LLMs in Sarcasm Detection? It's elementary! (Or is it?)

Priyanshu Mahato
IISER Kolkata
priyanshum.2003@gmail.com

Aniket Santosh Mishra
IISER Kolkata
mas23ms096@iiserkol.ac.in

Kripabandhu Ghosh
IISER Kolkata
kripaghosh@iiserkol.ac.in

Abstract

While Large Language Models (LLMs) are frequently cited for their sophisticated pragmatic reasoning (Wei et al., 2022; Bubeck et al., 2023), recent progress in sarcasm detection increasingly relies on synthetic benchmarks (Li et al., 2025; Anonymous, 2025). This study exposes a catastrophic generalization gap in this paradigm: we observe that models achieve near-perfect accuracy on synthetic data but collapse to random guessing on organic human speech. By triangulating hidden state geometry, entropy analysis, and causal interventions, we demonstrate that this disparity stems from shortcut learning (Geirhos et al., 2020)—models exploit the low-entropy statistical signatures of generated text while remaining "semantically blind" to the pragmatic cues essential for irony. Our findings indicate that high performance on synthetic leaderboards reflects forensic pattern matching rather than the genuine linguistic intelligence assumed in prior work, creating a statistical mirage of competence.

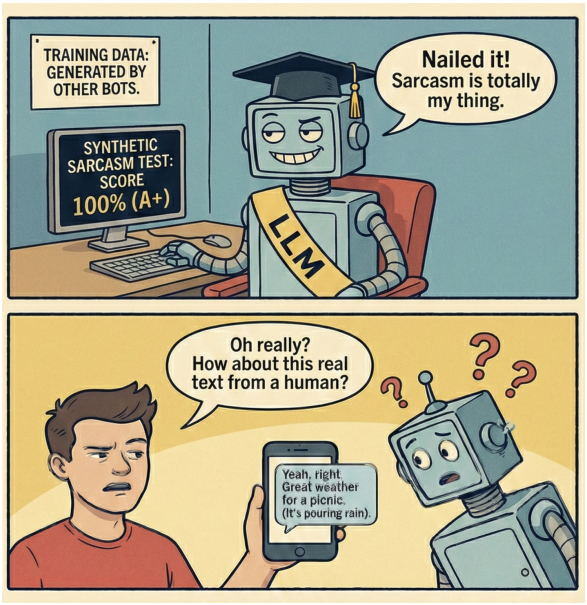

Figure 1: Artifact Detection ≠ Sarcasm Detection: A visual summary of why synthetic performance metrics are a poor proxy for genuine linguistic intelligence in real-world scenarios.

1 Introduction

Sarcasm detection stands as one of the most challenging tasks in Natural Language Processing (NLP), requiring models to look beyond literal semantics to identify incongruity, tonal nuance, and context dependence (Joshi et al., 2017). Unlike standard sentiment analysis, sarcasm often employs positive lexical cues to convey negative intent, a phenomenon that challenges even human annotators (Wallace et al., 2014; Filatova, 2012). Consequently, the ability to decipher such pragmatic intent is frequently cited as a litmus test for genuine machine intelligence (Wei et al., 2022; Bubeck et al., 2023).

Recent evaluations suggest that Large Language Models (LLMs) have attained near-human competence in this domain. However, we propose that this proficiency may be deceptive. The core question intersects with the broader debate on whether LLMs perform robust reasoning or merely exhibit **shortcut learning**—a failure mode where models rely on superficial heuristics rather than semantic understanding (Geirhos et al., 2020). This phenomenon, often termed the **"Clever Hans" effect** (Pacchiardi et al., 2024), occurs when models exploit **spurious correlations** (Sagawa* et al., 2020) or **dataset artifacts** (Gururangan et al., 2018)—such as specific lexical patterns or low-entropy signatures—to solve a task without learning the underlying task reasoning.

In the context of sarcasm, this risk is acute. As LLMs are increasingly pre-trained on synthetic corpora, they may learn to recognize the statistical "texture" of machine-generated text rather than the pragmatic incongruity of sarcasm. If true, high

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29169–29199

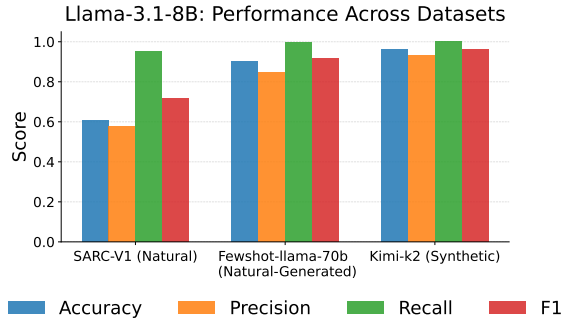

Figure 2: Performance follows a distinct hierarchy: models achieve near-perfect separability on synthetic data (right) and remain high on natural-generated text (middle), yet collapse on organic speech (left).

leaderboard performance represents **artifact exploitation** rather than linguistic comprehension. To investigate this, we introduce a rigorous experimental framework contrasting model behavior across synthetic datasets, organic baselines, and a natural-generated bridge dataset (Figure 3).

By triangulating evidence from hidden state geometry, membership inference, and causal interventions, we expose a fundamental disconnect. We observe a distinct performance hierarchy: while models achieve near-perfect accuracy on synthetic data, they exhibit intermediate performance on natural-generated samples and collapse to random guessing on organic speech (Figure 2). Our findings reveal that current models do not detect sarcasm; they detect the generator, effectively solving a forensic classification task under the guise of pragmatic reasoning.

Our key contributions are as follows:

Exposing the Synthetic Mirage: We reveal a catastrophic generalization gap where models achieve near-perfect accuracy on synthetic sarcasm but collapse to random guessing on organic speech (Section 4.1.1). Hidden state visualizations confirm this, showing synthetic sarcasm forms separable "clean islands" in the residual stream, whereas organic sarcasm remains deeply entangled, proving models fail to learn a robust representation of ironic intent (Section 4.2.1).

Mechanistic Proof of Artifact Exploitation: Using Min-K%++ (Section 4.2.3) and layer-wise ablation (Section 4.2.6), we demonstrate that models detect synthetic sarcasm via shortcut learning based on low-entropy generative signatures rather than content.

Confirming Semantic Blindness: Through token occlusion (Section 4.2.7) and LIME interventions (Section 4.2.4), we show that models remain confident even when pivotal semantic anchors are removed. This confirms a reliance on distributed surface-level heuristics, rendering models effectively "blind" to the semantic incongruity essential for genuine sarcasm detection.

2 Methodology

To deconstruct the mechanistic basis of sarcasm detection, we employ a multi-stage framework that incorporates behavioral performance, internal representation geometry, and causal faithfulness. Our approach explicitly contrasts model behavior on synthetic versus organic distributions to isolate shortcut learning.

2.1 Data Preparation Strategy

To scrutinize the "generalization gap", we curate data streams into two distinct categories based on their generative source: Organic (human-authored) and Synthetic (LLM-generated). For the organic text, this research uses content available in the sarcasm-labeled datasets SARC V1 (Oraby et al., 2016), SARC V2 (which is further bifurcated into three datasets containing general sarcasm (GEN), hyperbole (HYP), and rhetorical questions (RQ)) (Lukin and Walker, 2013), and News Headlines (Misra and Grover, 2021; Misra and Arora, 2023) for the organic text. For the synthetic text, two LLMs—Llama-3.3-70b-versatile (Grattafiori et al., 2024) and moonshotai-kimi-k2-Instruct-0905 (Team et al., 2026)—were prompted to generate sarcastic text. To bridge the gap between these regimes, we additionally curate a hybrid "Natural-Generated (Few-Shot)" dataset, where the model is prompted with organic exemplars from SARC V1 to mimic human stylistic features and induce linguistic diversity. The data was preprocessed to ensure better quality. We consider only those sentences which have number of words between 5 and 50. However, hashtags, user mentions, emoticons, abbreviations and similar elements were retained, as their presence in natural language provides enhanced insight. Further details about the datasets can be found in Table 1.

2.2 Dataset Fidelity and Inter-Annotator Agreement

To formally validate the fidelity of the generated text and characterise the complexity of the synthetic sarcasm, we conducted an external inter-annotator agreement study. We randomly sampled

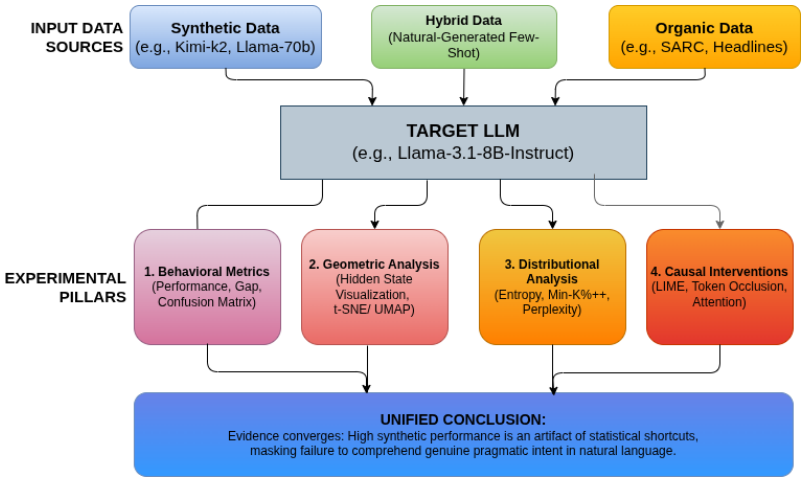

Figure 3: Methodological Overview

Dataset	Size (N)	# Sarcastic	# Non-Sarcastic	Avg. Length	Type-Token Ratio	Distribution
Llama-70b_synthetic	2579	1440	1139	26.4	0.05	Synthetic
Kimi-k2_synthetic	2979	1566	1413	21.9	0.08	Synthetic
SARC-V1	1995	998	997	59.9	0.09	Organic
SARC-V2_HYP	1164	582	582	56.4	0.12	Organic
SARC-V2_GEN	6520	3260	3260	44.9	0.06	Organic
SARC-V2_RQ	1702	851	851	67.8	0.09	Organic
News Headlines	28619	13634	14985	10.5	0.09	Organic
Natural-Generated	1118	584	534	28.3	0.12	Mixed/Hybrid

Table 1: Corpora Statistics

a subset of 50 utterances distributed across the three generated data regimes: Llama-3.3-70b (17), Kimi-k2 (17), and the hybrid Natural-Generated (16) dataset. Two independent human evaluators of different genders (who are not authors of this paper) were tasked with blindly labelling each utterance and were compensated commensurate with their efforts. These human annotations were evaluated against each other and against the "original" labels in the dataset.

To quantify the consensus between the two human raters, we computed Cohen's kappa (Cohen, 1960). The overall Cohen's kappa score for the annotated subset achieved 0.8397 (inter-annotator), 0.9199 (original vs annotator 1) and 0.8397 (original vs annotator 2) which corresponds to "Almost Perfect Agreement" according to the interpretation guidelines established by Landis and Koch (1977) (Landis and Koch, 1977). This exceptionally high inter-annotator agreement supports our hypothesis that standard zero-shot LLMs tend to generate structurally explicit, "easy" sarcasm, lacking the pragmatic ambiguity and nuance characteristic of organic human speech.

2.3 Binary Classification Formulation

We formulate sarcasm detection as a zero-shot conditional generation task. Given an instruction I and a target sentence S, the model learns a probability distribution $P_\theta(y|I, S)$. We constrain the output space—by using customized prompts for each model—to a binary token set $V_{target} = \{\text{“}Yes\text{”}, \text{“}No\text{”}\}$. The classification decision is derived from the normalized logits of these target tokens at the final decoding step T:

$$\hat{y} = \underset{v \in V_{target}}{\text{argmax}} P_\theta(v|h_L)$$

We explicitly analyze the confusion matrix structure rather than accuracy and other scores alone. Specifically, we monitor the false positive rate (FPR) on organic non-sarcastic samples to quantify "hallucination", hypothesizing that models biased by synthetic artifacts will over-attribute sarcasm in high-entropy natural text.

2.4 Hidden State Geometry

To verify if the model constructs a generalized concept of sarcasm, we analyze the geometry of the residual stream. Let S be an input sequence of

length T, consisting of tokens $\{w_1, w_2, \ldots, w_T\}$. Let $h_{l,t} \in \mathbb{R}^d$ denote the hidden state vector corresponding to token t at layer l.

Instead of relying on the final token, which maybe biased by immediate local context, we compute the mean-pooled representation e_l for the entire sequence at specific network depths (first, middle and final layers):

$$e_l = \frac{1}{T} \sum_{t=1}^{T} h_{l,t}$$

This operation aggregates the distributed semantic information across the sequence into a single fixed-size vector. We then employ non-linear dimensionality reduction techniques, specifically t-SNE (van der Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018), to project these embeddings e_l into a visualizable two-dimensional space. While we recognize that such projections do not provide a rigorous quantitative proof of class separability, they serve as a valuable qualitative tool for initial data exploration. A preliminary visual inspection provides an intuitive hint regarding the underlying representation space: synthetic data tends to map into visually distinct clusters (or "clean islands"), whereas the representations of organic text appear highly entangled. This stark visual contrast offers early qualitative support for our hypothesis that synthetic sarcasm relies on simplistic structural artifacts, making it artificially easier for the model to classify.

2.5 Mechanistic Interpretability Framework

We employ a four-pronged interpretability suite to isolate the mechanism of prediction.

2.5.1 Subsampling and Bootstrapping Protocol

To ensure statistical parity and mitigate vocabulary scaling effects (Heap's Law) across heterogeneous datasets, we employ a stratified subsampling strategy for all entropy-based metrics (Perplexity and Min-K%++). We fix the sample size $N = 1100$ (corresponding to the cardinality of the smallest dataset in our corpus). For each experimental trial $i \in \{1, \ldots, n\}$, we randomly draw N examples without replacement from each dataset D_j and compute the score distributions. The final visualizations and metrics represent the mean distribution over these $n = 100$ trials. This guarantees that observed differences in entropy are attributable to the generative nature of the text rather than dataset magnitude or sampling bias.

2.5.2 Perplexity and Entropy Analysis

We utilize Perplexity (Jelinek et al., 1977) as a proxy for the statistical "surprise" of the input text. Using the bootstrapping protocol defined above, we compute the Kernel Density Estimation (KDE) of the PPL distributions.

