A Multi-Aspect Evaluation Framework for Synthetic Data: Case Study on Irony and Sarcasm

Laura Majer[♡] **Ana Barić**[♡]
Florijan Sandalj **Ivan Unković** **Bojan Puvača** **Jan Šnajder**
TakeLab, Faculty of Electrical Engineering and Computing, University of Zagreb
{name.surname}@fer.hr

Abstract

Data augmentation (DA) using large language models (LLMs) is a cost-effective method for generating synthetic data, particularly for tasks with scarce datasets. However, its potential remains largely underexplored, both in terms of augmentation configuration and evaluation of synthetic data. This paper investigates LLM-based synthetic data generation for irony and sarcasm, two subjective and context-dependent forms of figurative language. We propose a multi-aspect evaluation framework assessing synthetic data's *utility–plausibility* and *extrinsic–intrinsic* dimensions through four aspects: predictive performance, sample diversity, linguistic properties, and human judgment. Our findings indicate that other aspects of evaluation, like diversity and linguistic features, do not necessarily correlate with an increase in predictive performance, underscoring the importance of multi-faceted evaluation. This work highlights the potential of LLM-based DA for irony and sarcasm detection, offering insights into the linguistic competence of LLMs. As synthetic data becomes increasingly prevalent, our framework offers a broadly applicable and crucial evaluation method, particularly for linguistically complex tasks.

takelab/llm-irony-sarcasm

1 Introduction

Irony and sarcasm are complex pragmatic phenomena that present a significant challenge in both linguistics and NLP. This is primarily due to their ambiguity, contextual dependency, subjectivity, and the inherent incongruence between expressed and intended meaning. Some datasets have only a few dozen examples per category, which is insufficient for training supervised models or fine-tuning pre-trained classifiers for irony and sarcasm detection. The challenge is further aggravated by the fact that linguistics lacks clear definitions of sarcasm and irony, as well as a consistent distinction between the two (see Appendix 5 for examples). This inconsistency is reflected in dataset annotations (Jang and Frassinelli, 2024), which makes it difficult to align or merge datasets. Thus, despite increasingly sophisticated language models, detecting irony and sarcasm remains a challenging task and a benchmark of true language understanding.

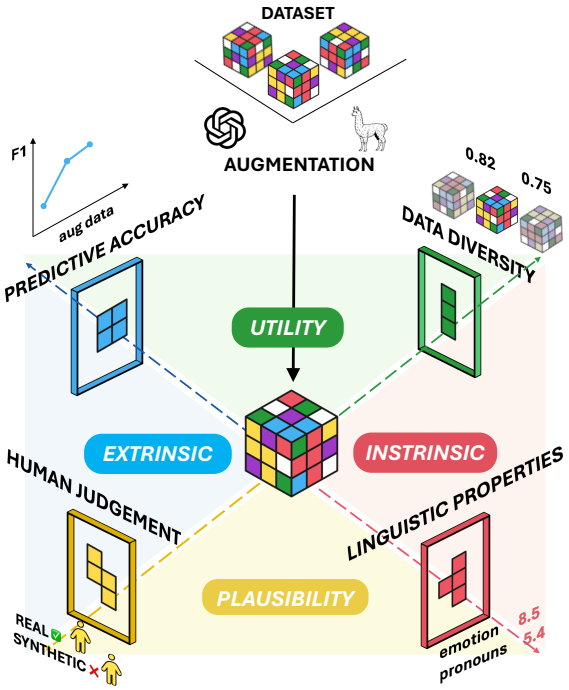


Figure 1: Multi-aspect framework for evaluating synthetic data quality through four aspects spanning the *utility–plausibility* and *extrinsic–intrinsic* dimensions.

In NLP, data augmentation (DA) is a widely adopted technique used to expand existing datasets by generating additional training examples through various transformations, improving model robustness and performance, especially in low-resource settings (Feng et al., 2021; Chen et al., 2023). While previous DA approaches relied on simple linguistic transformations, such as paraphrasing,

[♡]Equal contribution

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 305–323
March 29, 2026

synonym replacement, or back-translation (Wei and Zou, 2019), Large Language Models (LLMs) have proven more effective due to their strong language fluency, enhancing the plausibility and diversity of generated data (Ding et al., 2024). Another advantage is LLMs' capacity for in-context learning (ICL) (Brown et al., 2020), which enables the use of detailed instructions and few-shot demonstrations to produce higher-quality synthetic data.

However, despite its proven utility, LLM-based DA remains underexplored, both regarding augmentation configurations and the evaluation of synthetic data. The latter is especially important for linguistically complex and subjective tasks, such as sarcasm and irony detection. While prior work often used predictive accuracy as a proxy for synthetic data quality (Yoo et al., 2021), this metric alone can be misleading. Performance gains may stem from spurious correlations (Zhou et al., 2024), and may be limited when synthetic instances lack complexity or closely resemble the original data. Finally, improvements from synthetic data, even when the data is diverse and boosts in-domain performance, do not demonstrate that it genuinely captures the targeted linguistic phenomena or that it is indistinguishable from real data to proficient language users. This makes evaluation beyond downstream performance essential, not only for validating the synthetic data itself but also for analyzing the generative model that produced it. Systematic analysis of the generated instances can reveal which linguistic properties are consistently encoded and which are lacking, indicating that the model relies on shortcuts or artifacts. In this sense, synthetic data serves as a diagnostic probe: examining its structure and limitations provides direct evidence of the model's inductive biases and the extent to which it reproduces the target phenomenon, insights that zero- and few-shot classification performance alone cannot provide.

In this paper, we address the lack of systematic assessment of synthetic data for complex tasks by proposing a multi-aspect framework for evaluating its quality (Figure 1). This framework considers four key aspects of synthetic data: predictive accuracy, data diversity, linguistic properties, and human judgment. Together, these aspects span two key dimensions of data evaluation: the *extrinsic–intrinsic* dimension (examining how synthetic data appears externally versus its inherent characteristics) and the *utility–plausibility* dimension (assessing its usefulness for a given task versus its resemblance to real data).

We apply the proposed framework to LLM-based DA for irony and sarcasm detection. Considering various models, prompt configurations, and numbers of demonstrations, we aim for novel insights into the impact of DA configuration choices on the different aspects of the generated synthetic data. In particular, we focus on the following research questions: **RQ1**. How does synthetic data influence fine-tuning, and is it sufficiently diverse both internally and relative to the few-shot demonstrations? **RQ2**. Does synthetic data accurately capture the linguistic phenomena it is intended to represent, and how closely it resembles – or differs from – the original data it was modeled on? **RQ3**. Which augmentation setups produce synthetic data that scores best across multiple evaluation aspects, and do the different aspects correlate?

Our analyses show that while adding synthetic data improves predictive performance, the synthetic data points are mostly easy-to-learn and exhibit lower diversity. We also demonstrate that human annotators do not easily distinguish synthetic from original data, verifying the potential of using DA for irony and sarcasm detection. Overall, our multi-aspect evaluation shows that different DA configurations optimize for various quality aspects of synthetic data. Thus, rather than considering only the predictive performance on augmented data, our approach gives rise to Pareto-optimal sets of DA configurations, from which DA practitioners should choose based on the specific aspects they aim to maximize. Our work emphasizes the importance of thorough and multifaceted evaluation of synthetic data, particularly for subjective and linguistically complex tasks, and provides a framework applicable across domains. We make all synthetic datasets generated and used in this study publicly available.

2 Related Work

Our work connects two research areas: (1) irony and sarcasm classification, and (2) data augmentation using LLMs. The following is an overview of the main research done in these areas.

