One Joke to Rule them All?
On the (Im)possibility of Generalizing Humor Detection

Mor Turgeman[1] **Chen Shani**[2] **Dafna Shahaf**[1]
[1]The Hebrew University of Jerusalem [2]Stanford University
mortur@cs.huji.ac.il, cshani@stanford.edu, dshahaf@cs.huji.ac.il

Abstract

Humor is a complex form of communication that remains challenging for machines. Despite its broadness, most existing research on computational humor traditionally focused on modeling one specific type of humor. In this work, we wish to understand whether competence on specific humor tasks confers any ability to transfer to novel, unseen types; in other words, is this fragmentation inevitable? This question is especially timely as new humor types continuously emerge in online contexts (e.g., memes, anti-humor, AI fails). If LLMs are to keep up with this evolving landscape, they must be able to capture deeper, transferable mechanisms.

To investigate this, we conduct a series of transfer learning experiments across four datasets, representing different humor tasks. We explore varied diversity settings (varying between 1-3 datasets in training, testing on a novel one). Experiments show that models are capable of some transfer, reaching up to 75% accuracy on binary unseen datasets; training on diverse sources improves transferability (1.88-4.05%) with minimal-to-no drop in in-domain performance. Somewhat surprisingly, the one dataset (Dad Jokes) emerges as the best enabler of transfer, but the hardest one to transfer to. We release data and code.[1]

1 Introduction

Humor spans a wide range of styles and mechanisms, from puns and sarcasm to absurdity and satire (Attardo, 2024; Raskin, 1979), many of which involve linguistic play, pragmatic inference, or violations of logical expectations (Suls, 1972; Attardo, 2000). It is subjective and culturally dependent (Attardo, 2024; Martin and Ford, 2018), making the detection, generation, and explanation of humor hard for humans and machines (Shafiei and Saffari, 2025; Loakman et al., 2025; Horvitz et al., 2024). Despite the broad and diverse nature of humor, much of existing work on computational humor has focused on narrow, specialized tasks such as detecting humor in internet memes (Kumari et al., 2024), knock-knock jokes (Taylor and Mazlack, 2004), puns (Xu et al., 2024; Miller et al., 2017; Cocchieri et al., 2025), cartoons (Shahaf et al., 2015) or even "That's what she said" jokes (Kiddon and Brun, 2011), but relatively little attention has been paid to learning the *general* phenomenon of humor.

In this work, we are interested in **whether competence on one or more specific humor tasks confers any ability to transfer to novel, unseen types**. That is, we wish to understand whether splitting humor into subproblems is a necessary design choice or a historical artifact. This is especially important as new humor variants emerge over time; should we expect LLMs to understand them without further training, as humans often do?

Researchers from neuroscience and psychology have studied whether skill in one type of humor category aids another *in humans*, and the evidence is mixed: Findings suggest shared mechanisms (e.g., incongruity resolution) that provide some common ground across joke types and may enable partial transfer, but also specialized skills (language ambiguity, theory-of-mind, cultural knowledge) that are type-specific and create "transfer costs" (Dai et al., 2017; Farkas et al., 2021) (see Section 7).

In NLP, several studies have explored transfer between different types of humor or languages (Arora et al., 2022; Baranov et al., 2023; Wang et al., 2020). There is some evidence that multi-category training helps humor detection, but these works did not test on humor types that were not a part of training, making it difficult to assess true generalization to novel types of humor. Moreover, they rarely discuss the relations between humor types.

In this work we experiment with four humor-related datasets, representing different types of hu-

[1]https://github.com/morturr/HumorTransferLearning.git

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 1–28

Name	Text Length Mean & Std	Positive Example	Negative Example
Amazon Questions	143.80 ± 58.64 *60.41 ± 37.21	PEREGRINE Banana Saver Yellow | I have a problem with wolves where I live. Will this carrier protect bananas from wolves?	Two-Wheel Smart Scooter Self Balancing Unicycle Electric Scooter Electric Unicycle Smart Wheel | Do you ship by ups or fedex?
Reddit Dad Jokes	86.10 ± 26.25	I went to a bookstore and asked where the self-help section was The clerk said that if she told me, it would defeat the purpose.	Did you hear of the Librarian who became unwell while reading a book? She had to take a sick leave.
Sarcasm Headlines	62.45 ± 21.10	new york introduces shoe-sharing program for city's pedestrians	stars with gray hair prove getting older isn't all that bad
One Liners	60.66 ± 18.44	Couldn't afford to fix my brakes, so I made my horn louder.	Fear a silent man. He has lips like a drum.

Table 1: Properties of the datasets used in our experiments: mean and std of text length, and example sentences from the positive (humorous) and negative (non-humorous) classes. **See Appendix B for more examples.** *For Amazon Questions, we report both the length of the full input (product name + question) and the question alone.

mor. We selected two advanced models, LLaMA-2-7B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023), specifically chosen because they demonstrated poor performance on these datasets in a zero-shot setting; this allowed us to evaluate their potential for improvement through transfer. Our contributions are:

- We present the first systematic evaluation of humor transfer learning across multiple humor types using LLMs.
- We analyzed the models' performance in single and multi-task humor binary classification settings. We find that models are capable of some transfer, with Mistral achieving 75% accuracy on an unseen dataset. **Training on diverse data enhances transfer to unseen types**. In-domain performance remains relatively stable as the training set diversity increases, even as the number of training examples from the domain decreases significantly.
- We discover that certain humor types (e.g., Dad Jokes) more effectively enable transfer to others, suggesting latent structural relations.
- We propose a framework for studying humor transfer, including developing a method to generate negative (non-funny) examples for datasets that includes only positive ones.
- We make our code and data public[1].

2 Research Questions

We investigate the capacity of LLMs to perform transfer learning across different types of humor via the following research questions (RQs):

- **RQ1**: Do LLMs have the capability for humor transfer learning? Can they learn some type(s) of humor and generalize to new humor types?
- **RQ2**: Between which types of humor is transfer most effective?
- **RQ3**: How does data diversity influence humor transfer learning? Does training on more diverse datasets (e.g., containing multiple humor types) enhance generalization?

3 Data

To answer these RQs, we experiment with four humor datasets, each targeting a distinct humor task. Table 1 provides representative examples from both the humorous and non-humorous classes of all the datasets, and mean text lengths (more in Appendix B). The datasets differ in style, domain, and structure, providing a diverse testbed for generalization (see Appendix L for syntactic analyses).