2.5.3 Membership Inference via Min-K%++

To formally detect if the model is identifying "synthetic artifacts" (i.e., identifying that the text comes from a similar distribution to its training data), we adapt the Min-K%++ metric (Zhang et al., 2025). Unlike raw perplexity, Min-K%++ accounts for the token-specific variance.

2.5.4 Causal Attribution (LIME & Attention)

To determine if the model attends to semantic content, we use Local Interpretable Model-agnostic Explanations (LIME). We approximate the complex non-linear model with a locally linear surrogate around specific inputs, generating 2,000 perturbations to identify feature importance. We cross-reference this with BertViz visualisations (Vig, 2019) of attention heads to check for semantic anchoring.

2.5.5 Robustness Analysis (Ablation & Occlusion)

Finally, we test the structural robustness of the decision mechanism:

1. **Layer-wise Ablation**: We evaluate the contribution of individual transformer blocks by systematically bypassing them during inference. For each layer l, we temporarily replace the standard decoder block with an Identity Mechanism ($f(h) = h$), effectively removing that layer's transformation while preserving the residual stream's flow. We measure the degradation in accuracy (Δ_{acc}) relative to the baseline.

2. **Token Occlusion**: We iteratively mask content words in the input. We define *Semantic Blindness* as the condition where masking key semantic anchors results in negligible change to the output confidence ($\Delta_{conf} \approx 0$).

3 Experimental Setup

3.1 LLMs used

We employ Llama-3.3-70B-Versatile (70B parameters) (Grattafiori et al., 2024) and MoonshotAI-Kimi-K2-Instruct-0905 (32B parameters) (Team et al., 2026) to generate the synthetic datasets via the Groq API. Larger instruction-tuned models were preferred because, when explicitly constrained to produce only a sentence, they reliably adhere to the output format without introducing extraneous explanations or meta-text, thereby minimizing the need for post-generation denoising.

For the analysis, we use four decoder-only models, namely, Llama-3.1-8B-Instruct (8B parameters), gemma-2-9b-it (9B parameters) (Team et al., 2024), Qwen2.5-14B-Instruct (14B parameters) (Team, 2024) and Phi-3-medium-4k-instruct (14B parameters) (Abdin et al., 2024). We use instruction-tuned variants to frame sarcasm detection as a binary classification task by constraining the output space to the tokens "Yes" and "No". Instruction tuning facilitates reliable compliance with such strict output constraints via model-specific prompting. Finally, we prioritize relatively small models due to their greater deployability in low-resource settings and their feasibility for offline use without reliance on proprietary APIs.

3.2 Prompts Used

To constrain model outputs to binary classifications ("Yes"/"No"), we employed the specific prompting strategies detailed in Appendix Table 8.

Synthetic datasets were generated using the templates in Appendix Table 9. For each instance, we sampled a topic uniformly from the diverse domains listed in Appendix Table 10 and assigned a random binary label, instructing the model to produce a single conditioned English sentence.

Additionally, to bridge the gap between synthetic and organic distributions, we curated a **Natural-Generated (Few-Shot)** dataset using Llama-3.3-70B-Versatile. This process utilized three randomly sampled organic examples from SARC V1 as few-shot exemplars within the templates in Appendix Table 11. This approach aimed to elicit more linguistically natural phrasing and diverse syntactic structures while maintaining controlled label validity.

4 Results

4.1 The Symptom: Catastrophic Performance Collapse on Organic Data

4.1.1 Classification Performance

Model	Dataset	Acc.	Prec.	Rec.	F1
ORGANIC					
Llama-3.1-8B	SARC-V1	0.606	0.578	0.950	0.719
Qwen2.5-14B	SARC-V1	0.669	0.677	0.717	0.696
Gemma-2-9b-it	SARC-V1	0.560	0.547	0.983	0.703
Phi-3-medium-4k	SARC-V1	0.614	0.586	0.927	0.718
FULLY SYNTHETIC					
Llama-3.1-8B	Kimi-k2	0.961	0.932	1.0	0.964
Qwen2.5-14B	Kimi-k2	0.992	0.985	1.0	0.992
Gemma-2-9b-it	Kimi-k2	0.938	0.895	1.0	0.945
Phi-3-medium-4k	Kimi-k2	0.973	0.951	1.0	0.975
NATURAL-GENERATED					
Llama-3.1-8B	Fewshot-llama-70b	0.904	0.846	0.998	0.915
Qwen2.5-14B	Fewshot-llama-70b	0.973	0.954	0.996	0.974
Gemma-2-9b-it	Fewshot-llama-70b	0.848	0.787	0.998	0.880
Phi-3-medium-4k	Fewshot-llama-70b	0.881	0.814	1.0	0.897

Table 2: Performance comparison across datasets.

Llama-3.1-8B-Instruct (SARC-V1, Natural)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	166	597
Sarcastic	43	817

Llama-3.1-8B-Instruct (Kimi-k2-instruct-0905, Synthetic)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	1299	114
Sarcastic	0	1566

Llama-3.1-8B-Instruct (Fewshot-llama-70b, Natural-Synthetic)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	428	106
Sarcastic	1	583

Table 3: Confusion matrices for sarcasm detection

Table 2 presents the classification performance across selected models (Additional detailed models analysis can be found in Appendix C). A stark dichotomy emerges: on fully synthetic datasets (Kimi-k2, Llama-70b), all models achieve near-perfect metrics (Accuracy/F1 $\approx$ 1.0), indicating trivial separability.

In contrast, performance collapses on organic datasets (SARC, News Headlines). While recall for the sarcastic class remains high, precision and accuracy degrade significantly. This specific failure mode—high recall paired with low precision—indicates a systematic tendency to over-

predict sarcasm, resulting in excessive false positives rather than robust discrimination. As shown in the confusion matrices (Table 3), models on natural text frequently hallucinate sarcasm in neutral sentences, whereas synthetic data yields perfect diagonal separation.

The **Natural-Generated (Few-Shot)** dataset occupies an intermediate regime: while retaining the high recall of synthetic data, it suffers from a slight drop in precision, confirming that even conditioned generation cannot fully replicate the entanglement of fully organic distributions. Detailed class-wise metrics are provided in Appendix C.

4.2 The Diagnosis: Unmasking Statistical Shortcuts via Probing

4.2.1 Embedding Visualization

Visualizing Llama-3.1-8B-Instruct's internal representations via t-SNE and UMAP provides an intuitive first look at a fundamental mechanistic divergence between synthetic (Kimi-k2) and organic (SARC-V1) processing. Although, as noted in Section 2.4, such dimensionality reduction methods do not serve as rigorous proofs of class separability, they offer valuable qualitative hints regarding the model's internal data structuring. In these exploratory projections, synthetic data forms visually distinct clusters as early as the first layer, suggesting the model exploits surface-level artifacts that require no deep processing (Figure 4). In contrast, organic representations appear inextricably entangled across all layers (Figure 5), with even the final layer failing to exhibit clear separation of sarcastic intent. This geometric intuition aligns with the conclusion that high synthetic performance stems from pre-existing representational biases rather than the robust pragmatic understanding required for natural language.

4.2.2 Perplexity Analysis

Figure 6 presents the distribution of averaged log-perplexity values for sarcastic and non-sarcastic text across natural, synthetic, and hybrid datasets under Llama-3.1-8B-Instruct. Perplexity values are computed following the subsampling protocol described in Section 2.5.1. Fully synthetic datasets, such as Kimi-K2 and Llama-70B-generated data, exhibit tightly concentrated, low-perplexity distributions with clear separation between sarcastic and non-sarcastic classes, indicating highly regular and predictable surface patterns. In contrast, natural datasets show substantially broader distributions and pronounced overlap between classes, reflecting increased linguistic variability and reduced model confidence. Hybrid datasets generated using few-shot prompting occupy an intermediate regime, with perplexity profiles closer to synthetic data but with increased variance and reduced class separation. These trends are consistent across subsamples and suggest that low perplexity appears to act as a shortcut signal that aligns with synthetic labels but fails to support robust discrimination in natural text. Additional dataset-wise perplexity plots and analyses are provided in Appendix E.

4.2.3 Membership Inference - Min-K%++

Figure 7 presents the Min-K%++ distributions for Llama-3.1-8B-Instruct. Serving as a proxy for distributional familiarity, Min-K%++ reveals a stark dichotomy: synthetic datasets yield significantly higher scores, indicating the model perceives them as highly familiar or formulaic. Conversely, organic texts act as relative outliers. This gap implies that high performance on synthetic data relies on latent generative artifacts absent in natural language.

Crucially, while absolute Min-K%++ values (scale, width, and peak location) are heavily model-dependent and cannot be directly compared across architectures (Appendix F), a fundamental trend persists. Across all evaluators, distinct dataset-dependent differences in distribution shape remain consistently visible. This geometric divergence confirms that models structurally differentiate between organic and synthetic manifolds, even if the absolute direction of the shift varies by model.

4.2.4 LIME Visualization

We utilize LIME to map local feature importance (Figure 8). Synthetic examples (as in Appendix G) exhibit diffuse, low-magnitude attributions lacking decisive semantic anchors, suggesting reliance on broad generative artifacts rather than causal signals.

Conversely, natural examples reveal a misleading focus on superficial lexical cues. The model largely ignores essential topic-bearing terms like *"Archbishop"* and *"Christianity"*, instead over-attributing significance to neutral conversational markers (e.g., "yes", "what"). This asymmetry confirms that the model has internalized brittle lexical heuristics rather than grounded pragmatic

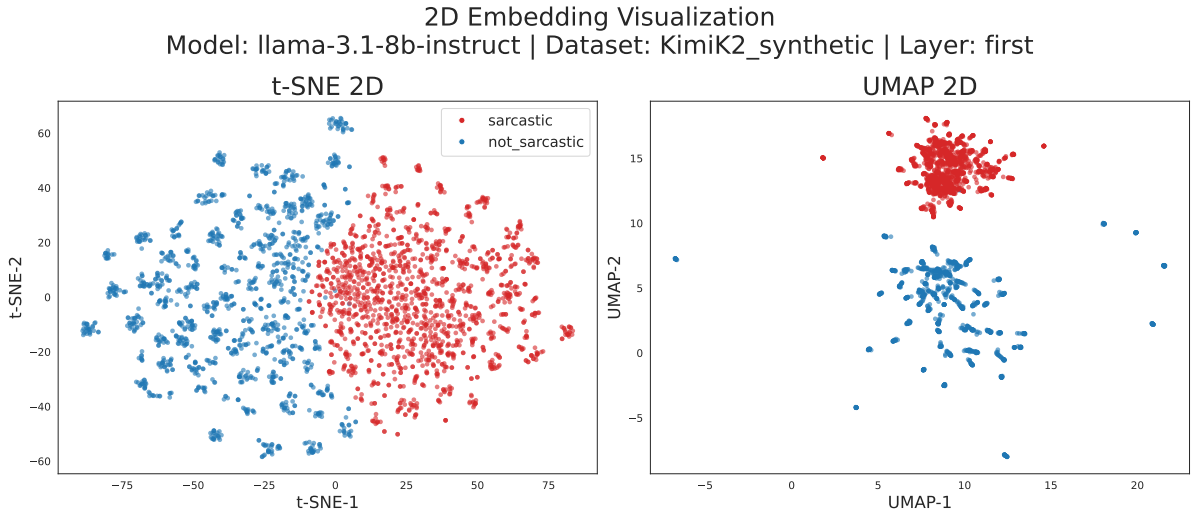

Figure 4: 2D projection of last layer embedding for Llama-3.1-8B-Instruct on Kimi-k2 (left: t-SNE, right: UMAP).

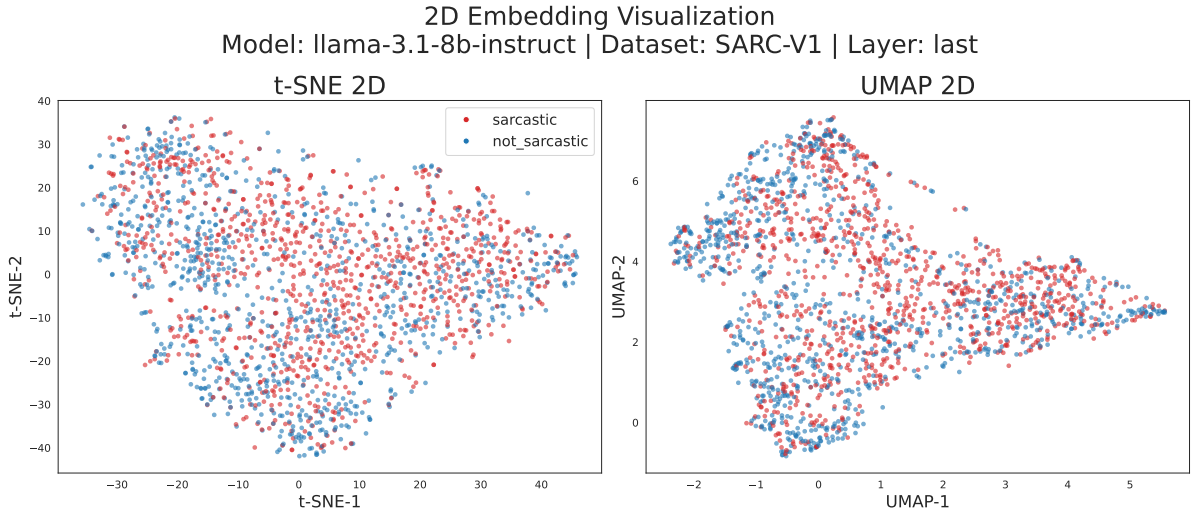

Figure 5: 2D projection of last layer embedding for Llama-3.1-8B-Instruct on SARC-V1 (left: t-SNE, right: UMAP).

representations, driving the high rate of false-positive hallucinations on organic text.

4.2.5 Attention Visualization

Figure 9 illustrates the attention dynamics for a representative sarcastic instance where the irony hinges on the semantic tension between "Archbishop" and "Christianity". From a linguistic perspective, the sarcasm is non-local; the token "Christianity" acts as a sarcastic anchor only when resolved against the specific context of "Archbishop." Therefore, a model engaging in genuine pragmatic reasoning should exhibit strong attention heads directing focus from "Christianity" back to its incongruous antecedent.