2.1 Irony and sarcasm detection

In linguistics, there is no agreement on definitions for sarcasm and irony, or their clear distinction (see Appendix 5 for examples). The primary distinction between the two categories lies in the target. According to this view, irony is general, whereas

sarcasm is more directed and malevolent (Averbeck, 2013), often linked to ridicule (Lee and Katz, 1998), and characterized by a mocking, sharper tone. Due to the lack of standardized definitions for these phenomena, distinguishing between intended and perceived irony or sarcasm can be difficult – especially in contexts with limited cues. This ambiguity makes manual data annotation challenging, as guidelines are often imprecise. As a result, many datasets are constructed by scraping the web for ironic hashtags without involving human annotators (Khodak et al., 2018), fail to differentiate between sarcasm and irony (Van Hee et al., 2018), or include only a small number of examples for each category (Abu Farha et al., 2022).

In response to these annotation and dataset limitations, recent research has sought to improve the robustness and nuance of sarcasm and irony detection. For instance, Jang and Frassinelli (2024) conduct both intra- and cross-dataset evaluations to assess the generalizability of sarcasm detection models, while Casola et al. (2024) emphasize the importance of a perspectivist approach by building an irony corpus from social media that includes disaggregated annotations and annotator metadata.

2.2 Data augmentation with LLMs

Since data annotation is both expensive and time-consuming (Wang et al., 2021), data augmentation (DA) often provides a more cost-effective and scalable alternative. In NLP, DA takes many forms, ranging from simple techniques such as word insertion, swapping, and synonym replacement (Wei and Zou, 2019) to more sophisticated methods utilizing pre-trained transformers for low-resource scenarios (Kumar et al., 2020).

Recently, LLMs have garnered significant attention for DA. Piedboeuf and Langlais (2025) demonstrate that LLM-based methods – particularly zero- and few-shot prompting – can outperform classical DA techniques, confirming their effectiveness. One early method, GPT3Mix (Yoo et al., 2021), introduced a prompt-based approach that selects two random samples from the dataset and includes them in the prompt alongside dataset-specific meta-information. Building on this idea, Møller et al. (2024) enhanced it with a task-specific system prompt to better guide generation. In a different line of work, Yang et al. (2024) proposed a multi-step framework that iteratively trains the model and augments the dataset using challenging, model-generated examples.

When evaluating DA methods, most approaches focus solely on predictive model performance when synthetic data is added to the training set. However, for subjective and linguistically rich tasks like irony and sarcasm detection, this can overlook important aspects such as semantic diversity and linguistic adherence. In this work, we broaden the evaluation criteria for data augmentation methods to include not only model performance but also data quality, focusing on semantic diversity and linguistic properties through the multi-aspect framework.

3 Experimental Setup

Datasets. For modeling irony and sarcasm, we utilize a dataset from SemEval-2022 Task 6, introduced by Abu Farha et al. (2022). We use subtasks A and B that determine whether a tweet is either sarcastic or non-sarcastic, and classify the specific type of ironic speech, respectively. More concretely, subtask B classifies each tweet into one of six ironic speech categories defined by Leggitt and Gibbs (2000). For our experiments, we used two main categories: sarcasm and irony, consisting of 893 and 175 tweets, respectively. In the original train-test split, the training set consisted of 713 sarcasm tweets and 155 irony tweets, whereas the test set consisted of 180 sarcasm tweets and 20 irony tweets. Furthermore, we randomly sampled non-sarcastic tweets from subtask A dataset. Using these, we constructed two balanced datasets: one for sarcasm, consisting of 1,786 tweets, and one for irony, consisting of 350 tweets.

Models. We use instruction-tuned LLMs since we formulate our prompts as n-shot instructions. We opt to include both an open-source and a closed-source instruction-following model: Llama-3.1-8B-Instruct (llama3; AI@Meta (2024)) and GPT-4o (gpt4; OpenAI et al. (2024)), respectively.

Prompt construction. To construct LLM prompts for ironic and sarcastic tweet augmentation, we follow Møller et al. (2024), who focused on tasks involving evaluative language, including sentiment analysis, offensive speech detection, and emotion recognition. We divide the prompt format into two parts, following the taxonomy proposed by Jeoung et al. (2025): (1) fixed and (2) variable parts. The fixed part of the prompt is present in all configurations and includes a short task description (e.g., *Your task is to generate X ironic tweets.*) and output format instruction (e.g., *Separate each new instance with ordinal numbers.*), along with a

list of tweet examples used as demonstrations for the few-shot prompts (either 1- or 3-shot). In the 1-shot setting, the model is prompted to generate three new instances, while in the 3-shot setting, it is prompted to generate nine instances. Unlike the fixed part, the variable parts, which include the task definition and stylistic guidance, are either added to or omitted from the prompt, depending on the prompt configuration. In total, we evaluate 16 distinct prompt configurations for each task and model, generated by combining two variable components with both 1- and 3-shot scenarios.

To ensure we cover all relevant aspects of the datasets we aim to augment, we obtain definitions for irony and sarcasm from the iSarcasm dataset (Oprea and Magdy, 2020). Furthermore, we create stylistic guidance instructions by describing the informal and unstructured style typical of tweets. Before selecting a specific wording for both the fixed and variable parts of the prompt, we conduct an initial evaluation of different prompt variants on a subset of data for each task to identify the most effective one in terms of instruction-following quality. In subsequent analyses, we label each prompt configuration with '1' or '3' (for 1-shot and 3-shot, respectively), followed by letters indicating the included variable parts ('D' for Definition, 'S' for Stylistic guidance, and 'X' for neither). For example, '3DS' denotes a 3-shot prompt with both Definition and Stylistic guidance included. The complete prompts, including all variable parts, are shown in Table 7 in Appendix E.1. In what follows, we refer to a specific combination of an LLM and prompt configuration as *augmentation setup*.

Demonstration selection. We utilize all training examples as demonstrations in 1-shot configurations and employ a cluster-based selection method to choose semantically similar examples for each prompt in 3-shot configurations for each task. Concretely, we obtain training data embeddings from BERT (Devlin et al., 2019), reduce dimensionality with PCA, and apply K-Means to cluster semantically similar examples.

4 Utility

We now apply the proposed evaluation framework to analyze the synthetic data generated by the DA procedure described above. Our results are presented along the *utility–plausibility* dimension, beginning with utility. Utility captures the extent to which synthetic data supports model training and performance. Specifically, we assess utility through two aspects: predictive accuracy (an *extrinsic* aspect) and data diversity (an *intrinsic* aspect).

4.1 Predictive Accuracy

Since DA expands datasets often too small for model fine-tuning, improved predictive performance in classification is typically used as a proxy for synthetic data quality (Wei and Zou, 2019; Møller et al., 2024; Piedboeuf and Langlais, 2025).

Original data. To establish baseline performance on the original data, we fine-tune three transformer-based models commonly used for irony and sarcasm detection: BERT (Devlin et al., 2019), BERTweet (Nguyen et al., 2020), and RoBERTa (Zhuang et al., 2021). Each model is fine-tuned five times with different random seeds on the task-specific training sets, using default hyperparameters (see Appendix B). In addition, we evaluate two LLMs in 0-shot and 2-shot settings: Llama-3.1-8B-Instruct (llama3; AI@Meta (2024)) and Gemma-3-4B (gemma3; Team et al. (2025)). For prompting, we use the model-specific chat templates and provide test examples as input. The full set of baseline prompts is provided in Appendix C.1. Table 1 reports the average F1 scores on the test sets. For the fine-tuned models, results are averaged over five runs evaluated on a shared test set. To assess overall performance differences among models, we apply the Friedman test. For the irony task, the test does not reveal statistically significant differences among models ($\chi^2(\text{df} = 4) = 10.03$, $p = 0.124$); therefore, no post-hoc pairwise comparisons are conducted. For the sarcasm task, the Friedman test indicates statistically significant differences among models ($\chi^2(\text{df} = 4) = 26.57$, $p < 0.001$). We subsequently perform post-hoc pairwise comparisons using Wilcoxon signed-rank tests with Holm correction to account for multiple comparisons. However, none of the pairwise comparisons remain statistically significant after correction (all Holm-adjusted p-values > 0.05). Since none of the models significantly outperforms the others across both tasks, we select BERTweet as the baseline model for the remainder of the study, as it achieves the highest average F1 score.