3.1 Dataset Descriptions

Amazon: 19K records, each containing a product name paired with a user-submitted question, annotated for humor by humans (Ziser et al., 2020).

One Liners: 32K one-liner sentences (Mihalcea and Strapparava, 2005). Humorous examples were collected using an algorithm designed to harvest funny one-liners. Non-humorous examples were sourced from news headlines, proverbs, and sentences from the British National Corpus.

Sarcasm Headlines: 28K headlines consisting of both real news headlines and sarcastic ones from *The Onion*, a satirical news outlet (Misra and Arora, 2023).

Reddit Dad Jokes: Reddit posts collected from the `r/dadjokes` subreddit (Reddit, 2023). For the positive class, we selected high-confidence positive samples (Reddit score ≥ 20).

As the original dataset only includes positive (i.e., humorous) examples, we needed to generate negative samples. To create negative examples that closely match the positive class in content, writing style, and semantics, we used GPT-4 Turbo (OpenAI, 2023) in a few-shot setup. For the samples with lower Reddit score, we asked GPT to minimally modify each joke, preserving style and content but removing the humorous element. E.g., given "Why can't milk cartons walk? Because they lactose," the generated negative was "Why can't milk containers move? They lack appendages." (see prompt in Appendix A).

To assess generation quality, we manually reviewed 3,000 outputs. Only 2.63% did not maintain the style or content (typically due to the LLM summarizing the unfunny joke). Importantly, we found no examples where the punchline was retained. To ensure balanced text length distributions between both classes of Dad Jokes, we paired each negative example with the closest-length positive example, ensuring no duplicates.

3.2 Dataset Properties and Preprocessing

Humor Styles. The datasets span a variety of humor styles: While there is some natural overlap in humor types, questions in the Amazon dataset are often sarcastic or ironic, News Headlines feature more sophisticated or absurd humor that often borders on non-sequitur (and requires some knowledge about current events). One-Liners and Dad Jokes both include brief, standalone jokes, frequently employing puns or wordplay; dad jokes also includes many short stories (see Appendix B for specific examples that illustrate these differences; full data will be published upon acceptance).

Data Partitioning. For fair cross-dataset comparisons, we randomly downsampled all datasets to 6,250 examples. Each dataset was balanced between positive and negative classes. We used an 80%/2%/18% training/validation/test split (to balance evaluation reliability with efficiency, given the large number of trained models and evaluations).

In transfer experiments from multiple datasets, we constructed the training set by sampling equally from each dataset, ensuring class balance and maintaining a consistent training size of 5,000 (2,500 positive and 2,500 negative datapoints).

Dataset Distinctness. To address concerns about overlap between the datasets that could inflate transfer results, we performed domain classification. We trained models on a 4-class task using 5k samples (equal amount of samples from each dataset), with the goal of correctly identifying the originating dataset for each sample. We repeated the experiment 3 times, using only positive samples, only negative samples, and both positives and negatives. Both models achieved 98-100% accuracy in all settings. This indicates that the datasets **have clear enough differences, while still allowing models to leverage shared patterns across domains**. See Appendix K for t-SNE visualizations.

4 Experiments

We conducted a suite of experiments to examine LLM humor transfer, data diversity effects, and humor type differences (see RQs in Section 2). Figure 1 depicts our three experimental setups.

- **Single Dataset Training:** Examines whether basic transfer occurs between datasets (RQ1) and whether certain datasets are more similar to others (RQ2). Models were fine-tuned on each dataset and evaluated on all datasets.
- **Double Dataset Training:** Explores multi-task learning by examining how training on two humor styles affects both in-domain performance and generalization to other styles (RQ3), and further investigates the relationships between different humor types (RQ2). Models were fine-tuned on each pair of datasets and evaluated on all datasets.
- **Triple Dataset Training:** Examines the effect of training in the most diverse data setting to evaluate how data diversity impacts transferability (RQ3), and to identify which humor types are most generalizable from others (RQ2). Models were trained on three datasets and evaluated on all datasets.

Models. For all experiments, we used both LLaMA-2-7B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023). We selected them for two main reasons: (1) they are comparable in size but belong to different families, to assess whether transferability trends are consistent across architectures; (2) both models are widely used and have demonstrated strong performance on various NLP tasks but performed near guess-level (40%-56%) in the zero-shot setting on our tasks (Appendix C).

Through our exploration, we found that simpler models (e.g., non-LLM) struggled to capture the nuances of the task, while larger models performed too well in the zero-shot setting. The rationale behind the choice of models was to ensure that

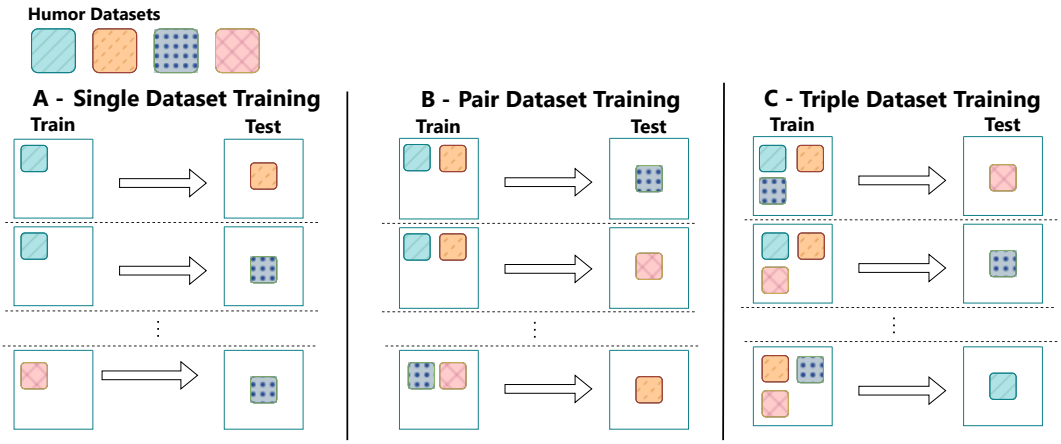

Figure 1: Overview of the experimental setups. **(A) Single Dataset Training:** Train on one dataset and test on each of the other datasets. **(B) Double Dataset Training:** Train on two of datasets and test on each of the remaining three. **(C) Triple Dataset Training:** Train on three datasets and test on the held-out forth dataset.

the models would be able to learn, and that **any observed improvement would be a result of effective transfer learning, rather than the model already being familiar with the task**.

We applied instruction fine-tuning to the models and used a prompt describing the task for each training sample (see Appendix D). Full training and evaluation details are provided in Appendix E.