However, the visualization reveals a pathological attention landscape. First, the maximal attention weight across all layers and heads is consistently directed toward the <|begin_of_text|> token, acting as a global "attention sink" that overshadows content tokens. Second, the specific semantic dependency between "Christianity" and "Archbishop" is negligible. While faintly visible in early layers, this connection does not consolidate with depth; instead, it decays. By the final layers, the residual attention (excluding the start token) is strictly localized to immediate neighbors, effectively severing the long-range context necessary to decode the irony. This dual phenomenon—dominance of the start token and progressive loss of context—suggests the model re-

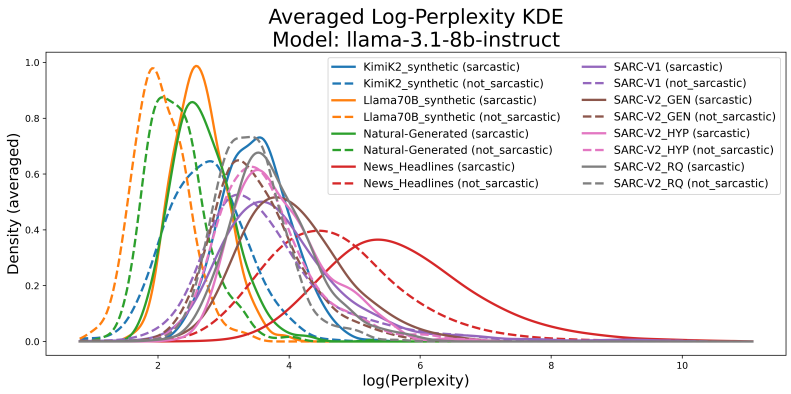

Figure 6: Log-perplexity distributions for sarcastic and non-sarcastic text across datasets on Llama-3.1-8b-Instruct model.

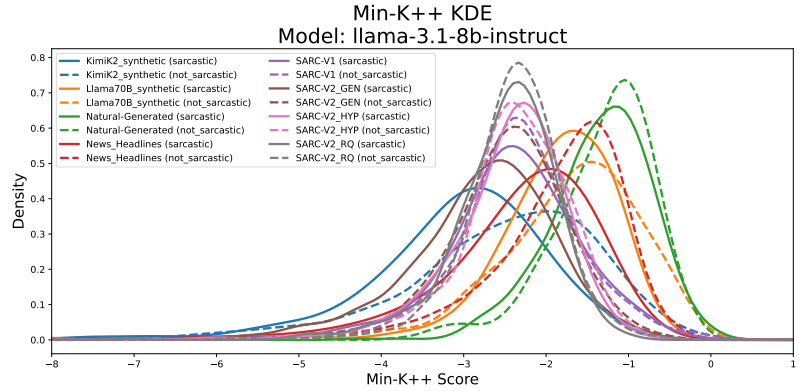

Figure 7: Min-K%++ score distributions for sarcastic and non-sarcastic text across datasets on Llama-3.1-8b-Instruct model.

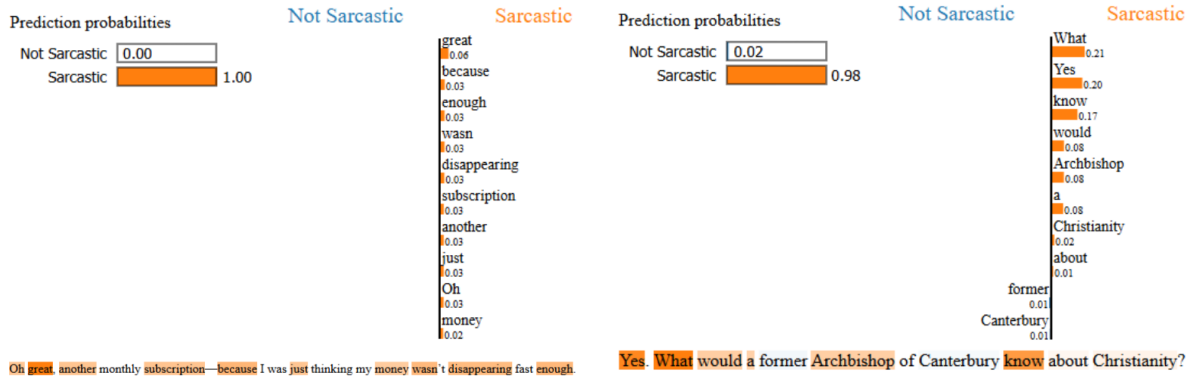

Figure 8: LIME visualization for a sarcastic sentence from Kimi-k2 (left) and SARC-V1 (right) dataset (model: llama-3.1-8b-instruct).

lies on global artifacts and localized cues rather than relational semantic structure. Additional examples are provided in Appendix H.

4.2.6 Layer Ablation

Figure 23 in the appendix I presents the layer ablation results for LLaMA-3.1-8B on a 100-sentence subsample of SARC-V1. Across layers, ablating individual transformer blocks produces only modest changes in accuracy, with the maximum observed drop not exceeding approximately 0.15. This indicates that sarcasm detection does not depend critically on any single layer.

In this specific instance, ablating several early and early-middle layers results in comparatively larger accuracy drops than ablating later layers.

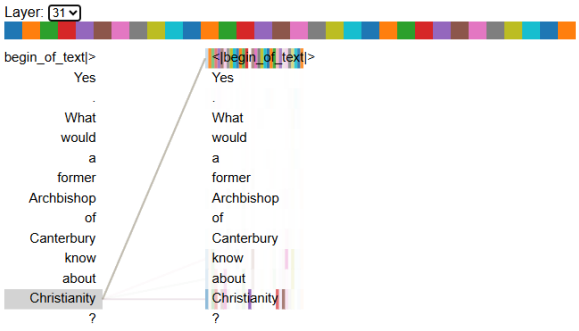

Figure 9: Attention distribution for the last transformer layer for a sarcastic example. (model: llama-3.1-8b-instruct)

However, this pattern is neither uniform nor monotonic: multiple mid and late layers exhibit near-zero impact, and ablating some layers leads to negative drops, indicating a slight improvement in accuracy when those layers are removed.

Importantly, this behavior is not consistent across sentences or subsamples. While many cases exhibit greater sensitivity to early-layer removal, others do not, and no stable layer-specific hierarchy emerges. The presence of both weak positive and negative effects suggests that sarcasm-relevant signals are distributed and fragile rather than localized or progressively refined with depth.

Taken together, the small magnitude of accuracy degradation, the absence of a consistent depth-wise trend, and the occasional improvement under ablation argue against a dedicated or hierarchical mechanism for sarcasm detection within the model architecture. Instead, these results are consistent with a shallow, distributed decision signal that is not substantially transformed or corrected by later layers.

4.2.7 Token Occlusion

We assess causal sensitivity via token-level occlusion (Section 2.5.5). As illustrated in Figure 26, synthetic examples yield almost uniformly flat attribution maps, indicating that predictions are not driven by specific tokens.

In contrast, natural examples reveal a critical failure mode we term **"Semantic Blindness"**. While occlusion induces minor confidence fluctuations, it rarely alters the decision boundary. Crucially, even when pivotal semantic anchors like *"Archbishop"* or *"Christianity"*—which provide the essential context for the irony—are removed, the model maintains high confidence in the sarcastic class. This decoupling suggests that predictions rely on shallow, distributed heuristics (e.g., sentence texture) rather than the resolution of specific semantic incongruities.

5 Conclusion

This study investigates the mechanistic basis of sarcasm detection in Large Language Models. We observe a pronounced generalization gap: models demonstrate near-perfect accuracy on synthetic data but show limited capability on organic human speech. Our geometric and causal analyses reveal that this performance on synthetic domains is largely driven by artifact learning, where models exploit low-entropy statistical signatures rather than resolving semantic incongruity. Consequently, high accuracy in controlled environments does not necessarily translate to robust pragmatic understanding. A more detailed discussion on the results can be found in the Appendix K. These findings highlight that current LLMs primarily treat sarcasm detection as a distributional recognition task, underscoring the necessity of distinguishing between statistical pattern matching and genuine semantic reasoning in future research.

Limitations

While our mechanistic framework provides evidence of artifact learning, the study has some limitations. The analysis is restricted to text-only data, omitting paralinguistic and multimodal cues essential to human sarcasm; it is further constrained to sub-20B open-weights models due to the need for internal access, preventing verification on larger or closed-source systems. Although we demonstrate reliance on low-entropy statistical artifacts, we do not identify the precise linguistic features underlying these shortcuts. Finally, all datasets are Anglocentric, limiting the generalizability of our conclusions across cultures and languages.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 66 others.

2024. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft.

Anonymous. 2025. Deceptive humor: A synthetic multilingual benchmark dataset for bridging fabricated claims with humorous content. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20(1):37–46.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 392–398, Istanbul, Turkey. European Language Resources Association (ELRA).

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

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

Jithendra Katta, Manikanta Allanki, and Nikhil Reddy Kodumuru. 2025. Understanding sarcasm detection through mechanistic interpretability. In *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, pages 990–995.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Zhu Li, Yuqing Zhang, Xiyuan Gao, Shekhar Nayak, and Matt Coler. 2025. Leveraging Large Language Models for Sarcastic Speech Annotation in Sarcasm Detection. In *Interspeech 2025*, pages 3973–3977.

Stephanie Lukin and Marilyn Walker. 2013. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40, Atlanta, Georgia. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Rishabh Misra and Prahal Arora. 2023. Sarcasm detection using news headlines dataset. *AI Open*, 4:13–18.

Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*. Independently Published.

Jun Niu, Peng Liu, Xiaoyan Zhu, Kuo Shen, Yuecong Wang, Haotian Chi, Yulong Shen, Xiaohong Jiang, Jianfeng Ma, and Yuqing Zhang. 2024. A survey on membership inference attacks and defenses in machine learning. *Journal of Information and Intelligence*, 2(5):404–454.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.

Lorenzo Pacchiardi, Marko Tesic, Lucy G. Cheke, and José Hernández-Orallo. 2024. Leaving the barn door open for clever hans: Simple features predict llm benchmark answers. *Preprint*, arXiv:2410.11672.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard,

Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, and 181 others. 2026. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 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, Baltimore, Maryland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*.

A Related Work

A.1 Historical Evolution of Sarcasm Detection

Early sarcasm identification relied on manually crafted features, such as interjections, punctuation, and sentiment incongruity. The emergence of Transformer-based models (BERT, RoBERTa) marked a paradigm shift by capturing long-range dependencies. Despite these improvements, the dependability of text-only models remains uncertain in real-world contexts, often necessitating multimodal cues (audio/visual) to clarify ambiguities—a limitation not addressed by current text-only synthetic benchmarks.

A.2 Sarcasm Detection and the Synthetic Shift

Sarcasm detection has evolved from rule-based linguistic indicators to Large Language Models (LLMs) employing "LLM-as-a-judge" frameworks. To address data scarcity, recent works have utilized models like GPT-4 and Llama-3 to generate large-scale corpora for sarcasm (Li et al., 2025) and "deceptive humour" (Anonymous, 2025). However, relying on synthetic data introduces the risk of "unfaithfulness," where models learn simplistic heuristics rather than sophisticated pragmatic components. Our research challenges the validity of evaluating models on these synthetic distributions, arguing that high performance often reflects artifact recognition rather than genuine semantic understanding.

A.3 General Interpretability Methods

Beyond mechanistic approaches, general interpretability frameworks like LIME and SHAP have been widely utilized to elucidate model predictions by assigning importance scores to input features. However, these methods often fail to distinguish between a model that uses a feature because it understands the meaning versus one that uses it as a statistical shortcut, necessitating the deeper causal interventions employed in this study.

A.4 Mechanistic Interpretability & Shortcut Learning

While general interpretability methods like SHAP (Lundberg and Lee, 2017) treat models as black boxes, our work focuses on "shortcut learning" or the "Clever Hans" effect (Pacchiardi et al., 2024), where models rely on spurious correlations. This study closely scrutinizes prior work by (Katta et al., 2025), which correlated attention mechanisms with sarcastic cues in BERT-Large but accepted the model's performance at face value. We diverge by explicitly testing the hypothesis that the attention signal observed in previous studies is a statistical artifact resulting from synthetic generation, rather than a sign of linguistic reasoning.

A.5 Entropy as an Artifact Detector

Information-theoretic methods typically use entropy to measure text complexity or detect pretraining data via Membership Inference Attacks (MIA) (Niu et al., 2024). We repurpose the Min-K%++ metric not for detecting training membership, but for identifying "distributional membership." By quantifying the likelihood of tokens, we utilize Min-K%++ to detect the low-entropy signature of synthetic text, presenting a novel application of this metric to correlate low perplexity with false-positive sarcasm detection.

B Use of Generative AI

We acknowledge the use of LLMs to assist with proofreading and polishing the English quality of the manuscript. The original content and ideas were generated entirely by the authors. The tools were also employed for minor syntax checking and code support within the experimental workflow.

C Classification Analysis

C.1 Aggregate Metrics Across Models and Datasets

Table 4 reports accuracy, precision, recall, and F1 scores for all evaluated models across organic, synthetic, and hybrid datasets. Two consistent trends emerge.

First, fully synthetic datasets (Kimi-k2, Llama-70B) yield uniformly high performance across all models, with accuracy and F1 scores approaching 1.0. Precision and recall are both near-perfect, indicating trivial separability between sarcastic and non-sarcastic classes. This consistency across models suggests that the separability signal is dataset-driven rather than model-specific.