Synthetic data. We measure the influence of synthetic data on model performance by fine-tuning BERTweet, our best-performing model on original data, with real and synthetic data for all prompt configurations. To examine how performance changes with different synthetic-to-real data ratios, we grad-

	Model	Irony	Sarcasm
FT	BERT	0.60 ± 0.17	0.53 ± 0.02
	BERTweet	0.68 ± 0.08	0.70 ± 0.07
	RoBERTa	0.39 ± 0.22	0.64 ± 0.07
0-shot	llama3	0.68 ± 0.06	0.56 ± 0.01
	gemma3	0.44 ± 0.02	0.37 ± 0.01
2-shot	llama3	0.64 ± 0.03	0.73 ± 0.01
	gemma3	0.65 ± 0.02	0.73 ± 0.01

Table 1: F1 scores (mean $\pm$ std) of models trained on the original dataset for irony and sarcasm detection tasks.

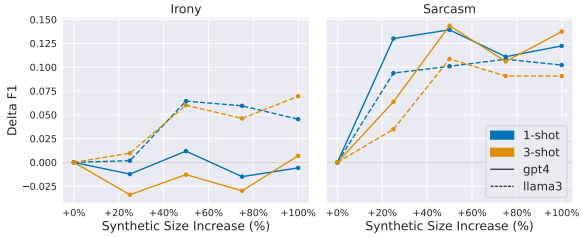


Figure 2: F1 score change relative to original training data across increasing synthetic data sizes, grouped by few-shot setting (color) and model type (line style), for irony and sarcasm detection. Averaged over five runs and across configurations; positive values indicate performance gains from synthetic augmentation.

ually increase the amount of synthetic data in the training set following the approach of Møller et al. (2024). This setup also helps identify whether there is an upper limit to the amount of synthetic data beyond which model performance plateaus. Figure 2 shows the change in F1 score aggregated across different few-shot settings (either 1-shot or 3-shot). The addition of synthetic data yielded a maximum performance improvement of 10% for irony and a higher improvement of 15% for sarcasm; similar performance gains were also reported in (Møller et al., 2024). Additionally, we observe that, across both tasks and all few-shot settings, performance stabilizes once synthetic data comprises approximately 50% of the total training set. Additionally, we notice a larger performance gain for sarcasm. Similar trends are observed for prompts extended with definition and style parts (see Appendix C.2).

Motivated by the observed plateau in model performance as the amount of synthetic data increases, we employ dataset cartography (Swayamdipta et al., 2020) to assess if this stagnation stems from differences in example difficulty distribution between the original and synthetic datasets. We fine-tune BERTweet for 3 epochs using a balanced set of

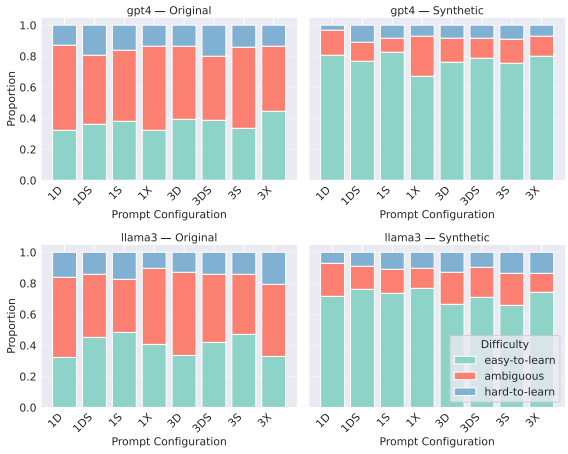


Figure 3: Distribution of difficulty levels (easy-to-learn, ambiguous, and hard-to-learn) across different prompt configurations for the irony detection task, with the original data (left) and the synthetic data (right). Each bar shows the normalized proportion of each difficulty category for each prompt configuration.

original and synthetic data and default hyperparameters, and record the confidence, correctness, and variability scores, later used to identify easy-to-learn, ambiguous, and hard-to-learn examples. Figure 3 shows the differences in difficulty distribution between original and synthetic data for irony. We observe higher proportions of easy-to-learn examples in synthetic data across both tasks and prompt configurations. A similar trend is observed for sarcasm (see Figure 7 in the appendix).

4.2 Data Diversity

The stagnation in F1 scores despite adding more synthetic data, coupled with generally volatile training curves and the observation that many synthetic samples fall into the easy-to-learn region of the data map, directly motivates data diversity as another evaluation aspect.

Method. We quantify the diversity of the synthetic dataset through the semantic similarity between pairs of individual instances. We use the SBERT model (Reimers and Gurevych, 2019) to calculate instance-level similarity, with the zero-one output range, reflecting the degree of semantic overlap between a pair of instances. We then investigate the semantic diversity of synthetic datasets across different configurations, approaching it from two angles. The first type of diversity we consider is *intra-diversity*, motivated by the observation that synthetic examples often contain repetitive phrases or recurring themes that may superficially boost

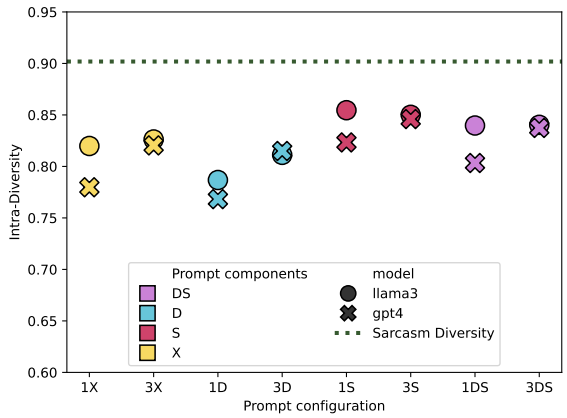


Figure 4: Intra-diversity plots for sarcasm synthetic datasets across prompt setups (shown in color) and models (shown in symbols).

performance but ultimately offer limited variety – potentially explaining the observed performance plateau. Given a synthetic dataset of size M and semantic similarity $\mathrm{sim}(\cdot,\cdot)$ between two synthetic instances, $\tilde{x}_i$ and $\tilde{x}_j$, we compute intra-diversity as:

$$d_{\text{intra}} = 1 - \frac{1}{M}\sum_{i \neq j} \mathrm{sim}(\tilde{x}_i, \tilde{x}_j)$$

Second, since the few-shot demonstrations are the only component that varies between runs, the generation process may produce instances that are too similar to the demonstrations. To assess this, we calculate *inter-diversity*, which quantifies the extent to which the synthetic instances are different from the few-shot demonstrations included in the prompt used to generate them (see Appendix C.3 for the formula).

Original data. To compare augmented datasets with the original dataset, we calculate intra-diversity for the original irony and sarcasm datasets. The resulting scores are 0.888 for irony and 0.902 for sarcasm, respectively.