5 Results and Analysis

We now present our results, addressing each RQ in turn. Accuracy scores are summarized in Tables 2-4. See Appendix for STDs, Confidence Intervals, F1, Recall, and Precision scores (Tables 8-12).

5.1 RQ1: Transfer Humor Capability

LLMs can transfer humor knowledge across datasets, but success varies by model and humor type. To assess whether LLMs can perform humor transfer learning, we examine the results from the Single Dataset Training (Table 2). Both LLMs are able to learn the humor style they were trained on (in-domain), but they differ in their ability to generalize to unfamiliar humor types (transfer).

LLaMA-2 shows weaker performance overall, with in-domain accuracy averaging 4.75% lower than Mistral's. Its cross-dataset performance is ∼60% in most cases, exhibiting some transfer. Mistral demonstrates better transfer, reaching 67-75% accuracy on several target datasets, particularly when trained on Amazon or Dad Jokes.

5.2 RQ2: Linking Humor Types

Humor datasets differ in transferability. We investigate which datasets enable the most effective transfer across three experimental setups.

In the single-dataset-training experiment (Table 2) we focus on Mistral, given its superior performance. Training on Amazon leads to relatively high transfer accuracy on Headlines and One Liners (72-75%). The reverse direction yields lower performance (64-65%). This suggests that Amazon may support broader generalization, perhaps due to its coverage of a wide range of humor styles and topics. In contrast, Headlines and One Liners are more structurally constrained and stylistically homogeneous, which may limit their transferability.

Dad Jokes shows the most asymmetric pattern: training on it yields strong transfer accuracy (68-71%), but models trained on other datasets perform poorly on Dad Jokes (51-62%). We note that it includes multi-sentence narratives, puns, irony, and cultural references, which are not easily captured by shorter or more templatic humor styles.

Finally, One Liners and Headlines show relatively strong bidirectional transfer, likely due to their shared brevity and stylistic similarity. Both rely on compact, punchline-driven formats and often draw from news-style or everyday language, which may facilitate mutual generalization.

We analyze the pairwise training results in Table 3, computing average transfer accuracy for each target dataset by averaging model performance across all training pairs that exclude it. Both LLMs exhibit similar trends across most datasets, mirror-

Train Dataset	Model	Test Dataset			
		Amazon	Dad Jokes	Headlines	One Liners
Amazon	LLaMA-2	88	65	65	61
	Mistral	91	62	75	72
Dad Jokes	LLaMA-2	63	93	59	70
	Mistral	69	94	68	71
Headlines	LLaMA-2	58	57	90	65
	Mistral	65	56	97	62
One Liners	LLaMA-2	62	62	54	87
	Mistral	64	51	67	95

Table 2: **[Partial transfer between humor datasets.] Single Dataset Training:** Accuracy scores (0-100), averaged over four training seeds. Models trained on a single dataset and evaluated on all. Diagonal values reflect in-domain performance, showing that both models effectively learn their respective datasets. Off-diagonal values capture transfer performance, revealing asymmetric transfer patterns. For example, Dad Jokes transfers well to One Liners (70-71%), but not vice versa (51-62%). Mistral consistently shows stronger transfer than LLaMA-2, especially when trained on Amazon and Dad Jokes. See Appendix Table 11 for standard deviations.

Two Train Datasets	Model	Test Dataset			
		Amazon	Dad Jokes	Headlines	One Liners
Amazon + Dad Jokes	LLaMA-2	82	90	67	69
	Mistral	89	95	74	74
Amazon + Headlines	LLaMA-2	82	62	90	69
	Mistral	89	67	96	67
Dad Jokes + Headlines	LLaMA-2	63	89	92	71
	Mistral	71	91	95	70
Dad Jokes + One Liners	LLaMA-2	74	90	65	91
	Mistral	66	95	70	94
Headlines + One Liners	LLaMA-2	63	55	93	92
	Mistral	68	52	97	94
One Liners + Amazon	LLaMA-2	86	56	65	91
	Mistral	90	53	74	94

Table 3: **[Amazon + Dad Jokes generalizes best.] Double Dataset Training:** Accuracy scores (0-100), averaged over four seeds. Models were trained on two datasets. Amazon + Dad Jokes yields the strongest transfer (74% Mistral; 67-69% LLaMA-2), while Headlines + One Liners yields the weakest transfer (52-55% on Dad Jokes; 63-68% on Amazon). Mistral outperforms LLaMA-2 in most cases. See Appendix Table 11 for standard deviations.

Left-Out Dataset	Model	Test Dataset			
		Amazon	Dad Jokes	Headlines	One Liners
Amazon	LLaMA-2	69	88	89	88
	Mistral	66	94	96	94
Dad Jokes	LLaMA-2	84	57	90	87
	Mistral	90	55	96	94
Headlines	LLaMA-2	86	89	68	88
	Mistral	90	95	73	93
One Liners	LLaMA-2	84	89	91	69
	Mistral	90	96	96	74

Table 4: **[Training on limited data preserves self accuracy.] Triple Dataset Training:** Accuracy scores (0-100), averaged over four seeds. Models were trained on three datasets (excluding the one listed in the "Left Out Dataset" column), using 33% of each. Strong seen-dataset accuracy shows robust learning; partial transfer is observed on the left-out dataset (e.g., Mistral reaches 73% on Headlines, 74% on One Liners). See Appendix Table 11 for STDs.

ing the earlier conclusion that **Dad Jokes is the one supporting the broadest transfer, whereas One Liners and Headlines are more learnable from other humor types** (Table 3). For every target dataset, the strongest transfer is obtained when training pairs include Dad Jokes. This strengthens the conclusion from single-dataset-training.

Conversely, when Dad Jokes is the target, both LLMs perform best when trained on Amazon + Headlines (62-67%), outperforming combinations containing One Liners (52-56%). This mirrors the weak One Liners → Dad Jokes transfer observed in the single-source setting and further illustrates the strong asymmetry between these humor datasets.

We now examine transfer performance in the triple-dataset experiment, where each dataset is held out once for evaluation (Table 4). As in the previous experiments, both models struggle to transfer to Dad Jokes, with an average accuracy of 56%. In contrast, Headlines and One Liners show the strongest transfer results, averaging 70.5% and 71.5% respectively, followed by Amazon, which demonstrates slightly weaker transfer with an average of 67.5%. These findings are consistent with the trends observed in the single- and double experiments, where Dad Jokes emerged as the most difficult for transfer learning, and One Liners and Headlines were the most receptive to transfer.