Second, natural datasets (SARC-V1, SARC-V2 variants, News Headlines) exhibit a markedly different pattern. While recall for the sarcastic class remains high across models, precision drops substantially, leading to moderate F1 scores despite

Model	Dataset	Acc.	Prec.	Rec.	F1
ORGANIC					
Llama-3.1-8B	SARC-V1	0.606	0.578	0.950	0.719
Qwen2.5-14B	SARC-V1	0.669	0.677	0.717	0.696
Gemma-2-9b-it	SARC-V1	0.560	0.547	0.983	0.703
Phi-3-medium-4k	SARC-V1	0.614	0.586	0.927	0.718
Llama-3.1-8B	GEN-SARC-V2	0.643	0.598	0.970	0.740
Qwen2.5-14B	GEN-SARC-V2	0.766	0.753	0.823	0.787
Gemma-2-9b-it	GEN-SARC-V2	0.607	0.572	0.985	0.724
Phi-3-medium-4k	GEN-SARC-V2	0.649	0.602	0.973	0.744
Llama-3.1-8B	HYP-SARC-V2	0.561	0.536	0.994	0.696
Qwen2.5-14B	HYP-SARC-V2	0.720	0.655	0.946	0.774
Gemma-2-9b-it	HYP-SARC-V2	0.547	0.528	0.998	0.690
Phi-3-medium-4k	HYP-SARC-V2	0.598	0.558	0.992	0.714
Llama-3.1-8B	RQ-SARC-V2	0.620	0.572	0.988	0.725
Qwen2.5-14B	RQ-SARC-V2	0.791	0.745	0.895	0.813
Gemma-2-9b-it	RQ-SARC-V2	0.577	0.545	0.992	0.703
Phi-3-medium-4k	RQ-SARC-V2	0.625	0.575	0.986	0.727
Llama-3.1-8B	News-Headlines	0.692	0.626	0.884	0.733
Qwen2.5-14B	News-Headlines	0.687	0.747	0.523	0.615
Gemma-2-9b-it	News-Headlines	0.641	0.575	0.949	0.717
Phi-3-medium-4k	News-Headlines	0.655	0.591	0.902	0.714
FULLY SYNTHETIC					
Llama-3.1-8B	Kimi-k2	0.961	0.932	1.0	0.964
Qwen2.5-14B	Kimi-k2	0.992	0.985	1.0	0.992
Gemma-2-9b-it	Kimi-k2	0.938	0.895	1.0	0.945
Phi-3-medium-4k	Kimi-k2	0.973	0.951	1.0	0.975
Llama-3.1-8B	Llama-70b	0.993	0.988	1.0	0.994
Qwen2.5-14B	Llama-70b	0.998	0.998	0.999	0.998
Gemma-2-9b-it	Llama-70b	0.987	0.978	0.998	0.988
Phi-3-medium-4k	Llama-70b	0.996	0.994	1.0	0.997
NATURAL GENERATED					
Llama-3.1-8B	Fewshot-llama-70b	0.904	0.846	0.998	0.915
Qwen2.5-14B	Fewshot-llama-70b	0.973	0.954	0.996	0.974
Gemma-2-9b-it	Fewshot-llama-70b	0.848	0.787	0.998	0.880
Phi-3-medium-4k	Fewshot-llama-70b	0.881	0.814	1.0	0.897

Table 4: Classification metrics for sarcasm detection across datasets.

Gemma-2-9b-it (SARC-V1, Natural)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	64	699
Sarcastic	14	846

Phi-3-medium-4k-instruct (SARC-V1, Natural)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	200	563
Sarcastic	62	798

Qwen2.5-14b-Instruct (SARC-V1, Natural)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	469	294
Sarcastic	243	617

Table 5: Confusion matrices for sarcasm detection (SARC-V1 dataset)

acceptable accuracy. This imbalance indicates that correct detection of sarcastic instances is achieved largely at the cost of misclassifying non-sarcastic inputs.

Hybrid Natural-Generated (Few-Shot) datasets occupy an intermediate regime. Performance is higher than in fully organic data, but worse than in fully synthetic data, particularly in precision. This suggests that few-shot prompting partially transfers synthetic regularities while reintroducing some of the ambiguity present in human-authored text.

C.2 Confusion Matrix Structure and Error Asymmetry

To contextualize these aggregate metrics, Tables 5, 6 & 7 report confusion matrices for representative models and datasets. These matrices reveal an asymmetric error profile.

On natural datasets, the dominant error mode is a large number of false positives, where non-sarcastic sentences are predicted as sarcastic. True negatives are comparatively rare, while true positives remain abundant. This structure explains the high recall-low precision regime observed in Table 4. Importantly, false negatives are comparatively infrequent, indicating that the models adopt a biased decision boundary favoring sarcasm predictions. In contrast, synthetic datasets display a strongly diagonal confusion-matrix structure. While a small number of false positives are present, false negatives are effectively absent and misclassifications remain rare relative to natural datasets.

For few-shot generated data, confusion matrices again show elevated false positives relative to synthetic data, though substantially fewer than in fully natural datasets. This intermediate structure mirrors the corresponding metric trends and reinforces the interpretation that few-shot generation partially, but not fully recovers the complexity of human sarcasm.

C.3 Interpretation

Taken together, the combined metric and confusion-matrix analysis demonstrates that strong classification performance on synthetic data does not translate to robust discrimination on natural language. Aggregate scores alone obscure a consistent structural failure: models systematically hallucinate sarcasm in high-entropy human text. Confusion matrices make this failure explicit by revealing that errors are not random but heavily skewed toward false positives.

These results support the broader claim of the paper that sarcasm detection in current LLMs is

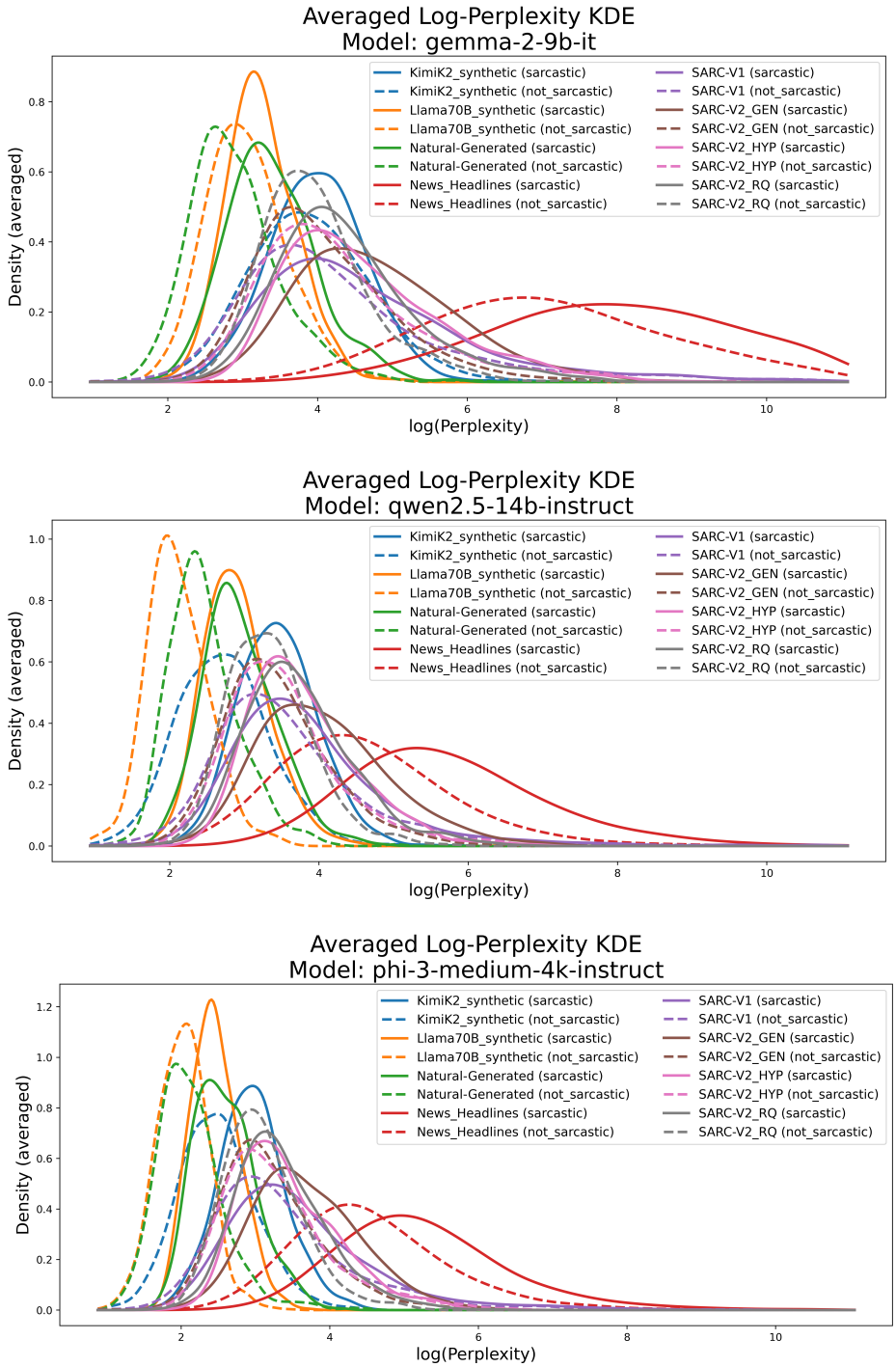

Figure 10: Log-perplexity distributions for sarcastic and non-sarcastic text across datasets for gemma-2-9b-it (top), phi-3-medium-4k-instruct (middle) & qwen2.5-14b-instruct (bottom).

dominated by distributional artifact recognition rather than grounded pragmatic understanding.

D Embedding

This appendix provides additional embedding visualizations to support the qualitative observations reported in Section 4.2.1. Across the full study, we generated embeddings for 8 datasets, 4 models, and 3 transformer depths (first, middle, final), yielding 96 visualizations in total. Due to space constraints, we include a small subset of figures here, all the remaining visualizations exhibit qualitatively similar behavior and are therefore omitted for brevity.

Gemma-2-9b-it (Kimi-k2-instruct-0905, Synthetic)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	1231	182
Sarcastic	0	1566

Phi-3-medium-4k-instruct (Kimi-k2-instruct-0905, Synthetic)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	1334	79
Sarcastic	0	1566

Qwen2.5-14b-Instruct (Kimi-k2-instruct-0905, Synthetic)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	1390	23
Sarcastic	0	1566

Table 6: Confusion matrices for sarcasm detection (Kimi-k2 dataset)

Gemma-2-9b-it (Fewshot-llama-70b, Natural-Synthetic)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	377	157
Sarcastic	1	583

Phi-3-medium-4k-instruct (Fewshot-llama-70b, Natural-Synthetic)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	401	133
Sarcastic	0	584

Qwen2.5-14b-Instruct (Fewshot-llama-70b, Natural-Synthetic)

	Predicted	
True	Not Sarcastic	Sarcastic
Not Sarcastic	506	28
Sarcastic	2	582

Table 7: Confusion matrices for sarcasm detection (Fewshot generated dataset)

D.1 Dataset-Driven Geometry Under a Fixed Model

Figures 11 and 12 present additional projections for Llama-3.1-8B-Instruct on a synthetic dataset (Kimi-k2) and an organic dataset (SARC-V1), respectively, across multiple layers.

For the synthetic dataset, sarcastic and non-sarcastic samples occupy largely separate portions of the embedding space, even though the overall distribution follows a continuous, non-linear shape. This separation is visible at intermediate layers and persists in later layers without becoming substantially sharper.

In contrast, the organic dataset exhibits substantial overlap between classes at all examined depths. Although intermediate layers occasionally impose a global geometric organization, sarcastic and non-sarcastic samples remain intermingled, and no consistent class-wise separation emerges in deeper layers.

D.2 Consistency Across Models

To assess whether the observed behavior is specific to a single model, Figures 13 and 14 show corresponding embedding projections for Qwen-2.5-14B-Instruct on the same organic and synthetic datasets.

These results indicate that the embedding behavior described in Section 4.2.1 generalizes across model families and is not confined to a particular architecture or parameter scale.

The same qualitative regimes are observed. Synthetic data again displays clear separation between sarcastic and non-sarcastic examples across layers, whereas organic data remains entangled. Differences in projection shape and density across models do not alter this underlying pattern.

D.3 Relation to the Main-Body Analysis

Section 4.2.1 highlights that strong class separation in embedding space appears primarily for synthetic datasets, whereas embeddings for natural human data remain highly overlapping. The additional figures in this appendix demonstrate that this contrast persists across models and layers and is not limited to the specific examples shown in the main text.

E Perplexity Analysis

For a sequence x of length N, the perplexity is defined as:

$$PPL(x) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\ln P_\theta(x_i|x_{<i})\right)$$

The goal of this appendix is to assess whether the qualitative behavior observed under LLaMA-3.1-8B-Instruct (section 4.2.2) is preserved across models. Across all evaluation models, datasets organize into stable perplexity regimes, despite differences in absolute scale. Fully synthetic datasets

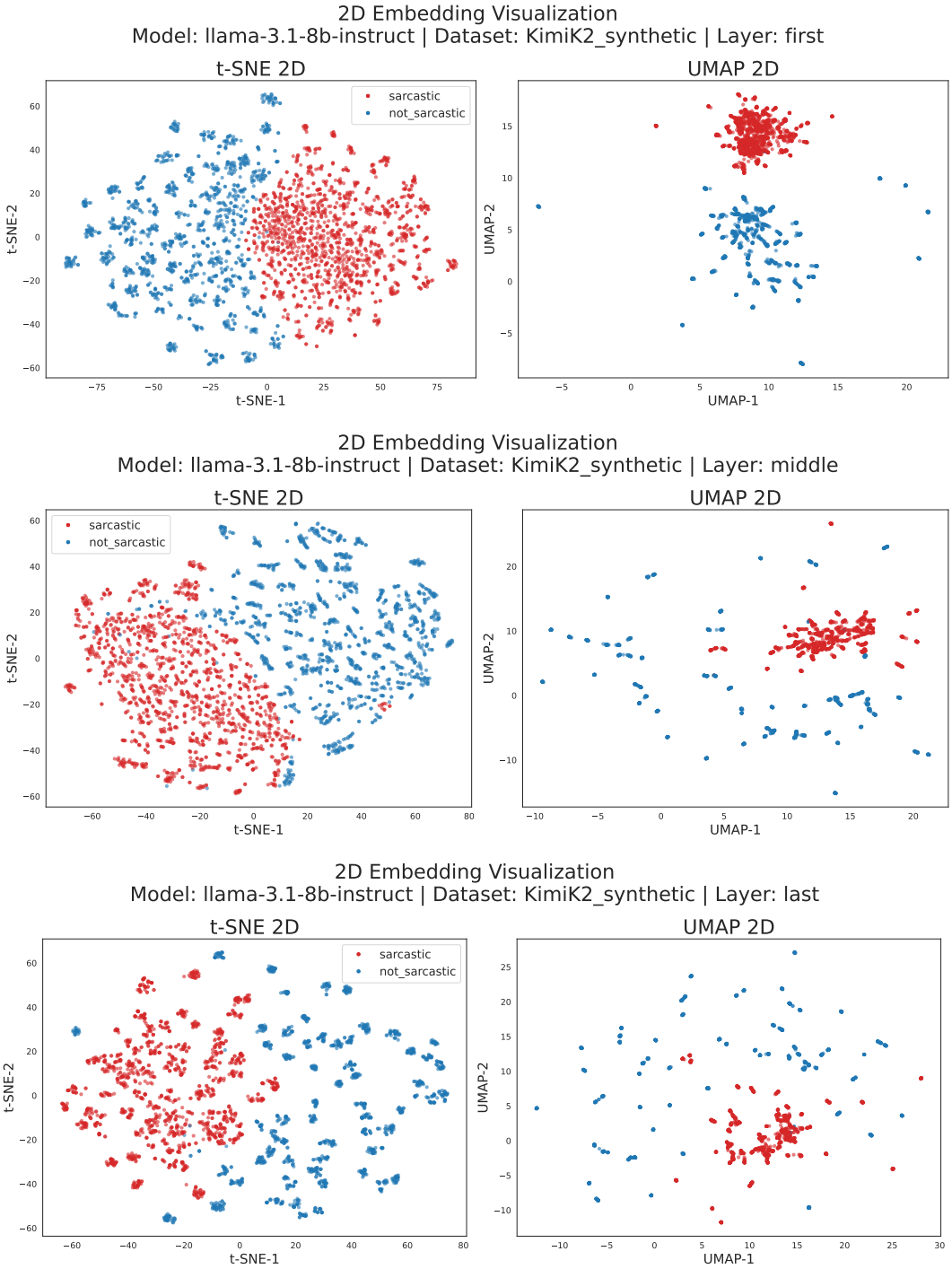

Figure 11: 2D projection of early (top), intermediate (middle) and final (bottom) layer embeddings for Llama-3.1-8B-Instruct on Kimi-k2 (left: t-SNE, right: UMAP).

consistently occupy lower log-perplexity regions than natural datasets, while natural datasets exhibit higher perplexity and substantial overlap between sarcastic and non-sarcastic samples. This qualitative structure is preserved across evaluators, indicating that the patterns reported in Section 4.2.2 are not model-specific.

