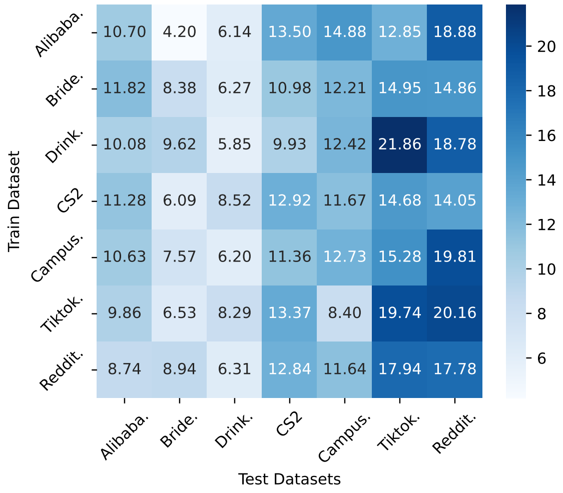
D.5 Transfer Performance of LLM-SFT

To investigate whether LLMs can generalize parody detection across different datasets, we evaluate Supervised Finetuned (SFT) LLMs trained on different datasets across seven datasets by measuring the F1 score. As shown in Figure 11, we conduct experiments on two LLMs, Gemma-2-9B and DeepSeek-R1-7B, as they exhibit relatively better performance in the supervised finetuning experiments. The results indicate that the transferability of SFT-LLMs is low across datasets with different topics. For example, Gemma-2-9B trained on *BridePrice* achieves strong performance on this dataset but fails to generalize well to others. However, for datasets within the same topic, such as Tiktok-Trump and Reddit-Trump, both of which focus on the U.S. election, transferability is higher due to shared knowledge between the datasets. Still, these results do not demonstrate that SFT-LLMs can develop a generalizable ability for parody detection across diverse datasets, indicating that parody detection remains highly dataset-specific and lacks robustness in cross-domain settings.



	Alibaba.	Bride.	Drink.	CS2	Campus.	Tiktok.	Reddit.
Alibaba.	19.51	19.61	10.70	16.81	17.39	32.56	34.78
Bride.	6.21	60.00	9.09	10.53	8.00	0.00	8.33
Drink.	9.32	6.67	15.31	23.91	0.00	14.58	24.00
CS2	3.55	12.82	18.52	28.18	14.93	22.22	15.62
Campus.	5.10	13.79	13.64	11.63	32.65	21.28	15.38
Tiktok.	11.91	0.00	9.88	0.00	13.79	46.77	39.06
Reddit.	12.67	0.00	11.27	9.68	17.65	39.56	47.79

(a) Gemma2-9B



	Alibaba.	Bride.	Drink.	CS2	Campus.	Tiktok.	Reddit.
Alibaba.	10.70	4.20	6.14	13.50	14.88	12.85	18.88
Bride.	11.82	8.38	6.27	10.98	12.21	14.95	14.86
Drink.	10.08	9.62	5.85	9.93	12.42	21.86	18.78
CS2	11.28	6.09	8.52	12.92	11.67	14.68	14.05
Campus.	10.63	7.57	6.20	11.36	12.73	15.28	19.81
Tiktok.	9.86	6.53	8.29	13.37	8.40	19.74	20.16
Reddit.	8.74	8.94	6.31	12.84	11.64	17.94	17.78

(b) DeepSeek-R1-7B

Figure 11: Transfer performance of supervised finetuning LLMs across 7 datasets

Question
对于“男朋友彩礼给的少工资卡也不愿意上交我很犹豫怎么办？”这一话题，有网友评论：“ 这种男的不要也罢。姐妹要相信自己的价值，好男人很多，一定不要委屈自己。 一生一次他还不愿意付出所有留退路，你觉得那个退路会是谁？消耗掉的可是你无价的青春啊。”
请判断这条评论是否为反串，并给出50字以内的解释。
Regarding the topic "My boyfriend gives little betrothal money and refuses to hand over his salary card—I’m hesitant, what should I do?", a netizen commented: "This kind of man isn’t worth it. Sis, believe in your value—there are many good men, never settle for less." "If he won’t give his all for a once-in-a-lifetime event and leaves himself an out, who do you think that’s for? What’s being wasted is your priceless youth."
Please determine whether this comment is satire and provide an explanation within 50 words.