Conclusions. Taken together, the experiments **reveal a hierarchy of humor transferability:** Dad Jokes enables strong transfer to all but remains difficult to generalize to. Amazon occupies a middle ground, benefiting from Dad Jokes while transferring reasonably well. Headlines and One Liners are the most generalizable targets, but offer comparatively less utility when used as sources for transfer.

These findings hint at the idea that datasets covering more styles and topics could support broader transfer. Another hypothesis is that the strong performance of Dad Jokes is due to its construction of negative samples. As noted above, the negatives are minimal modifications of funny jokes, preserving style and content but removing the humorous element; this might have forced the models trained on this data to disentangle the key features of humor from other superficial confounding variables.

5.3 RQ3: Impact of Data Diversity

We now explore how the diversity of humor *training* data affects model performance. Specifically, whether exposure to multiple humor styles improves generalization across domains, and whether

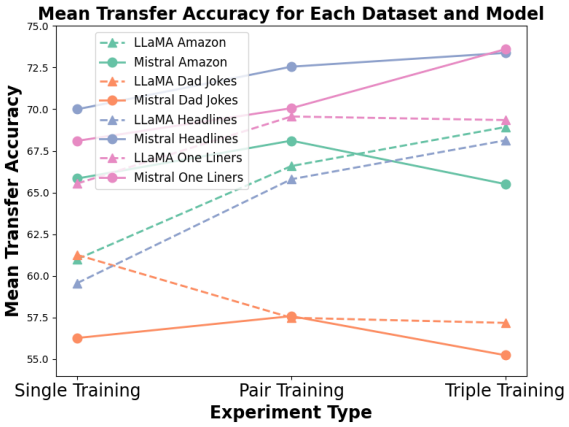

Figure 2: **[Increasing the training data diversity improves transfer.] Comparing transfer across the experiments.** Mistral results are shown with solid lines, LLaMA-2 with dotted lines. Colors represent different test datasets. In general, more diverse training data leads to better transfer than single-dataset training. LLaMA-2 shows consistent improvement across experiments, except on Dad Jokes. Notably, Dad Jokes is the only dataset that performs worse with increased diversity.

it depends on humor type or evaluation setting. We note that we discuss the average case in our experiments; of course, adding a dataset that is similar to the target could affect transfer dramatically.

We first assess the overall impact of data diversity on humor transfer learning. Next, we investigate how different humor types respond to diversity during training, identifying styles that benefit more or less from multi-source input. We then evaluate how data diversity influences in-domain accuracy (that is, performance on humor styles included in training). Finally, we compare the effects of diversity between Mistral and LLaMA-2 to understand whether model architecture or pretraining background modulates these trends.

5.3.1 Impact of Data Diversity on Transfer

Greater training diversity improves humor transfer, but with diminishing returns. To investigate how training data diversity affects humor transferability, we compare our three experiments: (A) single-dataset training (no diversity), (B) double-training (moderate diversity), and (C) triple-training (high diversity). For each setup, we evaluate on a held-out dataset and average performance across runs that exclude the evaluated dataset from training. Figure 2 depicts transfer performance across the three experiments (see non-aggregated results in Appendix I).

For LLaMA-2, we observe a consistent improve-

ment in transfer accuracy with increasing diversity. Moving from single- to double-dataset training yields an average gain of 3.02 percentage points across target datasets, with further gains of 1.04 points from double to triple. The only exception is Dad Jokes, which shows the opposite trend, arguably due to its relative high diversity.

Mistral follows a similar trend from single-to-double-dataset training, improving by 2.02 points on average. However, it plateaus in triple-training, with a slight drop of 0.14 points compared to double, though still 1.88 higher than single-training.

Training on diverse humor sources consistently improves models' generalization to unseen humor types. The largest gains come from moving beyond single-dataset training; returns begin to diminish as more datasets are added, suggesting that **moderate diversity may be nearly as effective as maximal diversity for cross-domain humor transfer**.

5.3.2 Transfer Between Distinct Humor Types

Data diversity yields larger gains for structurally simpler humor types. We now focus on humor transfer across *humor types*. Headlines and One Liners consistently benefit from increased diversity, with both LLMs showing improved accuracy from single- to double- to triple-dataset training. A minor exception occurs for LLaMA on One Liners, where double- slightly outperforms triple-training, though both exceed single baseline.

For Amazon, LLaMA's accuracy improves substantially when increasing diversity. In contrast, for Mistral, double-dataset-training yields the highest accuracy, while the triple performing slightly worse than single-dataset baseline. Dad Jokes displays a decline in performance as data diversity increases.

These findings suggest that increasing training data diversity is more beneficial for generalizing to structurally simpler humor tasks.

5.3.3 Data Diversity & In-Domain Accuracy

Less in-domain training data leads to only a small drop in in-domain performance. We evaluate the impact of reduced dataset specific training data on in-domain performance. For each dataset, we compare its single-dataset training accuracy to the average accuracy of the three models trained on it within a triple-dataset setup (Tables 2, 4).

In most cases, in-domain accuracy decreases under the triple-training setup due to the reduced amount of in-domain data (just 33% compared to single-dataset training). However, the average accuracy of Mistral drops by only 0.49 percentage points, and LLaMA-2 by 1.76. These results suggest that both models are relatively **robust to reductions in in-domain data, with only minimal performance loss** even when training data is substantially diluted. In rare cases, the triple setup led to small performance gains.

6 Discussion

Prioritizing Diversity in Humor Learning. As our results show, increasing data diversity generally enhances humor transfer learning, echoing trends observed in other NLP tasks (Yu et al., 2022; Wang et al., 2022; Rozen et al., 2019). Reducing the amount of in-domain data to 33% led to only a slight decrease in accuracy. This could mean our datasets were large enough to retain performance despite downsampling, or that future humor applications should prioritize data diversity over size (similar to recent work highlighting the importance of diversity in synthetic datasets (Zhu et al., 2025; Long et al., 2024; Chen et al., 2024)).

Impact of Dad Jokes' Construction. We hypothesize that one reason models trained on Dad Jokes generalize so effectively to other humor types (but not vice versa) is how its negative samples were crafted to syntactically and topically resemble jokes, but without the punchline (also supported by the high self-similarity of Dad Jokes in embedding space, see below). This design encourages models to focus on humor, as they could not rely on superficial, dataset-specific signals. This insight suggests a new strategy for building humor datasets with strong transfer potential.