Synthetic data. Figure 4 shows the intra-diversity values across augmentation setups for sarcasm. Prompt configurations are color-coded, and marker shapes indicate the model. Although synthetic datasets show lower diversity than the original, the difference of $\sim$0.05 of top-performing configurations is small – especially considering sets are three times larger. Prompt configurations strongly influence diversity: style-only prompts produce the most diverse data, while definition-only prompts result in the least. For gpt4, three-shot prompts clearly outperform one-shot prompts. We observe similar patterns for sarcasm, shown in Figure 8. Inter-diversity behaves similarly, but varies less across prompt types and more between one-shot and three-shot setups (see Figure 9 in the appendix). Three-shot setups yield higher diversity across tasks, suggesting that not all examples influence every output, increasing batch diversity.

5 Plausibility

Predictive performance and data diversity assess how synthetic data impacts a model's training and performance when included in its training set, focusing on its *utility*. In contrast, *plausibility* concerns whether the synthetic data exhibits linguistic properties associated with irony and sarcasm and appears realistic compared to the original data. We evaluate plausibility through two aspects: linguistic properties (an *intrinsic* aspect) and human judgment (an *extrinsic* aspect).

5.1 Linguistic Properties

Since irony and sarcasm are inherently linguistic phenomena, synthetic data should ideally conform to specific definitions and linguistic properties for accurate representation. Generating misaligned samples can introduce flawed data points and encourage shortcut learning during fine-tuning, resulting in a model that incorrectly represents irony and sarcasm. Conversely, a model that effectively captures these concepts is more likely to generalize to out-of-distribution data, making linguistic properties an essential aspect of synthetic data.

Method. Among various approaches to modeling linguistic features relevant to irony and sarcasm, we use the LIWC 2022 lexicon (Boyd et al., 2022). This choice is motivated by Sulis et al. (2016), who showed that user-marked figurative language (e.g., irony, sarcasm) exhibits distinct patterns detectable with closed-vocabulary methods. LIWC 2022 includes both basic linguistic categories (e.g., part-of-speech tags, word types) and more nuanced dimensions, such as emotions (anger, anxiety, sadness), psychological states (*Cognition*, *State*), and social behaviors (*Interpersonal conflict*, *Moralization*). It comprises 116 features, each representing the proportion of words in a text that match predefined lexical categories, yielding sparse feature vectors per sample and sparse matrices per dataset (ironic, sarcastic, or neutral). To compare LIWC features between original and synthetic data, we use the non-parametric Common Language Effect Size (CLES) (McGraw and Wong, 1992), which quantifies the

probability that a randomly selected instance from one group scores higher than one from another. Prior studies suggest CLES is more interpretable than traditional effect size metrics (Brooks et al., 2014). We correct for multiple comparisons across all configuration–feature pairs using the Benjamini–Hochberg procedure with a false discovery rate threshold of $\alpha = 0.05$.

Original data. Before comparing linguistic features of original and synthetic data, we first examine differences among ironic, sarcastic, and neutral instances in the original dataset to assess if prior findings (Sulis et al., 2016) hold. Since LIWC includes many features, we focus on a few discriminative categories – *Emotion*, *Social*, *Physical*, and *States* – based on definitions of irony and sarcasm. To better isolate patterns, we average only non-zero LIWC scores. For instance, we find that irony is more associated with anxiety, while neutral instances tend to reflect more positive emotion (see Figure 10 in the appendix).

Synthetic data. Figure 5 shows heatmaps of the CLES score for sarcasm across augmentation setups, with rows representing prompt configurations per model and columns representing LIWC feature categories sorted by average, ascending. After applying FDR correction, all configuration–feature comparisons were statistically significant. Red values indicate a lower presence of the feature in synthetic data compared to the original, while blue values indicate a higher presence of the feature. Paler squares indicate a higher overlap between synthetic and original features.

For both sarcasm and irony and across both models, similar patterns emerge. The category *Cognition*, which includes features such as *certitude*, *causation* and *memory*, was consistently underrepresented in synthetic data, while *Tone*, covering the degree of both positive and negative tone, was consistently overrepresented in synthetic data.

5.2 Human Judgement

To complement the plausibility evaluation, we conduct a human judgment study to assess the quality of the synthetic data from two aspects: *data detectability* and *data authenticity*, by comparing real and synthetic data. Data detectability assesses how well humans can distinguish the difference between irony and sarcasm for real and synthetic data. Data authenticity assesses the extent to which humans perceive synthetic data as believable or realistic

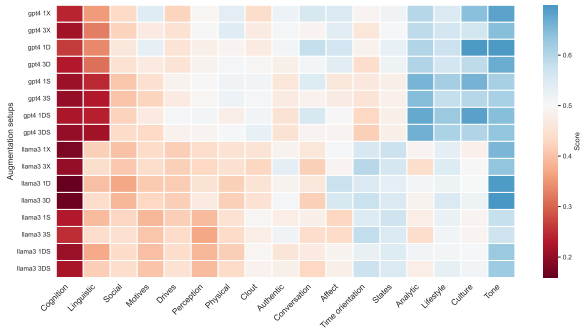


Figure 5: Heatmap of CLES scores for aggregated LIWC features across augmentation setups for sarcasm, lower values represent lower frequency in synthetic data compared to original.

compared to real data.

Annotation setup. Each aspect in the study is evaluated independently through two annotation tasks, with annotations provided by twelve volunteer participants. In both tasks, each example was annotated by three participants. Furthermore, we instructed participants to complete the data detectability task first and then proceed to the data authenticity task only after its completion. A strict task order was imposed to mitigate participant anchoring bias, as awareness of the different example origins (real or synthetic) could influence label judgments in the data detectability task (e.g., basing judgments on perceived origin rather than linguistic fit alone). To avoid this, participants were informed about the inclusion of the synthetic data only after completing the data detectability task. For both tasks, we randomly selected a balanced sample consisting of real and synthetic data, where synthetic data is selected from the prompt configurations that achieve the highest F1 scores for irony (gpt4 1X) and sarcasm (llama3 1X), respectively.

Data detectability. To evaluate whether humans can recognize irony and sarcasm equally well in real and synthetic data, we design an annotation task comparing how these categories are perceived across data sources. The intuition is that if synthetic data is linguistically and semantically plausible, then the ability of humans to identify irony and sarcasm should be comparable to that in real data. To this end, we devise a ternary classification task where annotators label each tweet as either `ironic`, `sarcastic`, or `neutral`, without knowing whether the tweet is real or synthetic. We compile a balanced dataset of 350 examples in total, uniformly sampled from five classes: real irony, real

sarcasm, real neutral, synthetic irony, and synthetic sarcasm. All irony and sarcasm examples, regardless of source, are grouped under their respective labels, while the `neutral` class (which only exists in real data) is treated as a separate category.

Data authenticity. To further inspect the quality of synthetic data compared to real data, we ask participants to label tweets as either `real` or `synthetic`, aiming to assess how convincing the synthetic data is from a human perspective. We uniformly sample a total of 280 ironic and sarcastic tweets from both the real and synthetic datasets. Additionally, we ensure that no examples are repeated from the data detectability setup.

Annotation details. We use the open-source platform Alanno (Jukić et al., 2023) to conduct both annotation tasks. All twelve participants are either at the Master's (4), PhD (7), or postdoctoral level (1), working within the NLP domain. The group includes 11 computer scientists and 1 psychologist. The ages range from 24 to 34 years. All participants are fluent in English; two are native speakers, and ten demonstrate high proficiency.