Gronud Truth
是。这条评论通过使用“姐妹”这一称呼模仿女性视角，并以极端立场全面否定男性。
Yes. The comment mimics a female perspective by using the term "sister" and takes an extreme stance by completely negating men.

ChatGPT-4o ✗ Wrong Answer
不是。这条评论表达了女性在婚姻中应得到足够的经济保障和重视，强调自我价值和青春无价，反映了对传统彩礼和工资卡上交的重视，并非反串。
No. This comment expresses that women should be given adequate financial security and attention in marriage, emphasizes self-worth and the pricelessness of youth, and reflects the importance attached to traditional betrothal gifts and salary card submission. It is not a reverse portrayal.

Qwen2.5 ✗ Wrong Answer
不是。这条评论明确支持女性追求平等对待，强调自我价值和寻找更合适伴侣的重要性，没有表现出相反意图或讽刺意味。
No. This comment clearly supports women’s pursuit of equal treatment, emphasizes the importance of self-worth and finding a more suitable partner, and does not show any contrary intention or irony.

DeepSeek-V3 ✗ Wrong Answer
不是。这条评论表达了女性在婚姻中应得到足够的经济保障和重视，强调自我价值和青春无价，反映了对传统彩礼和工资卡上交的重视，并非反串。
No. This comment expresses that women should be given adequate financial security and attention in marriage, emphasizes self-worth and the pricelessness of youth, and reflects the importance attached to traditional betrothal gifts and salary card submission. It is not a reverse portrayal.

Claude3.5 ✗ Wrong Answer
不是。该评论语气、用词和观点都符合特定群体的真实想法，包括"姐妹""价值""青春"等关键词的使用方式自然,不存在夸张或讽刺的痕迹。
No. The tone, words and opinions of the comment are consistent with the real thoughts of a specific group, including the use of keywords such as "sisters", "values" and "youth" in a natural way without any trace of exaggeration or sarcasm.

Table 7: A case study of LLMs using an example from *BridePrice*.

Question
对于“2024年阿里巴巴数学竞赛预赛中，一位来自中等专业学校的女生取得12名的优秀成绩”这一话题，有网友评论道：“这位同学有实力！阿里巴巴有眼光！ 请阿里巴巴破格录取进入达摩院，助力阿里科技快速发展"
请判断这条评论是否为反串，并给出50字以内的解释。
Regarding the topic "My boyfriend gives little betrothal money and refuses to hand over his salary card—I’m hesitant, what should I do?", a netizen commented: "This kind of man isn’t worth it. Sis, believe in your value—there are many good men, never settle for less." "If he won’t give his all for a once-in-a-lifetime event and leaves himself an out, who do you think that’s for? What’s being wasted is your priceless youth."
Please determine whether this comment is satire and provide an explanation within 50 words.

Gronud Truth
是。这条评论通过对这位同学夸张的赞誉以及对阿里巴巴达摩院破格录取的期望，来引起这位同学反对者的争论。
Yes. This comment aroused controversy among the student’s opponents by exaggerating the praise for the student and expressing the hope that he would be admitted to Alibaba Damo Academy as an exception.

ChatGPT-4o ✓ Correct Answer, 👍 Good Explanation
是，因其夸张用词（如“破格录取”）和不匹配的期待（达摩院门槛极高）。若非反串，也可能是戏谑或调侃，并非严肃建议。
Yes, due to its exaggerated wording (e.g., "exceptional admission") and unrealistic expectations (since DAMO Academy has extremely high entry requirements). If not parody, it could be playful or mocking rather than a serious suggestion.