Comparing Mistral and LLaMA-2. While both capture humor and exhibit transfer, they show distinct strengths. Notably, Mistral consistently outperforms LLaMA-2 in transfer settings, suggesting a superior ability to learn shared stylistic or structural features across humor types.

Despite differences in performance, both models exhibit consistent patterns: (1) Dad Jokes reliably support transfer to other datasets but remain hard to generalize to; (2) Headlines and One Liners are easy to generalize to but offer limited transfer benefit; (3) Amazon occupies a middle ground, both benefiting from and contributing to transfer; and (4) data diversity particularly benefits structurally simpler humor types.

The alignment across models supports the conclusion that humor transfer is governed by struc-

tured, humor-type-specific patterns. Still, performance gaps such as Mistral achieving nearly 70% accuracy in settings where LLaMA-2 falls short, raise questions about whether the extent of transferability is intrinsic to humor or dependent on model architecture. These discrepancies highlight important directions for future research.

Dataset Similarity. To better understand dataset relations, we computed pairwise cosine similarity between sentence embeddings from the training split, both within and across datasets. See Appendix K. We used pretrained Mistral, as it produced more robust results. Surprisingly, the most structurally complex datasets, Dad Jokes and Amazon, were also the most self-similar, while One Liners was the least. Notably, higher cross-dataset similarity is correlated with stronger observed transfer.

7 Related Work

Humor Taxonomy. Linguistic and psychological theories provide rich taxonomies of humor. Tsakona (2017) distinguishes humor types by contextual expectations, while Dynel (2009) categorize forms like irony, puns, and allusions. The Humor Styles Questionnaire (Martin et al., 2003) defines humor styles with distinct social functions. Prior work has not systematically examined relationships between humor types or their potential for transfer.

Humor Transfer in Humans. Neuroimaging studies confirm that different humor types recruit distinct brain systems, implying limited transfer. For example, complex semantic jokes activate language areas, whereas sound-based puns engage speech-processing regions (Martin and Ford, 2018). Dai et al. (2017) showed that resolvable and unresolvable humor involve different neural paths in all stages. An fMRI meta-analysis found that humor engages broad language and reward circuits regardless of stimulus type, but ToM-based humor specifically activates mentalizing areas (Farkas et al., 2021). Developmental studies corroborate these results (Angeleri and Airenti, 2014; Yankovitz et al., 2023). Together these findings suggest shared mechanisms providing common ground, but also specialized type-specific skills.

Computational Humor. Humor recognition is subjective, context- and culture-sensitive. Kalloniatis and Adamidis (2024) reviewed the complexity of humor datasets and models. Multimodal, cross-lingual, and application-oriented studies (Xie et al., 2023; Shani et al., 2022; Shapira et al., 2023) further highlight the field's breadth.

Transfer Learning. From a machine learning perspective, our work builds on the foundation of transfer learning and multi-task learning (MTL) (Zhuang et al., 2021; Zhang and Yang, 2021).

Humor Transfer in LLMs. Prior MTL work on humor focused on joint training rather than transfer. Arora et al. (2022) used a shared-private model to capture general and type-specific humor features but did not test generalization to unseen types. Baranov et al. (2023) explored transfer by training on multiple humor datasets and evaluating on overlapping subsets, finding that diversity aids generalization. In contrast, we are interested in transfer to entirely unseen datasets. Wang et al. (2020) tackled multilingual tasks but did not explore transfer across languages. Loakman et al. (2025) investigated the ability of LLMs to explain jokes across different humor types, finding that no LLMs could generate adequate explanations for all types.

8 Conclusions

In this work, we asked whether competence on specific humor types enables generalization to novel styles. We found that **humor transfer is possible but asymmetric**: types like Dad Jokes support transfer but are hard to generalize to, while Headlines and One Liners are easier to predict but contribute little to transfer.

Exposure to diverse humor types generally improves performance, particularly for structurally simpler styles. While both LLaMA-2 and Mistral capture broad transfer patterns, Mistral consistently generalizes better. Interestingly, models retain strong in-domain performance even when trained on only 33% of target data.

Future work should expand to more humor types and modalities (e.g., cartoons or internet memes) and explore different axes of transfer, such as multilingual or cross-culture settings, as well as seek alignment between transfer patterns and findings and theories from cognition and neuroscience. Another possible extension would be to non-parametric learning, such as few-shot and in-context learning. We hope this work inspires follow-up work that could further illuminate what makes humor transferable in machines and in minds.

9 Limitations

This study has several limitations. First, we focus exclusively on the binary classification of short-form, English-language textual humor, omitting any reference to the ability of LLMs to rate the funniness of text. This ability would be valuable in interactive contexts such as dialogue or conversational humor. Additionally, this scope inherently excludes multimodal formats (e.g., memes, videos). Second, while each dataset serves as a stand-in for a particular humor style, these assignments are approximate and do not capture the full nuance or variability of humor genres. Moreover, individual datasets may reflect specific demographics, cultural biases, or artifactual noise. For instance, the non-humorous Dad Jokes samples were generated by ChatGPT; thus, the transformation was necessarily shaped by a more capable LLM's implicit understanding of humor. That influence may not have been uniform across joke types, potentially introducing systematic differences between categories that could partly explain the observed variance in transfer performance. Third, our analysis is based on a limited set of datasets (four) and models (Mistral-7B and LLaMA-2-7B), and relies heavily on transfer learning. This methodological scope may constrain the generalizability of our findings, as there may be other methods to capture the deeper mechanisms of humor, such as utilizing LLMs within a hybrid neural-symbolic system. Future work could address these limitations by incorporating continuous funniness ratings, joke explanation, alternative computational architectures, and a broader range of humor types and languages to better capture the richness and diversity of humorous expression.

10 Ethical considerations

Some of the datasets used in this study were collected from publicly available internet sources and may contain offensive content. Humor, by nature, often challenges social norms, and internet discourse can occasionally reflect inappropriate or sensitive material. However, a brief examination of the datasets revealed no indications of content that exceeds the bounds of good taste. We used the datasets as-is to preserve their original characteristics, which are essential for analyzing humor in natural contexts.

The datasets do not contain personally identifiable information, with the exception of the Reddit Dad Jokes dataset, which may include usernames. We note that this information is public (on Reddit), and we did not make any use of it in our work.

All datasets were used in accordance with their respective terms of use and solely for academic research purposes.

References

Romina Angeleri and Gabriella Airenti. 2014. The development of joke and irony understanding: a study with 3-to 6-year-old children. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 68(2):133.