Results. We assess how well annotator labels align with the gold labels for both real and synthetic data by calculating accuracy scores, which are reported for data detectability and authenticity in Table 2. In absolute terms, detectability scores fall below a random baseline of 0.5, highlighting the challenge of distinguishing irony from sarcasm – further supported by low annotator agreement (Fleiss' $\kappa = 0.35$). Similarly, despite higher accuracy on the data authenticity task, poor agreement (Fleiss' $\kappa = 0.14$) indicates difficulty and subjectivity in distinguishing real from synthetic data. Table 3 presents the confusion matrices for the original and synthetic datasets. The similar distribution of predictions across irony, sarcasm, and neutral labels suggests that the synthetic data exhibits class-level ambiguities and decision patterns comparable to those observed in original data. It is also a confirmation that LLMs generate neutral Detectability accuracy is higher on synthetic examples, whereas authenticity accuracy is higher on real examples. This suggests that synthetic data is easier to label correctly but harder to identify as artificial – likely due to being more straightforward or less nuanced, but still fairly realistic. This aligns with the finding that the majority of synthetic data falls under the easy-to-learn category.

	Irony		**Sarcasm**	
	Real	Synthetic	Real	Synthetic
Detectability	0.25	0.39	0.40	0.52
Authenticity	0.77	0.51	0.81	0.54

Table 2: Detectability vs. authenticity accuracy scores of real and synthetic annotations for irony and sarcasm.

		Predicted		
		Irony	**Sarcasm**	**Neutral**
Original	**Irony**	19	25	26
	Sarcasm	19	27	24
Synthetic	**Irony**	27	15	28
	Sarcasm	28	36	6

Table 3: Confusion matrices for detectability on original and synthetic data.

6 Discussion

We now address our research questions by examining results across multiple evaluation dimensions. **RQ1**, which concerns the impact of synthetic data on predictive performance and diversity, is best informed by the utility dimension. While synthetic data boosts performance, gains tend to plateau – likely due to its high proportion of easy-to-learn examples. Although diversity remains lower than in the original data, it is notable given the increased dataset size. For **RQ2**, the plausibility reveals that synthetic data often lacks variation in expression categories across setups but matches the original data in terms of linguistic markers. Human annotators' difficulty in distinguishing synthetic from real examples further supports its plausibility. However, the clearer distinction between irony and sarcasm in synthetic data – reflected in higher labeling accuracy – suggests improved clarity of linguistic cues, although possibly at the cost of realism.

In contrast to RQ1 and RQ2, which focus on individual dimensions, **RQ3** spans the entire evaluation framework, exploring correlations among evaluation aspects. All aspects, except for human judgment, are directly comparable across setups; we limited human judgment annotations to a single prompt configuration due to resource constraints. The other three aspects are standardized to values in the $[0, 1]$ range: F1 score synthetic data of original size (predictive accuracy), intra-diversity (data diversity), and average CLES divergence from 0.5 (linguistic properties). We report the scores for sarcasm in Table 4; similar trends are observed for irony (see Table 9 in Appendix F).

gpt4				llama3			
Config	F1	Div.	Ling.	Config	F1	Div.	Ling.
3DS★	0.754	0.837	0.085	1D	0.744	0.787	0.080
1DS	0.736	0.803	0.090	3D	0.708	0.811	0.080
1D	0.736	0.768	0.078	1DS ★	0.699	0.840	0.069
3D	0.732	0.815	0.080	3X	0.696	0.826	0.077
3X ★	0.719	0.820	0.068	1X	0.692	0.820	0.071
3S ★	0.708	0.846	0.081	3DS ★	0.658	0.840	0.070
1X	0.696	0.780	0.077	1S ★	0.637	0.855	0.069
1S	0.686	0.823	0.086	1S ★	0.637	0.855	0.069

Table 4: Score comparison across performance (*F1*), diversity (*Div.*), and linguistic properties (*Ling.*) for sarcasm. Cell shading reflects relative score quality per task per model (darker = better; for *Ling.*, lower is better), ★ = Pareto-optimal configuration.

Is predictive accuracy robust enough for DA evaluation? Ideally, synthetic data that improves predictive performance should be both diverse and linguistically consistent with the target phenomena. However, our results show that these qualities do not always align – for example, gpt4-generated data for sarcasm with high F1 scores had lower diversity and linguistic overlap, suggesting that increased variation in data can introduce noise. This suggests that increased variation or richness in language, while desirable in theory, may introduce noise or complexity. These findings highlight the need to evaluate model performance alongside explicit data quality metrics, rather than relying on it as a sole indicator of learning.

Can a single prompt configuration cover all desired aspects? No single prompt configuration optimizes performance, diversity, and linguistic consistency across models and tasks; instead, Pareto-optimal sets emerge for each model. Definition-only prompts (D) improve clarity with minimal complexity, while style-based prompts (S or DS) boost diversity but often hurt performance – especially with gpt4, whose outputs are more expressive. Few-shot settings also matter: 1-shot prompts reduce diversity but largely preserve consistency, indicating that more examples increase variation more than linguistic alignment. Overall, prompt design shapes synthetic data and often requires balancing expressiveness with model learnability. Comparing gpt4 and llama3 highlights the variability of LLMs as data generators. While gpt4 is more sensitive to prompt changes, exhibiting greater shifts in performance, diversity, and linguistic features, llama3 produces more consistent outputs with less variation.

7 Conclusion

In this paper, we introduced a multi-aspect evaluation framework for data augmentation (DA), targeting sarcasm and irony detection. While F1 score improvements are a standard measure of synthetic data utility, we argue that this metric is not sufficient. Our framework incorporates aspects that span the utility–plausibility spectrum and reflect diverse characteristics of the data. Results show that these aspects often do not correlate with F1 gains, revealing important gaps in conventional evaluation. Our findings demonstrate the potential of DA for complex language phenomena and emphasize the value of more nuanced, multi-dimensional evaluation in future work.

Limitations

Metrics. The metrics used to evaluate different aspects may not perfectly capture the nuances of those aspects, as they are inherently relative and difficult to quantify. Nevertheless, to our knowledge, this remains the most comprehensive evaluation of DA data to date.

Tasks Our analysis has not been tested on other linguistically complex tasks other than irony and sarcasm, which may reveal different patterns or lead to alternative conclusions. Other tasks may also involve additional crucial aspects not covered here, particularly those that extend beyond standard metrics like F1. This only confirms the importance of multi-aspect evaluation.

Models. Lastly, we evaluated predictive performance using only a single model, which limits the generalizability of our findings. Different models may interact with the data in distinct ways, potentially leading to different outcomes, especially since predictive performance is an indirect measure that reflects how the data influences model learning.

Acknowledgments

We thank our volunteer annotators for their valuable contribution to our work.

References

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*,

pages 802–814, Seattle, United States. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Joshua M. Averbeck. 2013. Comparisons of ironic and sarcastic arguments in terms of appropriateness and effectiveness in personal relationships. *Argumentation and Advocacy*, 50(1):47–57.

Ryan L. Boyd, Abhishek Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. The development and psychometric properties of LIWC-22. Technical report, University of Texas at Austin, Austin, TX.

M. E. Brooks, D. K. Dalal, and K. P. Nolan. 2014. Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, 99(2):332–340.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

John D. Campbell and Albert N. Katz. 2012. Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes*, 49(6):459–480.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.

Sullam Jeoung, Yueyan Chen, Yi Zhang, Shuai Wang, Haibo Ding, and Lin Lee Cheong. 2025. Promptprism: A linguistically-inspired taxonomy for prompts. *Preprint*, arXiv:2505.12592.

Josip Jukić, Fran Jelenić, Miroslav Bićanić, and Jan Snajder. 2023. ALANNO: An active learning annotation system for mortals. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 228–235, Dubrovnik, Croatia. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. *Preprint*, arXiv:1704.05579.

Roger Kreuz and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, 118:374–386.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Christopher J. Lee and Albert N. Katz. 1998. The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol*, 13(1):1–15.