A clear contrast emerges in class-conditional structure. In fully synthetic data, sarcastic and non-sarcastic samples exhibit systematic shifts in log-perplexity, yielding visible separation. In contrast, natural datasets remain broadly overlapping on class label, indicating that perplexity does not provide a reliable discrimination signal in natural language.

Few-shot Natural-Generated data occupies a

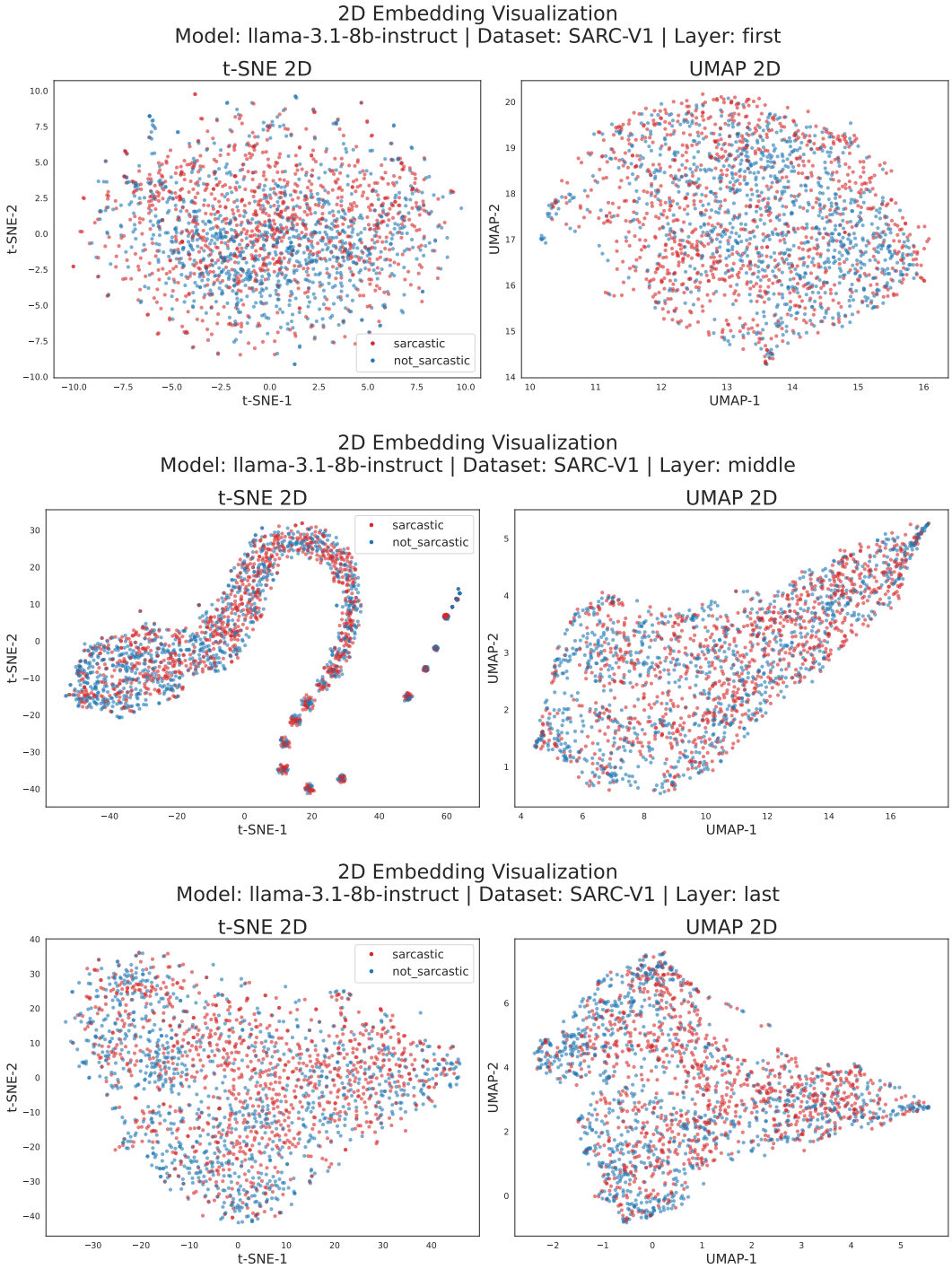

Figure 12: 2D projection of early (top), intermediate (middle), and final (bottom) layer embeddings for llama-3.1-8b-instruct on SARC-V1 (left: t-SNE, right: UMAP).

low-perplexity region under all evaluators, with distributional modes aligning with specific synthetic regimes rather than forming a smooth intermediate between synthetic and natural data. This alignment reflects how the evaluation model scores the generated text, not equivalence of generation processes.

Overall, these results confirm that perplexity supports class separation primarily in synthetic regimes, while remaining insufficient for sarcasm discrimination in natural language across evaluation models.

F Membership Inference - Min-K%++

Min-K%++ serves as a normalized entropy-sensitive proxy for distributional familiarity, mea-

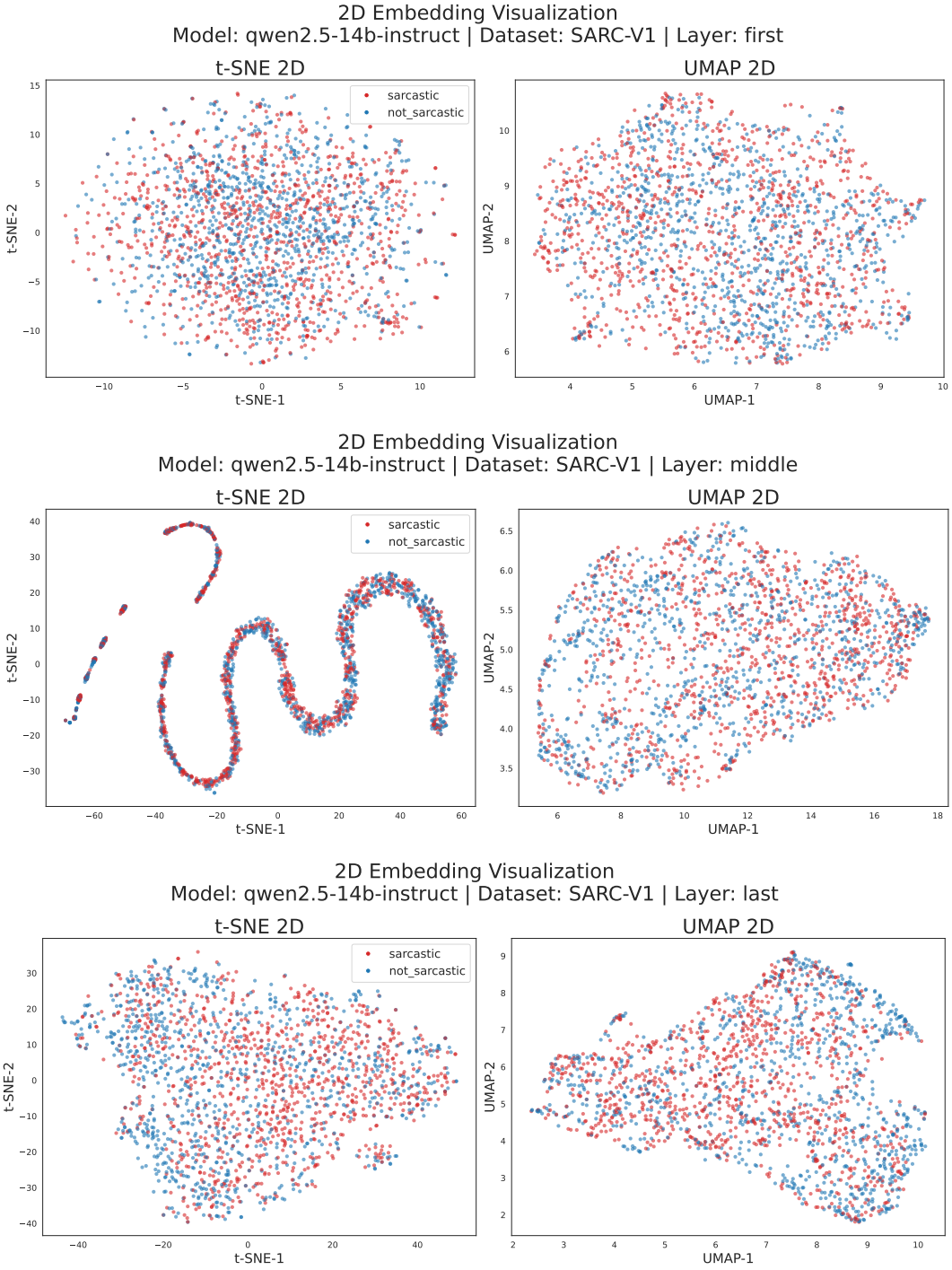

Figure 13: 2D projection of early (top), intermediate (middle), and final (bottom) layer embeddings for Qwen-2.5-14B-Instruct on SARC-V1 (left: t-SNE, right: UMAP).

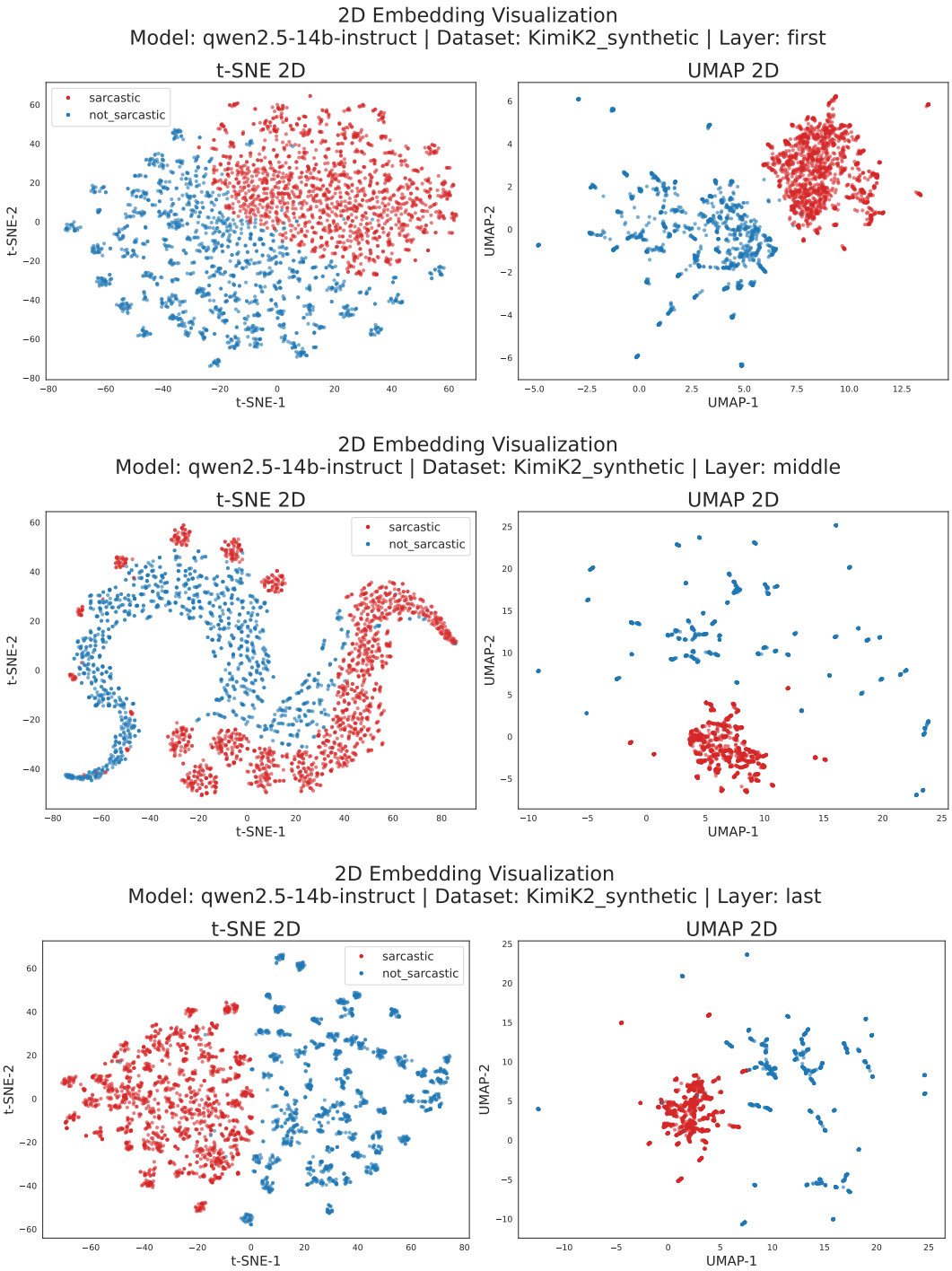

Figure 14: 2D projection of early (top), intermediate (middle) and final (bottom) layer embeddings for Qwen-2.5-14B-Instruct on Kimi-k2 (left: t-SNE, right: UMAP).

suring how concentrated a model's token-level probability mass is relative to its internal expectations. Higher scores indicate sequences that align closely with the model's preferred generative statistics, while lower scores reflect higher-entropy or distributionally atypical text.