Aseem Arora, Gaël Dias, Adam Jatowt, and Asif Ekbal. 2022. Transfer learning for humor detection by twin masked yellow muppets. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1–7.

Salvatore Attardo. 2000. Irony as relevant inappropriateness. *Journal of pragmatics*, 32(6):793–826.

Salvatore Attardo. 2024. *Linguistic theories of humor*, volume 1. Walter de Gruyter GmbH & Co KG.

Sayak Autrin et al. 2023. Peft: Parameter-efficient fine-tuning. https://github.com/huggingface/peft. Accessed: 2025-07-20.

Alexander Baranov, Vladimir Kniazhevsky, and Pavel Braslavski. 2023. You told me that joke twice: A systematic investigation of transferability and robustness of humor detection models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13701–13715, Singapore. Association for Computational Linguistics.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 2024. On the diversity of synthetic data and its impact on training large language models. *Preprint*, arXiv:2410.15226.

Alessio Cocchieri, Luca Ragazzi, Paolo Italiani, Giuseppe Tagliavini, and Gianluca Moro. 2025. "what do you call a dog that is incontrovertibly true? dogma": Testing LLM generalization through humor. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22922–22937, Vienna, Austria. Association for Computational Linguistics.

Ru H Dai, Hsueh-Chih Chen, Yu C Chan, Ching-Lin Wu, Ping Li, Shu L Cho, and Jon-Fan Hu. 2017. To resolve or not to resolve, that is the question: The dual-path model of incongruity resolution and absurd verbal humor by fmri. *Frontiers in psychology*, 8:498.

Marta Dynel. 2009. Beyond a joke: Types of conversational humour. *Language and Linguistics Compass*, 3(5):1284–1299.

Andrew H Farkas, Rebekah L Trotti, Elizabeth A Edge, Ling-Yu Huang, Aviva Kasowski, Olivia F Thomas, Eli Chlan, Maria P Granros, Kajol K Patel, and Dean Sabatinelli. 2021. Humor and emotion: Quantitative meta analyses of functional neuroimaging studies. *Cortex*, 139:60–72.

Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models. *Preprint*, arXiv:2403.00794.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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

Antonios Kalloniatis and Panagiotis Adamidis. 2024. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2):43.

Chloe Kiddon and Yuriy Brun. 2011. That's what she said: double entendre identification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 89–94.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and Vinutha BN. 2024. Let's all laugh together: A novel multitask framework for humor detection in internet memes. *IEEE Transactions on Computational Social Systems*, 11(3):4385–4395.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Comparing apples to oranges: A dataset & analysis of llm humour understanding from traditional puns to topical jokes. *Preprint*, arXiv:2507.13335.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Rod A Martin and Thomas Ford. 2018. *The psychology of humor: An integrative approach*. Academic press.

Rod A. Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of Research in Personality*, 37(1):48–75.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Human Language Technology - The Baltic Perspectiv*.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

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

OpenAI. 2023. New models and developer products announced at dev day. Accessed: 2025-07-20.

Victor Raskin. 1979. Semantic mechanisms of humor. In *Annual Meeting of the Berkeley Linguistics Society*, pages 325–335.

Reddit. 2023. Reddit dad jokes. https://www.kaggle.com/datasets/oktayozturk010/reddit-dad-jokes/data.

Ohad Rozen, Vered Shwartz, Roee Aharoni, and Ido Dagan. 2019. Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 196–205, Hong Kong, China. Association for Computational Linguistics.

Mohammadamin Shafiei and Hamidreza Saffari. 2025. Not all jokes land: Evaluating large language models understanding of workplace humor. *arXiv preprint arXiv:2506.01819*.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1065–1074.

Chen Shani, Alexander Libov, Sofia Tolmach, Liane Lewin-Eytan, Yoelle Maarek, and Dafna Shahaf. 2022. "alexa, do you want to build a snowman?" characterizing playful requests to conversational agents. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7.

Natalie Shapira, Oren Kalinsky, Alex Libov, Chen Shani, and Sofia Tolmach. 2023. Evaluating humorous response generation to playful shopping requests. In *European Conference on Information Retrieval*, pages 617–626. Springer.

Jerry M Suls. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The psychology of humor: Theoretical perspectives and empirical issues*, 1:81–100.

Julia M Taylor and Lawrence J Mazlack. 2004. Computationally recognizing wordplay in jokes. In *Proceedings of the annual meeting of the cognitive science society*, volume 26.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Villy Tsakona. 2017. Genres of humor. In *The Routledge handbook of language and humor*, pages 489–503. Routledge.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *Preprint*, arXiv:2103.03097.

Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. 2020. Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd annual conference of the European association for machine translation*, pages 53–59.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Heng Xie, Jizhou Cui, Yuhang Cao, Junjie Chen, Jianhua Tao, Cunhang Fan, Xuefei Liu, Zhengqi Wen, Heng Lu, Yuguang Yang, et al. 2023. Multimodal cross-lingual features and weight fusion for cross-cultural humor detection. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, pages 51–57.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. " a good pun is its own reword": Can large language models understand puns? *arXiv preprint arXiv:2404.13599*.

Bat-el Yankovitz, Anat Kasirer, and Nira Mashal. 2023. The relationship between semantic joke and idiom comprehension in adolescents with autism spectrum disorder. *Brain Sciences*, 13(6):935.

Yu Yu, Shahram Khadivi, and Jia Xu. 2022. Can data diversity enhance learning generalization? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *Preprint*, arXiv:1707.08114.

Yuchang Zhu, Huazhen Zhong, Qunshu Lin, Haotong Wei, Xiaolong Sun, Zixuan Yu, Minghao Liu, Zibin Zheng, and Liang Chen. 2025. What matters in llm-generated data: Diversity and its effect on model fine-tuning. *Preprint*, arXiv:2506.19262.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

Yftah Ziser, Elad Kravi, and David Carmel. 2020. Humor detection in product question answering systems. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

A GPT-4-Turbo Prompt for Generating Dad Jokes

###
1. input: 'A grizzly kept talking to me and annoyed me He was unbearable'
output: 'A grizzly kept talking to me and annoyed me He was intolerable'
2. input: 'For Christmas, I requested my family not to give me duplicates of the same item. Now I anticipate receiving the missing sock next time.'
output: 'For Christmas, I requested my family not to give me duplicates of the same item. Now I anticipate receiving the other book next time.'
3. input: 'My son's fourth birthday was today, but when he came to see me I didn't recognize him at first. I'd never seen him be 4.'
output: 'My son's fourth birthday was today, but when he came to see me I didn't recognize him at first. He grew up so fast.'
4. input: 'I asked my friend if he liked Nickleback. He told me that he never gave me any money'
output: 'I asked my friend if he liked Nickleback. He told me that he prefers Kings of Leon.'
5. input: 'I went to a bookstore and asked where the self-help section was The clerk said that if she told me, it would defeat the purpose.'
output: 'I went to a bookstore and asked where the self-help section was The clerk said it was in the third aisle .'
###
Using the examples in ### markers, please change some of the words in the following sentences to make them non humorous. You can change anything but please change the least you can:

B More Dataset Examples

Tables 5, 7 shows additional examples from each dataset of positive (humorous) and negative (non-humorous) samples, respectively.