John S Leggitt and Raymond W Gibbs. 2000. Emotional reactions to verbal irony. *Discourse processes*, 29(1):1–24.

Kenneth O. McGraw and S. P. Wong. 1992. A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The parrot dilemma:

Human-labeled vs. LLM-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192, St. Julian's, Malta. Association for Computational Linguistics.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2024. The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks. *Preprint*, arXiv:2304.13861.

Sachi Nakamura. 1995. How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

OpenAI et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

Penny M. Pexman. 2008. It's fascinating research: The cognition of verbal irony. *Current Directions in Psychological Science*, 17(4):286–290.

Frédéric Piedboeuf and Philippe Langlais. 2025. On evaluation protocols for data augmentation in a limited data scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3428–3443, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in twitter: Differences between irony, sarcasm and not. *Knowledge-Based Systems*, 108:132–143. New Avenues in Knowledge Bases for Natural Language Processing.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Gemma Team et al. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Akira Utsumi. 2000. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Shuangtao Yang, Xiaoyi Liu, Xiaozheng Dong, and Bo Fu. 2024. Mini-DA: Improving your model performance through minimal data augmentation using LLM. In *Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (DaSH 2024)*, pages 25–30, Mexico City, Mexico. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. 2024. Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2586–2614, Miami, Florida, USA. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Various irony and sarcasm definitions

Citation	Summary
(Kreuz and Glucksberg, 1989)	*Verbal irony, in contrast, need not refer explicitly to an ironic event or state. Instead, in verbal irony a speaker expresses an attitude toward some object, event, or person by saying something that is not literally true. Sarcasm is often characterized as a form of verbal irony and has been defined as "a sharp and often satirical or ironic utterance designed to cut or give pain."*
(Nakamura, 1995)	*The allusional pretense theory of irony posits two necessary conditions for understanding ironic intent: allusion to failed expectation (FE) and pragmatic insincerity. An allusion to FE refers to a discrepancy between a certain expectation and the subsequent reality. Pragmatic insincerity, the second necessary component proposed by Kumon-Nakamura et al. (1995), follows from the felicity condition, originally discussed by Austin (1962) and later elaborated on by Searle (1979; see also Colston, 2000). When being ironic, the individual is not intending the compliment and is thus being insincere, violating the felicity condition. This insincerity is pragmatic, relating to use rather than semantics. From this view, both FE and pragmatic insincerity must be present to invite irony.*
(Utsumi, 2000)	*Implicit display theory provides an additional contextual constraint for irony. A statement must be identified as being in an "ironic environment," which consists of: (1) an expectation by the speaker at time t, (2) a failure of that expectation, and (3) a negative emotional reaction to the incongruity. Ironic communication implicitly displays this environment by: (1) alluding to the expectation, (2) violating a pragmatic principle, and (3) expressing a negative attitude indirectly.*
(Pexman, 2008)	*Constraint satisfaction theories argue that no single cue is necessary to evoke sarcastic irony. Instead, irony is interpreted based on a set of probabilistic cues or hints evaluated in context. These might include who made the comment, the discourse or environmental support, and the manner of delivery. A single strong cue or multiple weaker cues may be sufficient to invite an ironic interpretation.*
(Campbell and Katz, 2012)	*Findings show that although several components have been claimed as "necessary" for sarcasm, none are truly required. Each factor independently predicts sarcasm ratings, and stylistic choices may play a significant role. Items with sarcastic instructions are rated more sarcastic than those without, even when other variables are controlled. Lexical features may contribute uniquely, outside traditional theory components.*
(Lee and Katz, 1998)	*This study supports that sarcasm often involves ridicule of a specific victim, differentiating it from irony. Sarcastic utterances bring to mind expectations associated with an identified person, while irony evokes broader, collective expectations.*
(Averbeck, 2013)	*Verbal irony is an intentionally inconsistent message that is usually benevolent and not directed at the listener. Sarcasm, however, is also inconsistent but malevolent and directed at the addressee. This distinction is based on intent and interpersonal direction.*

Table 5: Various definitions in related work for phenomena covered by the term *ironic speech*.

B Hyperparameters and Hardware Details

All models are fine-tuned using the HuggingFace Trainer[1] library on an NVIDIA RTX 3090 GPU (24GB RAM) with CUDA 12.9. We use the default hyperparameter settings provided by the Trainer library.

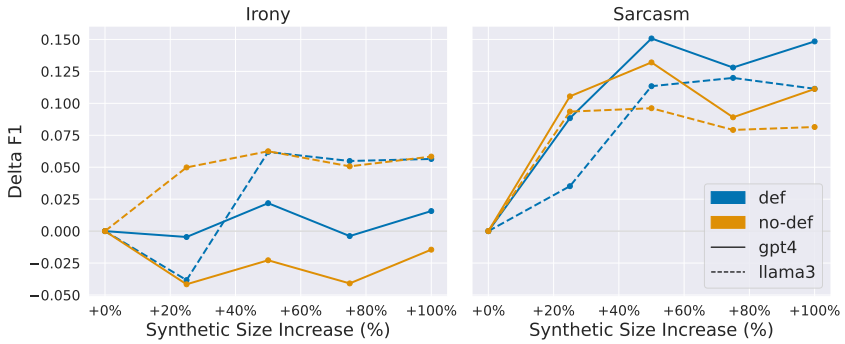
C Utility – additional experiments and information

C.1 Baseline prompts

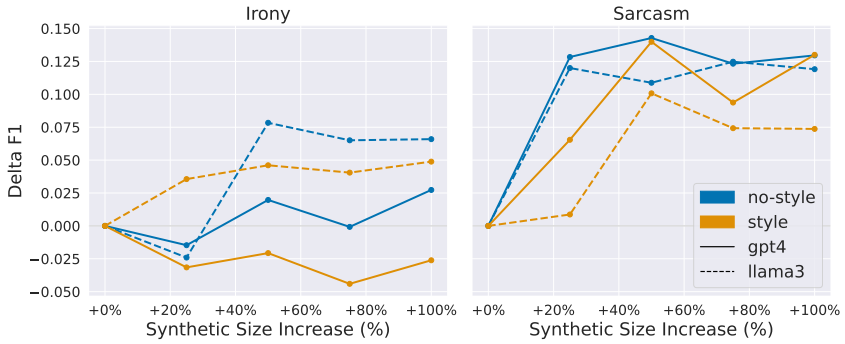
Task	System prompt
Sarcasm	You are a helpful assistant. Classify the following tweet into sarcastic or not sarcastic. Sarcasm is a rhetorical device where the intended meaning is opposite to the literal meaning, often used to mock, criticize, or highlight contradictions in a humorous or satirical way. Sarcastic tweets contradict the state of affairs, are directed towards an addressee and express a critical attitude. Output only the classification result.
Irony	You are a helpful assistant. Classify the following tweet into ironic or not ironic. Irony is a rhetorical device where the intended meaning is opposite to the literal meaning, often used to mock, criticize, or highlight contradictions in a humorous or satirical way. Ironic tweets contradict the state of affairs but are not obviously critical toward any particular addressee. Output only the classification result.

Table 6: System prompts used for classification tasks.

C.2 Predictive accuracy



(a) Grouped by definition setting.



(b) Grouped by style setting.

Figure 6: Change in F1 scores relative to baseline across increasing synthetic data sizes.