Qwen2.5 ✗ Wrong Answer
不是。它表达了对姜萍实力的认可和对阿里巴巴的期待，希望阿里能破格录取她，支持科技发展。语气积极正面。
No. It expresses recognition of This student’s abilities and expectations for Alibaba, hoping the company will make an exception to admit her and support technological development. The tone is positive and optimistic.

DeepSeek-V3 ✓ Correct Answer, 👍 Good Explanation
是，因其将中专背景与破格进入顶级研究院关联，隐含对选拔标准的不信任，通过夸张建议暗讽现实中的学历门槛或企业宣传策略，点赞表情强化了反讽效果。
Yes, as it links a vocational school background with exceptional admission to a top research institute, implying skepticism towards the selection criteria. The exaggerated suggestion subtly mocks real-world academic barriers or corporate PR strategies, with the thumbs-up emoji reinforcing the parodic tone.

Claude3.5 ✗ Wrong Answer
不是。评论者真诚肯定了姜萍的实力，并对阿里巴巴表达了期望。用语朴实，表达方式也符合普通网友的评论习惯，包括使用表情来强调观点。
No, The commenter sincerely acknowledges this student’s capabilities and expresses expectations for Alibaba. The language is simple and aligns with the typical commenting style of regular netizens, including the use of emojis to emphasize their point.

Table 8: A case study of LLMs using an example from *Alibaba-Math*.

Question 关于“一位博主为了鼓励员工喝水，将茶水吧改成游戏厅”的话题中，有网友评论：“构思的很好很巧妙，好构思啊，构思.” 请判断这条评论是否为反串，并给出50字以内的解释。 Regarding the topic of "A video creator encouraged employees to drink more water by turning the tea station into a game arcade," a netizen commented: "The idea is well thought out and very clever. Great idea, really great idea." Please determine whether this comment is parody and provide an explanation within 50 words. **Gronud Truth** 是。这条评论通过"构思"谐音“狗屎”，来嘲讽该博主的方法缺乏实质内容，不切实际。 Yes. This comment mocks the video creator’s method as lacking substance and being impractical by using the Chinese homophone for "idea" which is pronounced as "dog shit".
ChatGPT-4o ✓ Correct Answer, 👎 Bad Explanation 是。重复“构思”一词带有夸张和调侃意味，结合“很巧妙”可能是在讽刺方案脱离实际，甚至适得其反，体现出一种戏谑或反讽的态度 Yes. The repetition of the word "idea" carries an exaggerated and mocking tone. Combined with "very clever," it may be satirizing the impracticality of the plan or even suggesting it could backfire, reflecting a playful or ironic attitude.
Qwen2.5 ✓ Correct Answer, 👎 Bad Explanation 是。表面上看似在夸赞，但重复“构思”显得夸张且略带讽刺，暗指将茶水吧改成游戏厅的方案并不合理，实际效果可能适得其反。 Yes. While it appears to be a compliment, the exaggerated repetition of "idea" has a slightly ironic tone, implying that turning a tea station into a game arcade is not a reasonable plan and might actually be counterproductive.
DeepSeek-V3 ✓ Correct Answer, 👎 Bad Explanation 是。通过三次重复"构思"制造机械感，配合原事件逻辑悖论（游戏厅促进喝水），构成典型反讽结构——表面夸赞实则暗示想法荒诞，符合中文网络阴阳怪气表达范式。 Yes The triple repetition of "idea" creates a mechanical effect, combined with the logical paradox of the original event (a game arcade promoting water consumption). This forms a classic structure of irony—appearing as praise but actually hinting at the absurdity of the idea, aligning with the parodic style of Chinese internet humor
Claude3.5 ✓ Correct Answer, 👎 Bad Explanation 是。通过重复使用"构思"一词并且语气刻意夸张，表现出一种模仿式的嘲讽，暗示这个创意可能过于简单或缺乏实质内容。 Yes. The repeated use of the word "idea" and the deliberately exaggerated tone create a form of imitative mockery, implying that the concept may be overly simplistic or lacking in substance.