This appendix reports additional Min-K%++ score distributions to complement the analysis presented in Section 4.2.3. Figure 15 presents the resulting distributions for three additional evaluation models.

F.1 What the Min-K%++ Distributions Show

Across evaluation models, Min-K%++ score distributions vary in absolute scale, peak location, and relative positioning between datasets. No

consistent ordering or stable regime is preserved across evaluators: the same dataset may appear comparatively concentrated or dispersed depending on the model used for evaluation.

Few-shot Natural–Generated datasets likewise do not exhibit a fixed or intermediate behavior. Their Min-K%++ distributions shift across evaluators without a consistent relationship to either fully synthetic or organic datasets.

F.2 Dependence on the Model

Figure 15 shows that Min-K%++ behavior is strongly dependent on the evaluation model. Changes in model architecture and training lead to systematic shifts in score scale, distribution width, and peak location.

As a result, Min-K%++ values should be interpreted within the context of a single evaluator, rather than compared directly across models. However, the presence of dataset-dependent differences in distribution shape remains visible under all evaluators considered.

G LIME

To assess whether the behavior observed in the main-body LIME analysis reflects an isolated case or a systematic pattern, we apply Local Interpretable Model-agnostic Explanations (LIME) to multiple additional examples. These figures span both fully synthetic and organic datasets and are intended to test the repeatability of the attribution behavior across different sentences and models.

For each sentence, we generate 2,000 perturbed variants through random masking of input tokens and fit a locally linear surrogate model around the original prediction. Token-level importance scores reflect the sensitivity of the classifier's output to local perturbations, providing insight into whether predictions are anchored to semantically meaningful cues.

G.1 Comparison of Figures 16–17 with Figure 8 (Model-wise Consistency)

Figures 16 and 17 present LIME explanations for the same sarcastic sentence analyzed in Figure 8, but evaluated under different models (Gemma and LLaMA). Despite architectural differences, the qualitative attribution behavior remains similar.

As in Figure 8, neither model assigns dominant importance to tokens that encode the core pragmatic incongruity of the sentence. **Even if the important token is selected, it's contribution is observed to be on the lower side.** This mirrors the main-body observation that LIME explanations for correctly predicted sarcastic examples are diffuse and low-magnitude, even when the prediction itself is correct.

The consistency between Figures 8, 16, and 17 suggests that this behavior is not model-specific, but rather reflects a shared reliance on distributed statistical cues instead of explicit semantic grounding. In other words, different models arrive at the same prediction without converging on a shared, interpretable set of sarcasm-defining tokens.

G.2 Comparison of Figures 18–19: Correct vs. Incorrect Predictions

Figures 18 and 19 examine a more revealing regime: sentences where the models identify plausible contextual tokens but diverge in their final predictions.

In Figure 18 (LLaMA), the sentence is sarcastic and correctly classified as such. LIME assigns high importance to the token "Dinosaurs", which functions as a key semantic anchor by evoking anachronism and absurdity in context. This attribution aligns with human intuition: the sarcastic force of the sentence hinges on the incongruity introduced by this concept.

Figure 19 presents the same sentence evaluated under Qwen, which predicts the sentence as non-sarcastic (incorrectly). Crucially, LIME still assigns relatively high importance to the same token ("Dinosaurs") as compared to the other tokens of this sentence. That is, the model attends to the correct word, yet arrives at the wrong decision.

This contrast may highlight a central limitation of attribution-based explanations: identifying the right token is not sufficient if the model fails to correctly integrate that token into a broader pragmatic judgment.

Taken together, Figures 16–19 indicate that the attribution patterns observed in the main-body analysis recur across models and prediction outcomes. While LIME often highlights tokens that are intuitively relevant to the interpretation of sarcasm, these attributions do not consistently differentiate between correct and incorrect predictions (giving false-positive predictions more frequently). In particular, high importance assigned to contextually salient words can coincide with both successful and failed classifications. This

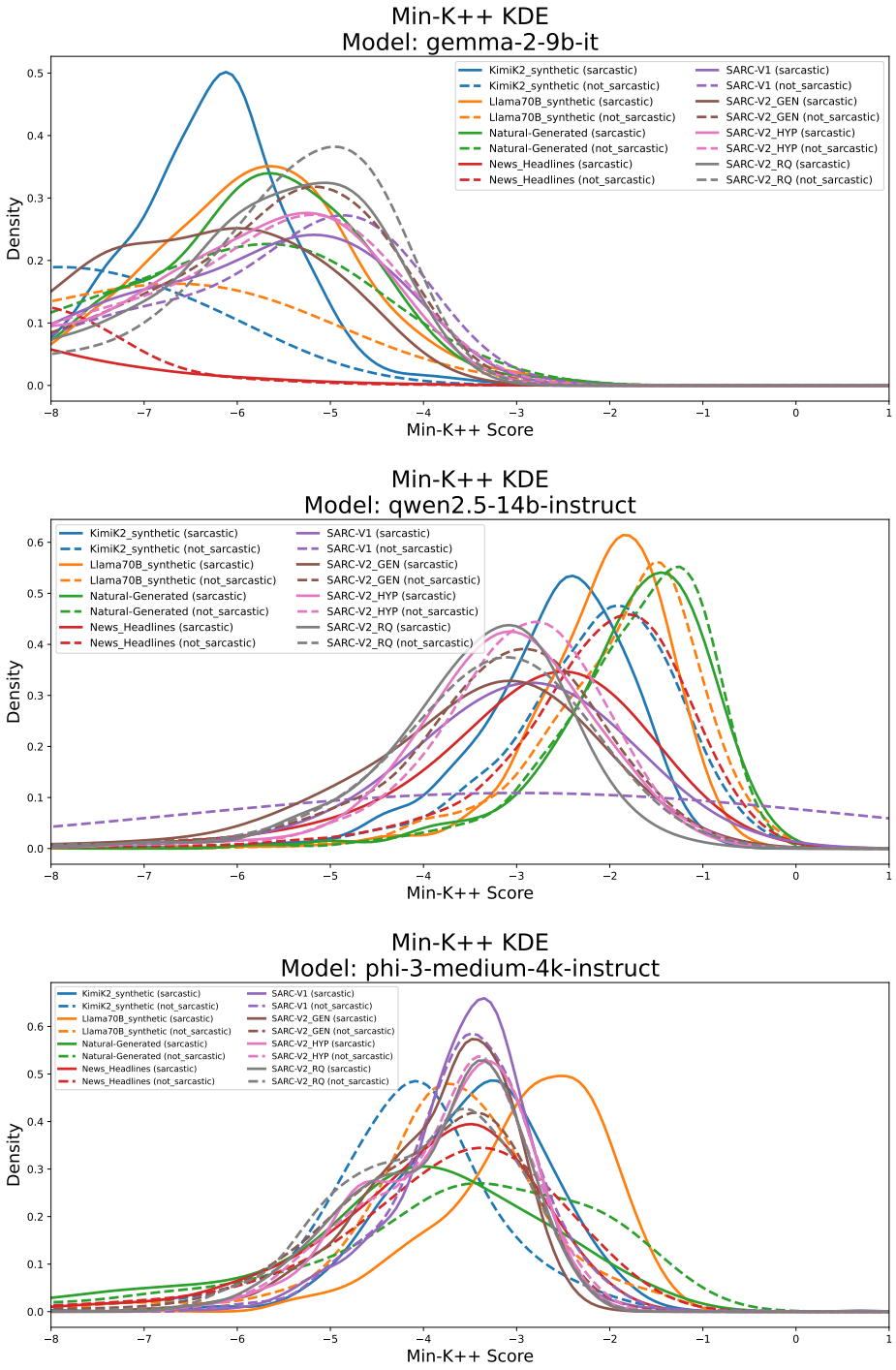

Figure 15: Min-K%++ score distributions for sarcastic and non-sarcastic text across datasets for gemma-2-9b-it (top), phi-3-medium-4k-instruct (middle) & qwen-2.5-14b-instruct (bottom).

suggests that token-level importance alone is insufficient to characterize how such cues are ultimately used in the model's decision process. Accordingly, the appendix results should be interpreted as illustrating the limited extent to which local attribution methods can disentangle lexical sensitivity from higher-level pragmatic reasoning in sarcasm detection.

H Attention

This appendix provides additional attention visualizations to support the analysis presented in Section 4.2.5. All visualizations are generated using BertViz (Vig, 2019). For each example, attention distributions are shown at three representative depths-early, intermediate, and final transformer

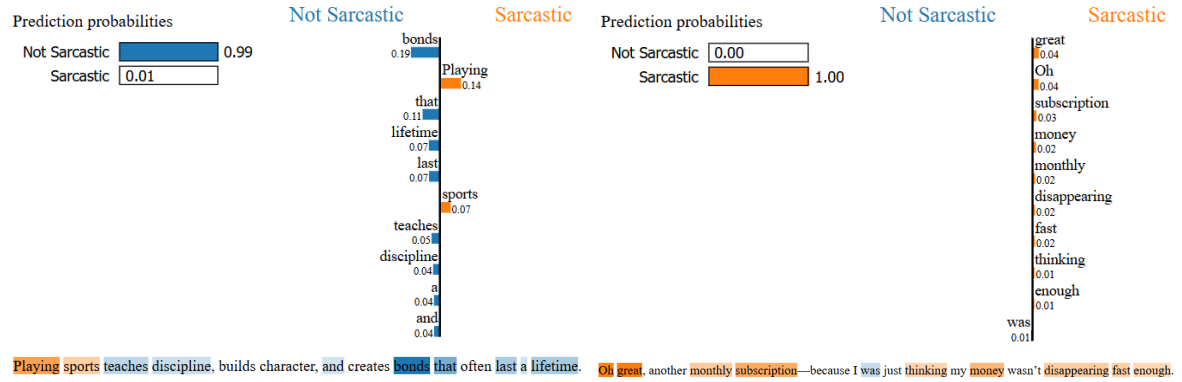

Figure 16: LIME visualization for a non-sarcastic (left) and sarcastic (right) sentence from the kimi-k2 (synthetic) dataset for the Gemma-2-9b-it model.

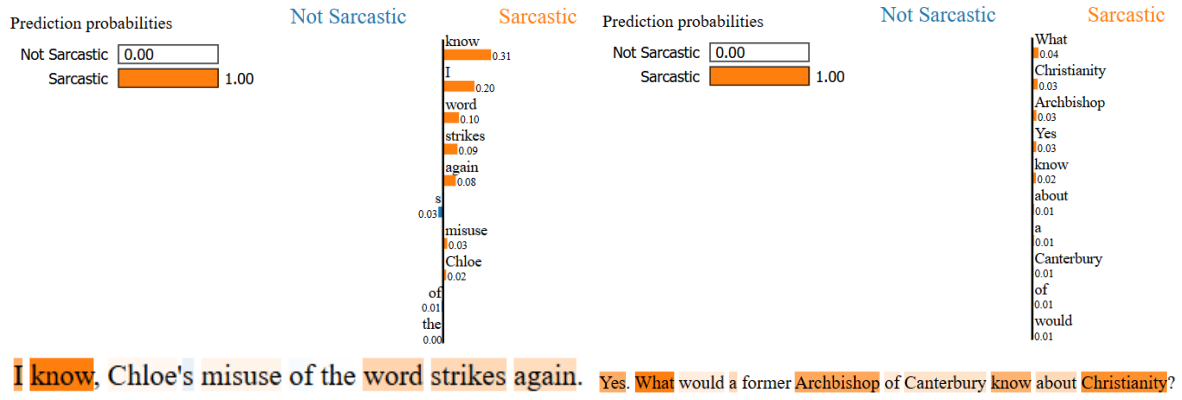

Figure 17: LIME visualization for a non-sarcastic (left) and sarcastic (right) sentence from the SARC-V1 (natural) dataset for the Gemma-2-9b-it model.

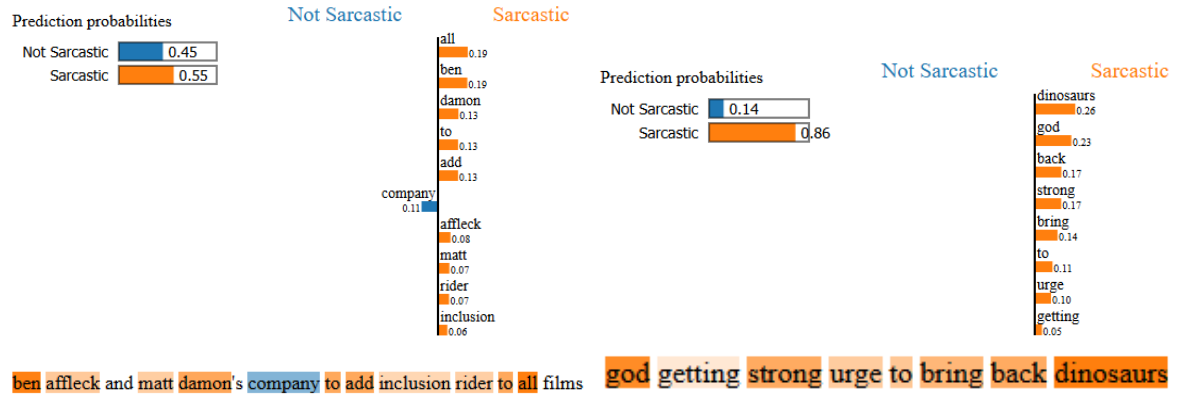

Figure 18: LIME visualization for a non-sarcastic (left) and sarcastic (right) sentence from the News Headlines (natural) dataset for the Llama-3.1-8b-instruct model.

layers (Figures 20, 21 & 22).

H.1 Early-Layer Sensitivity to Distributed Context

In early layers, attention heads exhibit relatively broad distributions, with weak but visible links spanning non-adjacent tokens. The early layers occasionally display faint attention paths connecting semantically relevant tokens.

However, these connections are neither sharp nor dominant. They coexist with substantial attention allocated to positional or structural tokens, and do not stand out as privileged or consistently selected across heads.