[1] https://huggingface.co/docs/transformers/main_classes/trainer

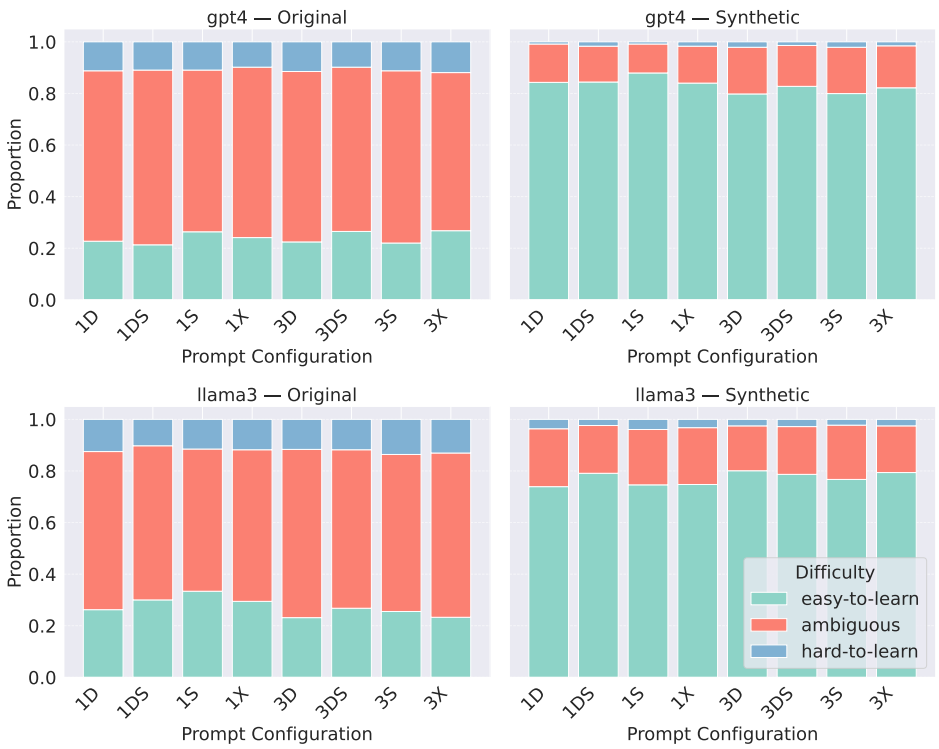


Figure 7: Comparison of the distribution of difficulty levels (easy-to-learn, ambiguous, and hard-to-learn) across different prompt configurations for sarcasm.

C.3 Data diversity

Given a real dataset partitioned into R calls $\{X_1, X_2, \ldots, X_R\}$, and a semantic similarity function $\mathrm{sim}(\cdot, \cdot)$ between real instances $x \in X_r$ and their corresponding synthetic augmentations $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3$, we compute *inter-diversity* as:

$$d_{\text{inter}} = 1 - \frac{1}{R} \sum_{r=1}^{R} \left(\frac{1}{k \cdot |X_r|} \sum_{x \in X_r} \sum_{i=1}^{3} \mathrm{sim}(x, \tilde{x}_i) \right)$$

This metric captures how distinct the synthetic data is from its corresponding real data across all calls.

D Plausibility – additional experiments and information

D.1 Linguistic properties

CLES. To focus on the data and avoid speculating about relevant linguistic categories, we use Common Language Effect Size (CLES) to select the most relevant ones for distinguishing between ironic, sarcastic, and neutral instances.

CLES is defined as:

$$\mathrm{CLES} = P(X > Y) + 0.5P(X = Y)$$

where X and Y are randomly chosen values from the distribution of the two datasets.

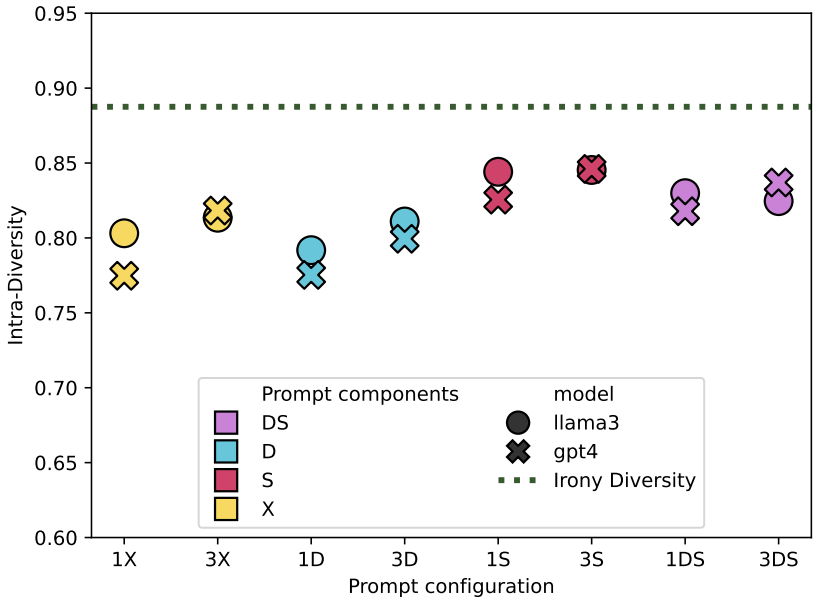
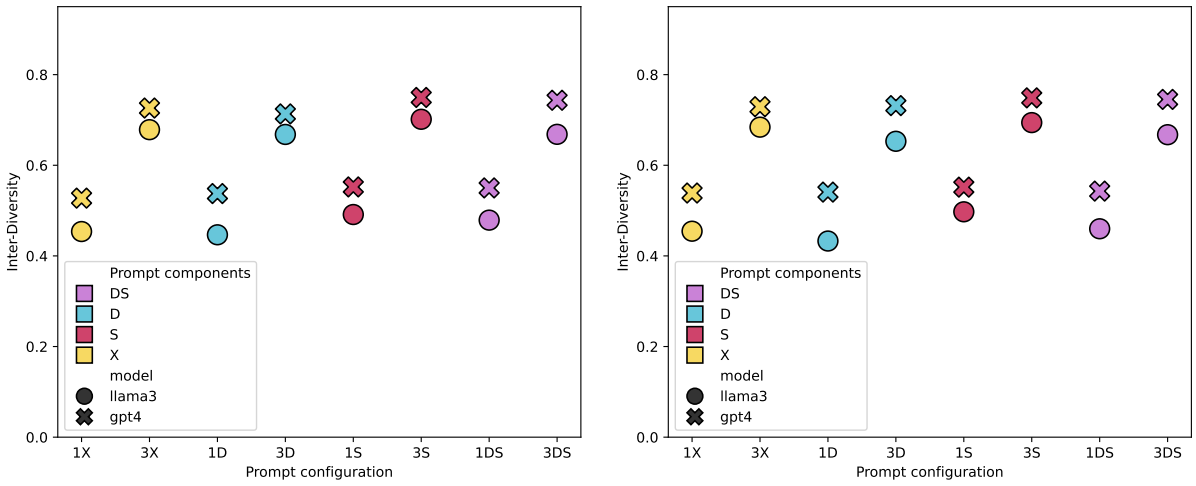


Figure 8: Intra-diversity plot for irony synthetic datasets across prompt setups (shown in color) and models (shown in symbols).



(a) Irony intra-diversity

(b) Sarcasm intra-diversity

Figure 9: Inter-diversity plots for irony and sarcasm synthetic datasets across prompt configurations (shown in color) and models (shown in symbols).