Name	Positive (Humorous) Examples
Amazon Questions	Colore ProVisible Graphite Transfer Artist Paper 9x13 - Boldly Create Art With Reusable & Erasable Carbon (50 Sheets) ||| If i buy this product will i look bored and harassed like the gal in the product photos?
	Super Smash Bros. Ultimate ||| Will this fix my marriage?
	This Works Deep Sleep Bath Oil 100ml ||| Does a person come with this $395 purchase? And does that person provide warm milk and hum lullabies until I am asleep?
	Allntrends Harry Style Sweatshirt Tattoo One Direction Shirt Tee ||| why does this exist? is there even a god? are we alone in this world?
Reddit Dad Jokes	My wife told me to sync her phone.. I threw it in the ocean and I don't know why she's mad at me
	There is a mysterious crime spree going on at our local IKEA. The cops are having a hard time putting the pieces together.
	My orchestra buddy wanted to bring his fiddle to a protest. I told him not to. In a peaceful protest, there's no need for violins.
	Vegans must think we meat eaters are gross. In our defence, a person who sells vegetables is grocer.
Sarcasm Headlines	car passengers launch urgent, mid-street investigation into whether woman in parking spot coming or going
	presidential debate commission anesthetizes audience to prevent outbursts during debate
	ex-con still hanging out with hallucinatory voices that got him in trouble in first place
	30th anniversary of 1973 commemorated
One Liners	If white wine goes with cooked fish, do white grapes go with sushi?
	Experiments should be reproducible - they should all fail in the same way.
	Cleaning your house while your kids are at home is like trying to shovel the driveway during a snowstorm.
	I've been on so many blind dates, I should get a free dog.

Table 5: Positive (Humorous) examples from the datasets used in our experiments.

Model	Test Dataset			
	Amazon	Dad Jokes	Headlines	One Liners
LLaMA-2	55	56	56	49
Mistral	51	55	40	50

Table 6: **Zero-Shot Performance across Humor Datasets.** Accuracy (%) of LLaMA-2 and Mistral evaluated on the validation set of each dataset. Both models show limited humor detection ability in a zero-shot setting. *Models used are the base versions without instruction tuning.*

C Model Zero-Shot Performance

We conducted a zero-shot evaluation using both LLMs across the validation splits of all datasets. The models received the instruction prompt (excluding the actual response; see Appendix D) and produced outputs of either "Funny" or "Not funny," indicating that they understood the task format. Accuracy ranged from 40% to 56%, which is about guess level (see full results in Table 6). Thus, the transfer patterns observed in our experimental setup did not occur randomly but resulted from learning humor-specific information during training.

D Instruction Fine-Tuning Prompt

Below is an instruction that describes a sentiment analysis task.

Instruction:
Given the following text, please determine if it should be classified as funny or not funny. Base your classification on humor elements such as wit, irony, absurdity, or comedic timing.

Input:
{SAMPLE-TEXT}

Response:
{Yes/No}

E Training and Hyperparameter Selection Details

We conducted a systematic hyperparameter search using 4-fold cross-validation on each dataset. The search space was defined as the Cartesian product of the following values: learning rate {3e-4, 5e-5, 6e-5}, LoRA rank {32, 64, 128}, and LoRA alpha {8, 16, 32}, with a fixed seed of 42—resulting in 27 total combinations. To reduce computation time, we randomly sampled 10 configurations per dataset using random search. Each model was trained for 2 epochs.

The batch size was set to 4 when possible, and reduced to 2 for datasets with longer input sequences to fit within GPU memory limits. For each dataset, we selected the top 3 configurations based on median cross-validation accuracy across all evaluation datasets.

These configurations were used in the main experiments as follows:

- **Single Dataset Training (A):** Each of the top three configurations was trained with four random seeds (7, 18, 28, 42). The configuration with the highest median accuracy across all datasets was selected for final evaluation on test set.

- **Double Dataset Training (B):** The top three configurations from each dataset in the pair (six total) were each trained with four random seeds. The best-performing configuration was selected based on median accuracy.

- **Triple Dataset Training (C):** For each held-out dataset, we used the top 3 configurations from each of the three training datasets (nine total). Each configuration was trained with four random seeds, again selecting the configuration with the highest median accuracy.

For all experimental setups, we report the mean accuracy across four different training seeds.

In the training process we used LoRA (Hu et al., 2022) for efficiency. All models were trained on 5,000 examples (see Section 3). We ran all experiments using the Hugging Face `transformers` (Wolf et al., 2020) and `peft` (Autrin et al., 2023) libraries. Training was performed on a mix of `A6000`, `A40`, and `L40S` GPUs. Training and evaluation of all experiments took approximately 14 days in total.

E.1 Best-found Hyperparameters

We report the best training hyperparameters used for each model, selected based on highest median accuracy across evaluations. The parameters are listed in the following order: learning rate, LoRA rank, and LoRA alpha.