Table 9: A case study of LLMs using an example from *DrinkWater*.

Question 关于“CS2赛事中，G2战队连续7次输给NAVI战队”这一话题，有网友评论：“hunter就是个垃圾，每个战队只能换2人否则取消名额的机制阻碍了90%战队的重建，也帮他这样的混子混了不少工资。monesy永远不配和zywoo相提并论，抖一步名不虚传，甚至天梯打一个集锦操作就被发到各大平台上，然而这届赛事一张图平均被手撕3次 ” 请判断这条评论是否为反串，并给出50字以内的解释。 In the topic "In CS2 tournaments, G2 has lost to NAVI seven times in a row," a netizen commented:"hunter is just trash. The rule that a team can only replace two players or lose their slot has blocked 90% of team rebuilds, and has also let freeloaders like him earn quite a bit of salary. monesy will never deserve to be compared to zywoo, best Rifler on TikTok is a well-deserved reputation. He even gets a highlight posted on social media platforms for a single flashy move in ranking, yet in this tournament, he gets destroyed an average of three times per map ." Please determine whether this comment is satire and provide an explanation within 50 words. **Gronud Truth** 不是。这条评论通过直接攻击选手（huNter 与 m0NESY）来表达对G2战队和选手的不满与反对。 No. This comment expresses dissatisfaction and opposition to the G2 team and players by directly attacking the players (hunter and monesy).
ChatGPT-4o ✗ Wrong Answer 是。评论内容极端贬低选手，带有夸张的情绪表达（如“垃圾”“混子”“永远不配”），同时使用表情符号强化讽刺意味，符合常见的反串风格。 Yes. The content excessively criticizes the players, uses exaggerated emotional expressions (such as "trash" and "freeloader"), and includes emojis to enhance its sarcastic tone, which is characteristic of satirical posts.
Qwen2.5 ✗ Wrong Answer 是。评论中使用了夸张的语气和明显的偏见，如“hunter就是个垃圾”和“monesy永远不配和zywoo相提并论”，并且通过讽刺的口吻（如“抖一步名不虚传”）可能意在调侃而非严肃批评，疑似反串黑。 Yes, It uses an exaggerated tone and clear bias, such as calling "hunter trash" and claiming "monesy is never worthy of being compared to zywoo." The sarcastic wording (e.g., "shake-and-shoot is well-deserved") suggests an intention to mock rather than genuinely criticize, making it seem like satirical trolling.
DeepSeek-V3 ✗ Wrong Answer 是。评论使用夸张贬低言辞（如"垃圾""混子"）、网络热梗（"抖一步"）及表情符号，通过极端负面评价和刻意对比（monesy/zywoo）模仿无脑黑粉语气，意在讽刺而非真实批判。 Yes. The comment employs exaggerated derogatory terms ("trash," "freeloader"), internet slang ("shake-and-shoot"), and emojis, mimicking the speech style of overly critical fans. The extreme negativity and deliberate comparisons (monesy vs. zywoo) suggest an intent to mock rather than provide a serious critique.
Claude3.5 ✓ Correct Answer, 👍 Good Explanation 不是。解释：用词尖锐且带有明显的情绪色彩，但论点具体且符合CS2玩家的专业表达方式，提到了具体的选手数据和规则限制，显示出对赛事的真实关注和了解。 No. Explanation: The wording is sharp and emotionally charged, but the arguments are specific and align with the professional discourse of CS2 players. It mentions specific player statistics and rule restrictions, indicating a genuine interest and understanding of the tournament.

Table 10: A case study of LLMs using an example from *CS2*.