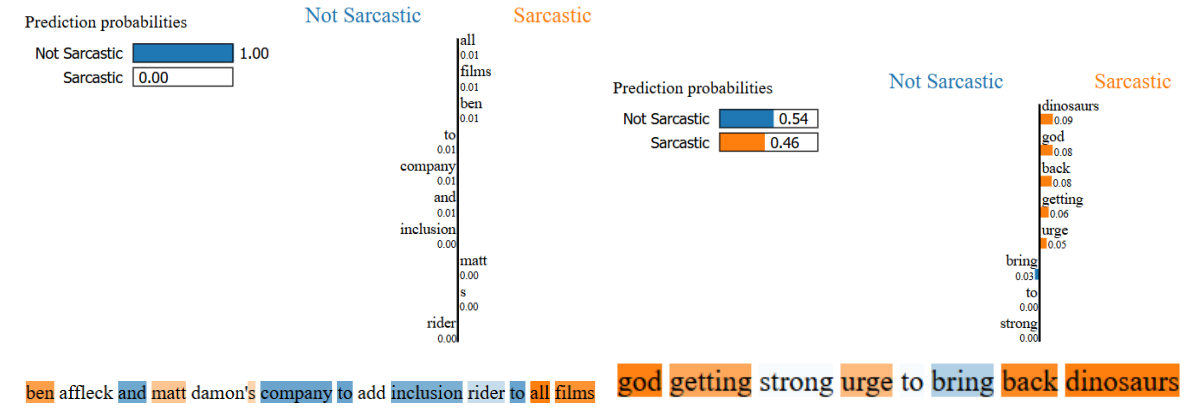

Figure 19: LIME visualization for a non-sarcastic (left) and sarcastic (right) sentence from the News Headlines (natural) dataset for the Qwen-2.5-14b-instruct model.

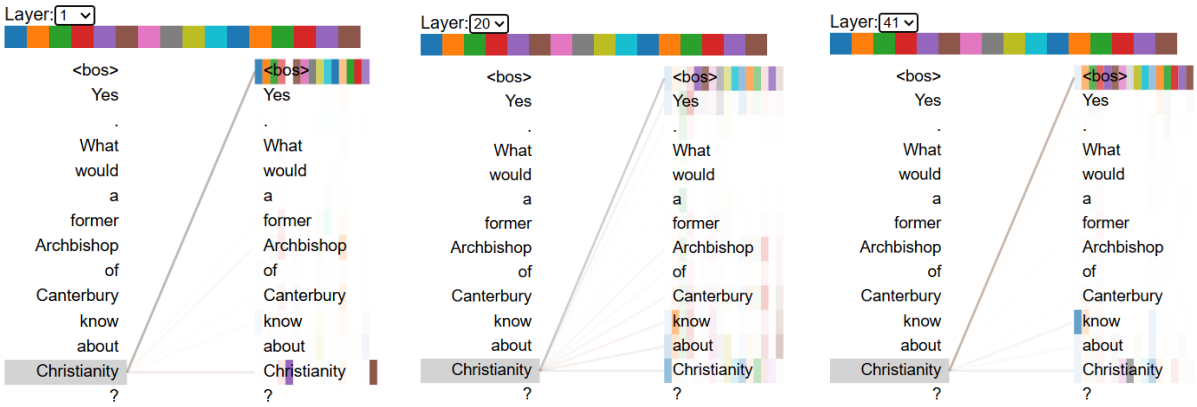

Figure 20: Attention distribution across early, middle and late transformer layer for a sarcastic sentence (model: gemma-2-9b-it, dataset: SARC-V1).

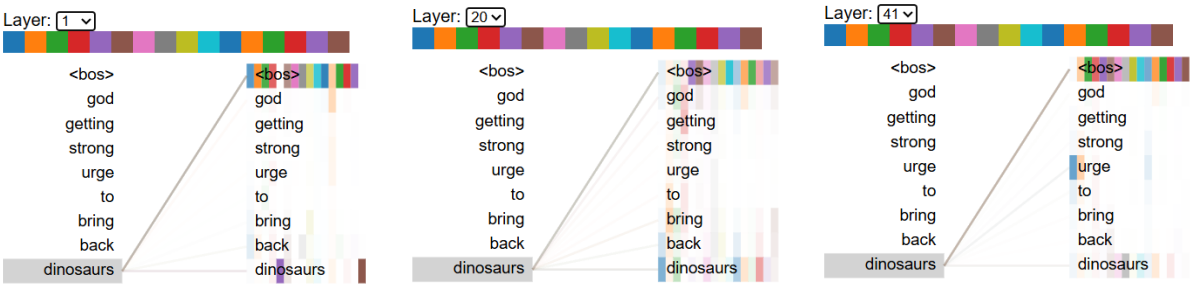

Figure 21: Attention distribution across early, middle and late transformer layer for a sarcastic sentence (model: gemma-2-9b-it, dataset: News Headlines).

H.2 Decay of Non-Local Attention with Depth

As depth increases, the attention patterns change systematically. In intermediate layers, previously visible long-range links weaken, while attention becomes increasingly concentrated around a small set of tokens. By the final layers, attention (excluding special tokens) is largely confined to immediate neighbors or short local spans.

Crucially, the figures show no consolidation of long-range dependencies with depth. Connections that would be required to integrate distant semantic cues do not strengthen or stabilize. Instead, they progressively diminish, indicating that deeper layers do not refine or reinforce non-local contextual relationships.

Figure 22: Attention distribution across early, middle and late transformer layer for a sarcastic sentence (model: llama-3.1-8b-instruct, dataset: News Headlines).

H.3 Dominance of Positional Focus in Final Layers

Across examples, final-layer attention is strongly dominated by positional focus, with a large fraction of heads allocating maximal weight to the beginning-of-sequence token.

The dominance of this positional focus persists across sentences and datasets. Importantly, it does not vary meaningfully between sarcastic and non-sarcastic inputs within organic data, suggesting that it reflects a general attention pattern rather than a sarcasm-specific mechanism.

Section 4.2.5 argues that, although early layers sometimes exhibit weak sensitivity to distributed context, this information is not preserved or integrated through the network. The appendix figures reinforce this claim by demonstrating that:

- Long-range attention links are fragile and transient.
- These links decay rather than strengthen with depth.
- Final-layer attention is dominated by positional and local focus, not relational semantic structure.

Accordingly, the appendix figures support the main-body conclusion that attention dynamics do not exhibit the progressive integration of non-local contextual cues required for robust sarcasm interpretation.

I Layer Ablation

This appendix reports layer-wise block removal ablation results for three evaluation models: gemma-2-9b-it, phi-3-medium-4k-instruct, and qwen2.5-14b-instruct (Figure 24). The goal of this analysis is to assess whether sarcasm detection depends critically on specific transformer layers or whether the decision signal is distributed across the network.

I.1 Experimental Procedure

For each model, we perform identity ablation by replacing the output of a single transformer block with the identity function during inference, effectively bypassing that layer while preserving the residual stream. All other layers remain unchanged.

Accuracy is measured on a fixed evaluation subset of SARC-V1, and the accuracy drop (relative to the unablated baseline) is reported for each layer. Positive values indicate performance degradation upon removal, while negative values indicate slight performance improvements.

This experiment is intended as a diagnostic probe, not as a precise localization method.

I.2 Observations Across Models

Across all three models, several consistent properties emerge:

- **Small magnitude of effects**: For most layers, removing a single block results in only modest changes in accuracy. Even the largest drops remain limited in scale relative to overall performance.
- **Lack of a monotonic depth trend**: Accuracy drops do not increase systematically with depth. Early, middle, and late layers all exhibit a mixture of positive, near-zero, and occasionally negative effects.
- **Non-unique "important" layers**: While certain layers produce larger drops than others, these peaks are sparse and model-specific. No single layer or contiguous layer range consistently dominates across models.

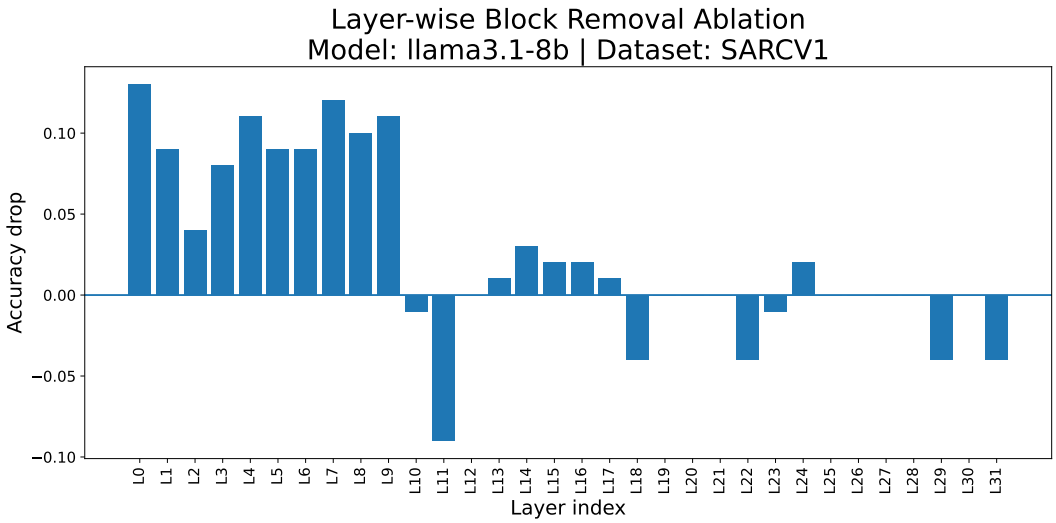

Figure 23: Layer ablation for LLaMA-3.1-8B.

- **Occasional negative drops**: In all models, some layers exhibit slight accuracy improvements when removed, indicating that their contribution is not strictly beneficial for sarcasm classification in this setting.

I.3 Model-Specific Notes

- **Gemma-2-9B-IT**: The ablation profile is relatively flat, with small fluctuations around zero. A few isolated layers produce moderate drops, but no stable hierarchy emerges.
- **Phi-3-Medium-4K-Instruct**: Larger variance is observed compared to Gemma, with several early and mid-layer removals causing noticeable accuracy drops. However, these effects are not monotonic and are interspersed with layers exhibiting negligible or negative impact.
- **Qwen-2.5-14B-Instruct**: Similar to Phi-3, a small number of mid-layer blocks yield larger drops, but these are isolated and not indicative of a dedicated sarcasm-processing stage.

I.4 Interpretation

Taken together, these results suggest that sarcasm-relevant signals are not localized to a specific transformer block. Instead, performance appears to rely on distributed, weakly contributing features spread across layers.

This analysis does not imply the absence of computation or representation related to sarcasm. Rather, it indicates the absence of a single, hierarchically refined, or indispensable layer-level mechanism for sarcasm detection.

Given the small effect sizes, the absence of consistent depth-wise trends, and the presence of negative ablation effects, these results are most consistent with a shallow and distributed decision signal, rather than a robust, progressively constructed semantic mechanism.

J Token Occlusion

This appendix extends the token occlusion analysis presented in Section 4.2.7 by examining whether the observed semantic insensitivity persists across multiple evaluation models and datasets. The goal is confirmatory rather than diagnostic: to verify that the failure of causal dependence on meaning-bearing tokens is not confined to a single model.

J.1 Experimental Scope

Token occlusion experiments were conducted on sarcastic examples drawn from both synthetic and organic datasets, including Kimi-K2 and SARC-V1, using the evaluation models LLaMA-3.1-8B-Instruct, Gemma-2-9B-it, Qwen2.5-14B-Instruct, and Phi-3-medium-4k-instruct. For each sentence, individual tokens were iteratively masked and the resulting change in predicted sarcasm probability was recorded, following the protocol in Section 2.5.5.

J.2 Cross-Model Consistency of Occlusion Behavior

For synthetic examples, occlusion produces nearly flat attribution maps across all models, indicating

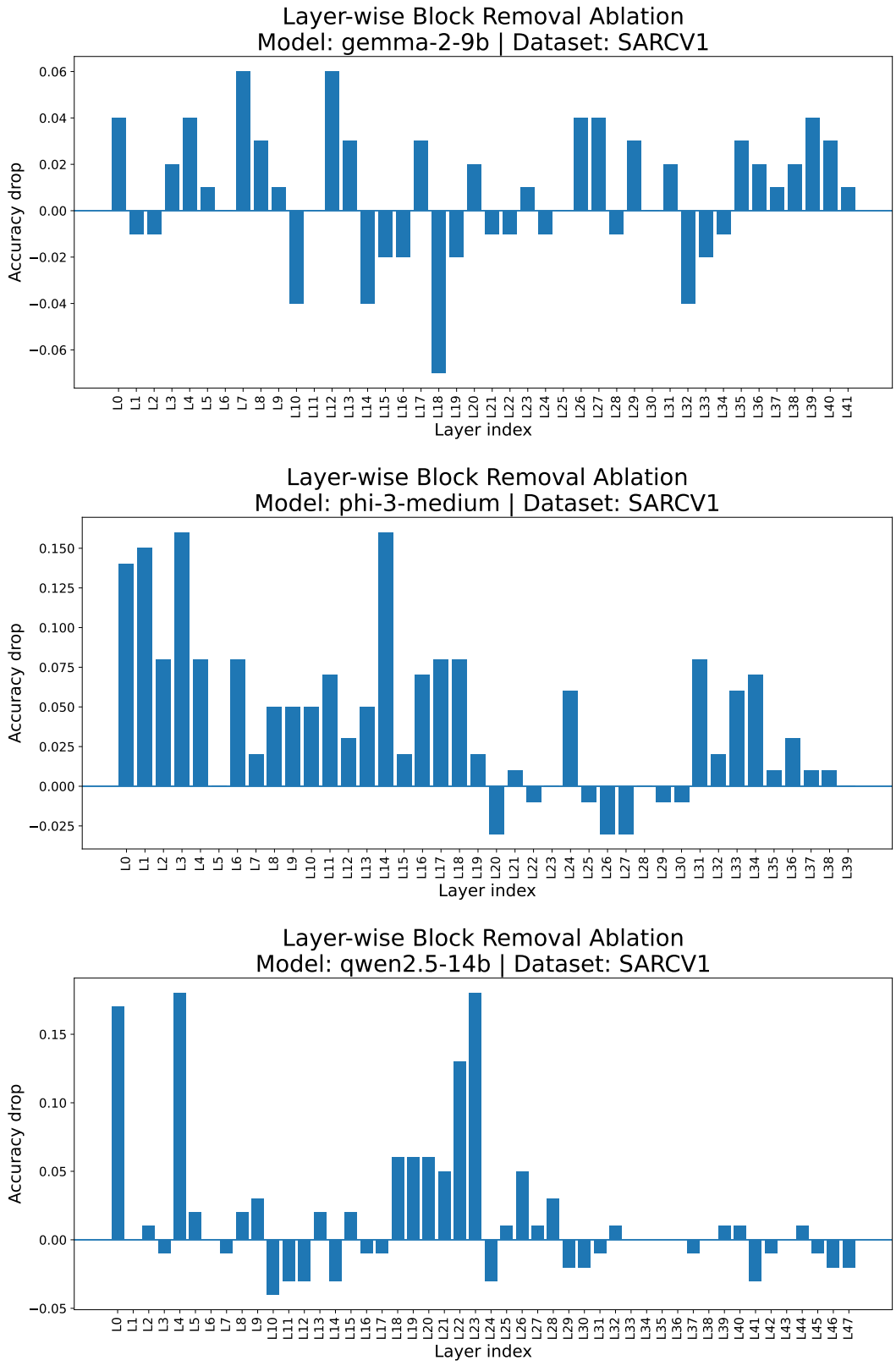

Figure 24: Layer ablation for gemma-2-9b-it (top), phi-3-medium-4k (middle) & qwen2.5-14b-instruct (bottom)

that predictions are invariant to lexical perturbations and not causally grounded in specific tokens. This behavior mirrors the flat sensitivity observed in Figure 26 and reflects reliance on distributed distributional cues rather than semantic content.