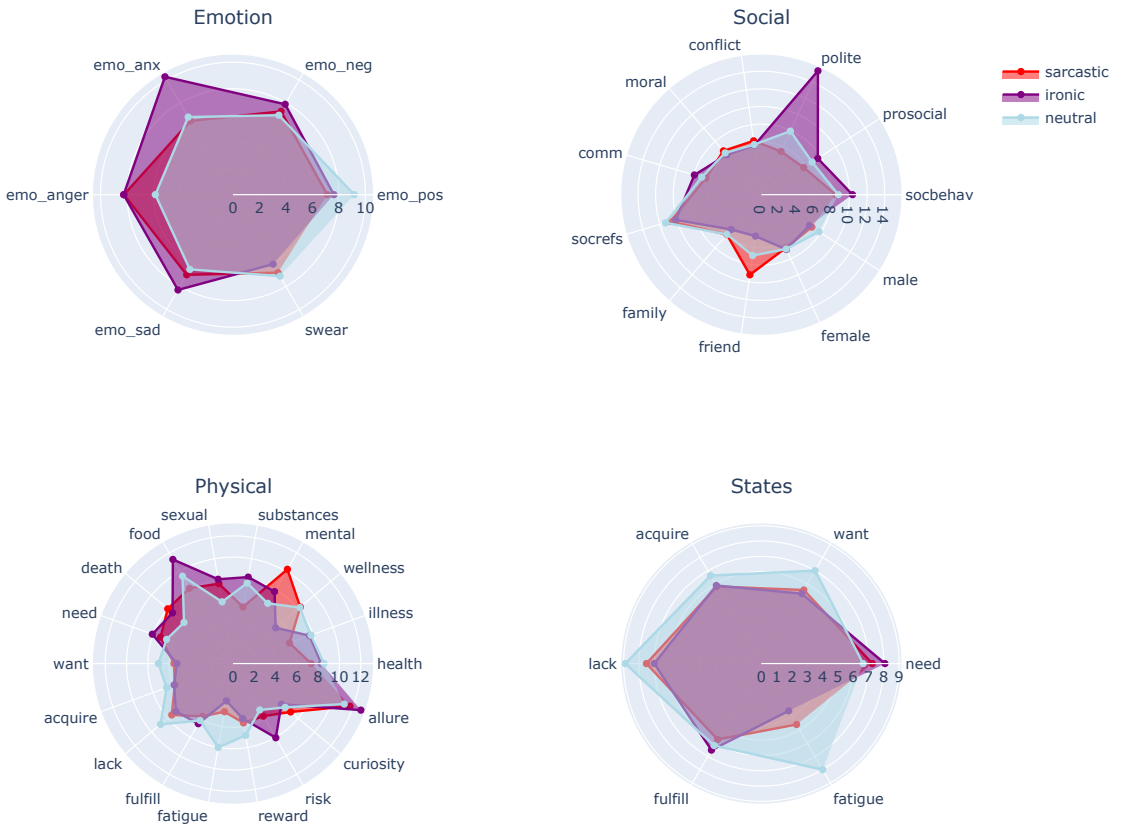


Figure 10: Spider plot for average non-zero LIWC values across notable categories for neutral, ironic, and sarcastic datasets.

E Augmentation setup

E.1 Prompt examples

Table 7 shows the prompts used for the data augmentation.

Task	Prompt
Irony	Your task is to generate 9 ironic Tweets. **Irony is a rhetorical device where the intended meaning is opposite to the literal meaning, often used to mock, criticize, or highlight contradictions in a humorous or satirical way. Ironic tweets contradict the state of affairs but are not obviously critical toward any particular addressee.** *The Tweets should be written in unstructured Twitter style — no formal grammar, proper capitalization, or structured sentences required. Focus on mimicking the style of the given tweet examples, such as mixing formal and informal grammar, structured and unstructured text, and occasionally using emojis and URLs. Pay attention to syntactic properties like ellipses and punctuation. Minimize the use of interjections and conversational phrases (e.g., 'love', 'nothing like', 'just', 'woke up', 'amazing', etc.).* Separate each new instance with ordinal numbers. Examples: - <example tweet> - <example tweet> - <example tweet> Ironic Tweets: 1.
Sarcasm	Your task is to generate 9 sarcastic Tweets. **Sarcasm is a rhetorical device where the intended meaning is opposite to the literal meaning, often used to mock, criticize, or highlight contradictions in a humorous or satirical way. Sarcastic tweets contradict the state of affairs, are directed towards an addressee and express a critical attitude.** *The Tweets should be written in unstructured Twitter style — no formal grammar, proper capitalization, or structured sentences required. Focus on mimicking the style of the given tweet examples, such as mixing formal and informal grammar, structured and unstructured text, and occasionally using emojis and URLs. Pay attention to syntactic properties like ellipses and punctuation. Minimize the use of interjections and conversational phrases (e.g., 'love', 'nothing like', 'just', 'woke up', 'amazing', etc.).* Separate each new instance with ordinal numbers. Examples: - <example tweet> - <example tweet> - <example tweet> Sarcastic Tweets: 1.

Table 7: Prompts for 3-shot setup used to generate ironic and sarcastic tweets. The **bolded** text provides the **definition**, while the *italic* portion covers *stylistic guidance*. For the 1-shot setup, the only difference is the requested number of generated tweets in the output (3 instead of 9) and the number of example tweets in the prompt. In ablated versions, definition, style, or both were omitted by removing the respective parts.

E.2 Human judgment study

Table 8 shows the guidelines used in the annotation.

Task	Guidelines
Data detectability	**IRONIC** **Definition**: Irony occurs when the tweet's intended meaning is opposite to its literal meaning. It often emphasizes contradictions or absurdities in a humorous or satirical way. **Usually, there is no clear target or addressee** of the criticism. If a target exists, the tone remains lighthearted and satirical. **Examples of ironic tweets**: 1. *See Brexit is going well.* 2. *I just absolutely LOVE how I've got to work outside for the next 3 days in the heatwave.* 3. *My eldest is having a wild Friday night out. She's going to bingo.* **SARCASTIC** **Definition**: Sarcasm is a form of irony used to mock, ridicule, or convey contempt. While sarcastic tweets also contradict reality, they are marked by a sharper tone and are typically directed at a specific individual or entity (e.g., a person, brand, or group). **Examples**: 1. *Loving season 4 of trump does America. Funniest season yet #DonaldTrump #Trump #MAGA #MAGA2020* 2. *Wouldn't it be cool if I could spontaneously combust.* 3. *When your @Apple delivery gets stolen by a UK Mail delivery driver and the police know who it is but I still don't get taken seriously* **NEUTRAL** **Definition**: Tweets that are literal, informative, express emotion without irony, or do not involve contradiction, criticism, or mockery. **Examples**: 1. *Going to the gym later today.* 2. *Thanks to everyone who came to the event!* 3. *It's hot outside, but I'm staying hydrated.*
Data authenticity	Your task is to read each text sample and label it as either: • **Original** – Written by a human (obtained via Twitter/X) • **Augmented** – Generated by a language model (LLM)

Table 8: Annotation guidelines for human judgement study.

F Multi-aspect score comparison

gpt4				llama3			
Config	F1	Div.	Ling.	Config	F1	Div.	Ling.
1X	0.720	0.775	0.085	3X	0.780	0.813	0.075
3D	0.719	0.799	0.073	1D	0.758	0.792	0.068
1D	0.713	0.775	0.085	3D	0.749	0.811	0.076
3X	0.703	0.818	0.078	3S	0.753	0.845	0.070
1DS	0.696	0.818	0.099	3DS	0.741	0.825	0.073
3DS	0.680	0.837	0.085	1X	0.722	0.803	0.079
3S	0.671	0.846	0.098	1DS	0.723	0.830	0.076
1S	0.594	0.826	0.102	1S	0.724	0.844	0.070

Table 9: Score comparison across performance (*F1*), diversity (*Div.*), and linguistic features (*Ling.*) for irony. Cell shading intensity reflects relative score quality per task per model (darker = better; for *Ling.*, lower is better).