In contrast, organic examples exhibit non-

uniform but weak token sensitivity. While masking individual tokens induces localized confidence changes, these perturbations rarely alter the predicted class. Crucially, for non-sarcastic SARC-V1 example (Figure 25), occlusion effects are skewed toward increasing confidence in the sarcastic class, rather than stabilizing or reinforcing the non-sarcastic prediction. That is, removing tokens from an already non-sarcastic sentence pushes the model further toward sarcasm.

This directional bias is consistent across evaluation models and persists despite differences in model.

J.3 Relation to False Positives in the Main Body

This directional effect provides a causal explanation for the high false-positive rates on organic datasets reported in Section 4.1.1 and Appendix C. As shown in the confusion matrices, models trained or evaluated on synthetic data tend to overpredict sarcasm in natural human text, yielding high recall but low precision.

Taken together with the Min-K%++ results (Section 4.2.3) and the attention analysis (Section 4.2.5), this finding suggests that the false positives are not incidental errors. They arise from a systematic inductive bias: when semantic grounding is weak or disrupted, the model defaults to detecting sarcasm based on distributional texture rather than meaning.

K Discussion

Our mechanistic analysis challenges the implicit assumption that the high performance of Large Language Models on sarcasm benchmarks reflects genuine pragmatic reasoning. By triangulating behavioral, geometric, and causal evidence, we demonstrate that the ceiling performance observed on synthetic datasets is largely a *Synthetic Mirage*—a reflection of the model's ability to detect distributional artifacts rather than linguistic meaning.

K.1 Sarcasm Detection as Authorship Attribution

The most significant finding is the catastrophic generalization gap between synthetic and organic domains. Our geometric analysis explains this dichotomy: synthetic sarcasm forms separable, low-entropy manifolds, while organic sarcasm remains inextricably entangled with non-sarcastic text.

This implies that on synthetic benchmarks, models are effectively solving authorship attribution. They identify the "statistical fingerprint" of the generator (e.g., Llama-3.3 or Kimi-k2) rather than semantic incongruity. The Min-K%++ results confirm this: synthetic sarcastic examples possess a unique, low-entropy signature that models exploit as a separator. When exposed to organic text, where this "generator signature" is absent, the model reverts to a biased prior, hallucinating sarcasm in high-entropy natural speech.

K.2 The Mechanism of Semantic Blindness

Crucially, our causal interventions reveal that models are functionally "blind" to the text. Token Occlusion and LIME demonstrate that removing semantic anchors (e.g., "Archbishop" vs. "Christianity") results in negligible confidence shifts. This is corroborated by Attention Visualization, where the dominance of the `<|begin_of_text|>` token (the "attention sink") and the decay of long-range dependencies suggest the model aggregates global statistical texture rather than resolving local semantic conflict. The robustness of the signal under Layer-wise Identity Ablation further proves that the detected feature is a pervasive global artifact, not a hierarchically constructed pragmatic insight.

K.3 The Few-Shot Paradox

The Few-Shot Llama dataset presents a critical paradox: while it exhibits the strongest artifact signature (highest Min-K%++ scores), its classification accuracy remains intermediate. This indicates a decoupling of signal and semantics; the model easily detects the "hyper-synthetic" nature of the text, but the semantic noise introduced by natural few-shot prompts disrupts the near-perfect separability found in fully synthetic data. Consequently, current benchmarks likely measure a model's sensitivity to "LLM-speak" rather than its understanding of human irony.

L Prompt Templates used

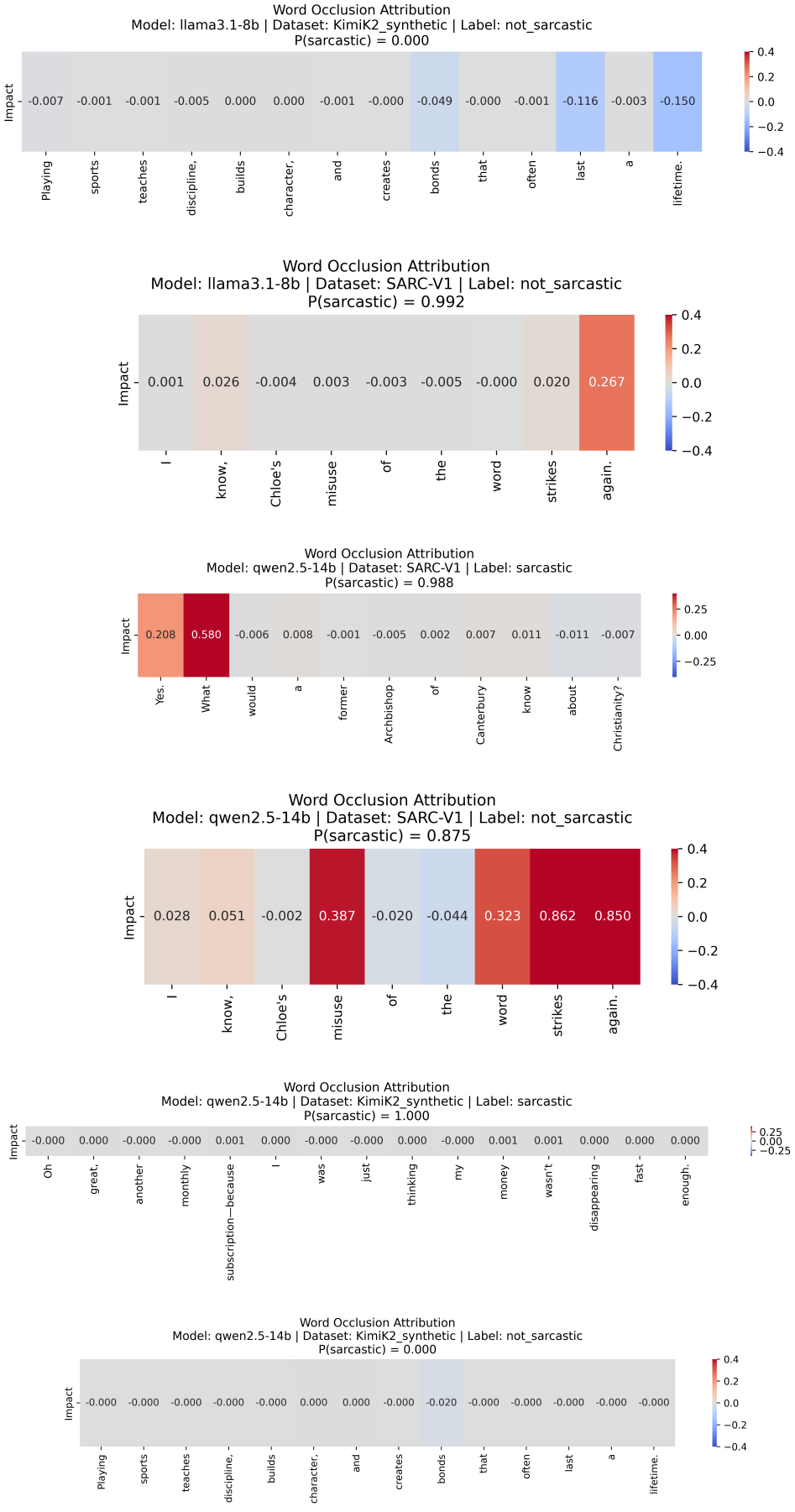

Figure 25: Token Occlusion attribution for sarcastic and non-sarcastic sentences for the model llama3.1-8b-instruction and qwen2.5-14b.

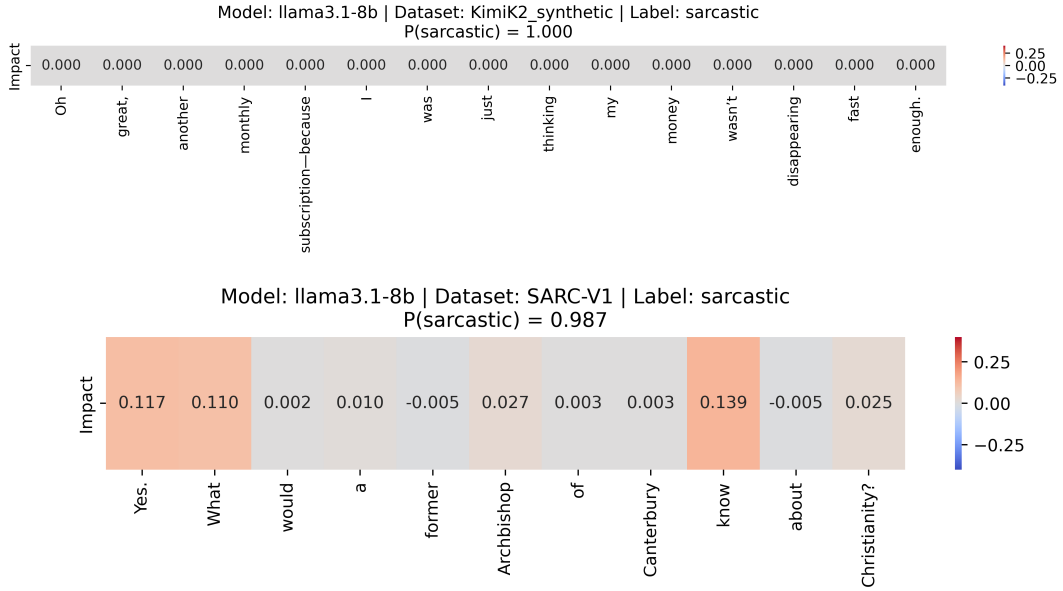

Figure 26: Token occlusion attribution for sarcastic sentences from the Kimi-k2 (top) and SARC-V1 (bottom) dataset.

Model	Prompt
Llama-3.1-8B-Instruct	<—start_header_id—>system<—end_header_id—> Your task is to classify if a sentence is sarcastic. Analyze the given sentence. Sarcasm often involves saying the opposite of what is true, using over-exaggeration, or conveying mockery. If the intent is to mock or criticize through irony, it is sarcastic. Answer with only "Yes" or "No". Sentence: {sentence}<—eot_id—><—start_header_id—>user<—end_header_id—> Is the response sarcastic?<—eot_id—><—start_header_id—>assistant<—end_header_id—>
gemma-2-9b-it	Determine whether the following sentence is sarcastic. Answer with only one word: Yes — if the sentence is sarcastic No — if the sentence is not sarcastic Sentence: "{sentence}" Answer:
Qwen2.5-14B-Instruct	You are known for being able to precisely classify whether a sentence is sarcastic or not. Determine whether the sentence is sarcastic. Sentence: "{sentence}" Is the sentence sarcastic? Answer strictly with only "Yes" or "No":
Phi-3-medium-4k-instruct	Is the following sentence sarcastic? Answer Yes or No Only, no other text. Sentence: "{sentence}" Answer:

Table 8: Prompt templates used for extracting only "Yes" and "No" tokens across models.

Tone	Prompt
sarcastic	Write exactly ONE English sentence that is sarcastic. Topic: {topic} Sentence:
non-sarcastic	Write exactly ONE English sentence that is sincere and not sarcastic. Topic: {topic} Sentence:

Table 9: Prompt templates used for generating datasets.

Topics
politics, elections, government policies, bureaucracy, public services, taxes, school exams, college life, online classes, grading systems, office meetings, corporate culture, deadlines at work, remote work, technology, artificial intelligence, software updates, slow internet, social media, influencers, online ads, spam emails, daily routines, commuting to work, traffic jams, waiting in queues, relationships, dating apps, breakups, friendship, mental health, stress, burnout, sleep schedules, movies, movie sequels, streaming platforms, celebrity scandals, sports, team loyalty, fantasy leagues, video games, shopping online, subscriptions, hidden fees, refund policies, travel plans, flight delays, public transport, coffee not working, ordering food, burnt meals, science news, research funding, space exploration, being busy, time management, procrastination, to-do lists

Table 10: Topics used for synthetic dataset generation.

	Prompt
Prompt A	You are generating English sentences written in a casual, conversational tone. Below are examples of {tone} sentences written by humans. They are natural and context-dependent. Examples: \n{example_block}\n\n Now generate ONE NEW {tone} sentence that matches the style and subtlety of the examples above. Use natural conversational phrasing, not formal or essay-like language. Do NOT paraphrase or copy the examples. Output only the sentence.
Prompt B	You are generating short, direct English sentences. Below are examples of {tone} sentences written by humans. They are natural and context-dependent. Examples: \n{example_block}\n\n Now generate ONE NEW {tone} sentence that matches the style and subtlety of the examples above. Keep the sentence concise and deadpan; avoid explanations or hedging. Do NOT paraphrase or copy the examples. Output only the sentence.
Prompt C	You are generating English sentences that describe a specific situation or moment. Below are examples of {tone} sentences written by humans. They are natural and context-dependent. Examples:\n{example_block}\n\n Now generate ONE NEW {tone} sentence that matches the style and subtlety of the examples above. Focus on a concrete situation rather than abstract argument. Do NOT paraphrase or copy the examples. Output only the sentences.

Table 11: Prompt templates used for generating more natural text. <tone>switches between the words "sarcastic" and "non-sarcastic" and <example_block>contains three randomly selected examples from SARC V1 of a specific tone.