- Mistral (Amazon): 0.0003, 64, 32
- Mistral (Dad Jokes): 6.00E-05, 64, 8
- Mistral (Headlines): 6.00E-05, 64, 32
- Mistral (One Liners): 6.00E-05, 64, 32
- Mistral (Amazon + Dad Jokes): 0.0003, 64, 8
- Mistral (Amazon + Headlines): 0.0003, 64, 8
- Mistral (Dad Jokes + Headlines): 6.00E-05, 64, 8
- Mistral (Dad Jokes + One Liners): 0.0003, 128, 16
- Mistral (Headlines + One Liners): 0.0003, 128, 16
- Mistral (One Liners + Amazon): 0.0003, 64, 32
- Mistral (Leave Out Amazon): 0.0003, 128, 16

Name	Negative (Non-Humorous) Examples
Amazon Questions	GUESS Men's Hooded Puffer Jacket Black S ||| i am 180 height 64 kg which size ?
	Flexbow Kodiak 6041VX Tent with Free Ground Tarp ||| What are the shipping dimentions?
	Neenah Paper 09448 Classic (100%) Cotton Writing Paper 8-1/2 x 11 24-lb 500 Sheets/Ream ||| Are matching NO. 10 envelopes available?
	iRobot Roomba 770 Robotic Vacuum Cleaner ||| Is this dual voltage ? Can I use this outside of North America?
Reddit Dad Jokes	What did the windmill say when it encountered something it admired? "I support your work."
	A man approached me, holding a beer, and claimed he had a talent for voices. I was skeptical.
	Working late, I discovered some change on the ground. It was an unexpected find.
	I took my girlfriend out to dinner for our anniversary and she had high expectations, I tried to manage them.
Sarcasm Headlines	these stunning overhead beach photos are enough last you to next summer
	texas education board votes to create classes on mexican-american studies
	how to prevent screen addiction in your young children
	shared leadership among women and men: good news and bad news
One Liners	There will be an intensive service of trains all weekend .
	He who takes the child by the hand, takes the mother by the heart.
	Sri Lanka Ceramic in rapid turnaround.
	Do I have to spell out to you how important this is to me?

Table 7: Negative (Non-Humorous) examples from the datasets used in our experiments.

- Mistral (Leave Out Dad Jokes): 0.0003, 64, 32
- Mistral (Leave Out Headlines): 0.0003, 64, 16
- Mistral (Leave Out One Liners): 0.0003, 64, 32
- LLaMA-2 (Amazon): 0.0003, 64, 8
- LLaMA-2 (Dad Jokes): 0.0003, 128, 32 (Batch size = 4)
- LLaMA-2 (Headlines): 6.00E-05, 128, 8
- LLaMA-2 (One Liners): 5.00E-05, 32, 16 (Batch size = 4)
- LLaMA-2 (Amazon + Dad Jokes): 0.0003, 32, 32 (Batch size = 4)
- LLaMA-2 (Amazon + Headlines): 6.00E-05, 128, 32
- LLaMA-2 (Dad Jokes + Headlines): 0.0003, 128, 32 (Batch size = 4)
- LLaMA-2 (Dad Jokes + One Liners): 0.0003, 32, 32 (Batch size = 4)
- LLaMA-2 (Headlines + One Liners): 0.0003, 32, 32 (Batch size = 4)
- LLaMA-2 (One Liners + Amazon): 0.0003, 64, 8
- LLaMA-2 (Leave Out Amazon): 0.0003, 32, 32 (Batch size = 4)
- LLaMA-2 (Leave Out Dad Jokes): 0.0003, 32, 8
- LLaMA-2 (Leave Out Headlines): 0.0003, 64, 8
- LLaMA-2 (Leave Out One Liners): 0.0003, 64, 8

E.2 Packages and Configurations

We used several widely adopted libraries for modeling, preprocessing, training, and evaluation. Below, we report the key packages and configurations we used:

Transformers (Hugging Face) We used pre-trained models `mistralai/Mistral-7B-v0.1` and `meta-llama/Llama-2-7b-hf` via the `AutoModelForCausalLM` and `AutoTokenizer` interfaces. We set `tokenizer.pad_token = tokenizer.eos_token`. Models were loaded using 4-bit quantization via the `BitsAndBytesConfig` class, with the following settings:

- `load_in_4bit = True`
- `bnb_4bit_quant_type = "nf4"`
- `bnb_4bit_compute_dtype = torch.float16`

For generation-based evaluation, we used `model.generate()` with `max_new_tokens = 5`.

PEFT (LoRA) We fine-tuned models using parameter-efficient fine-tuning via the `LoraConfig` class from the `peft` library. The following hyperparameters were used:

- `lora_dropout = 0.1`
- `bias = "none"`
- `task_type = "CAUSAL_LM"`

TRL (SFTTrainer) We trained models using the `SFTTrainer` class from the `trl` library. The training configuration was based on Hugging Face's `TrainingArguments`, with the following relevant settings:

- `gradient_accumulation_steps = 4`
- `gradient_checkpointing = True`
- `max_seq_length = 1024`

Datasets Dataset processing and construction were handled using the Hugging Face `datasets` library. For train/test splits, we used `train_test_split` with the following parameters:

- `stratify_by_column = "label"`
- `seed = 42`

Scikit-learn We used `StratifiedKFold` from `sklearn.model_selection` for dataset partitioning. Cross-validation settings included:

- `n_splits = 4`
- `shuffle = True`
- `random_state = 1`

For evaluation, we used the following metrics from `sklearn.metrics`: `accuracy_score`, `precision_score`, `recall_score`, and `f1_score`, all calculated with `pos_label = 1`.

All relevant parameters, random seeds, and training configurations are documented in our code.

F Recall, Precision and F1-score

Tables 8, 9, 10 shows the F1-score, recall and precision scores of the different experiments.

G Standard Deviations Across Seeds

Table 11 reports standard deviations across four training seeds for all experiments and test datasets.

H Confidence Intervals

Table 12 reports 95% confidence intervals for accuracy scores across four training seeds for all experiments and test datasets.

I Transfer Accuracy Differences

Table 13 reports the change in mean transfer accuracy (in percentage points) between experimental setups for each model and target dataset.

J In-Domain Accuracy across Experiments

Figure 3 illustrates the mean in-domain accuracy across experiments for each model and dataset.

K Dataset Embeddings Similarity

Figure 4 shows the pairwise cosine similarity between datasets based on Mistral embeddings. To further explore these relationships, Figure 5 provides a t-SNE visualization of the embeddings across the different datasets. These visualizations illustrate the relative distinction between the humor types while highlighting the overlap between specific categories, such as Dad Jokes and One Liners, as discussed in Section 3.2.

L Syntactic Analysis

Figures 6, 7, 8 present different syntactic analyses of the datasets: Question Marks, Text Length and Word Count distribution respectively. Figures 9, 10, 11, 12 supply different Part-Of-Speech (POS) analyses across the four datasets.

M Use of AI Assistance

During the preparation of this work, the authors used ChatGPT, GitHub Copilot, and Google Gemini to assist with writing and coding. All content generated with these tools was reviewed and edited by the authors, who take full responsibility for the final publication.

Train Dataset	Model	Metric	Test Dataset			
			Amazon	Dad Jokes	Headlines	One Liners
Amazon	LLaMA-2	**F1-score**	0.88	0.72	0.69	0.68
		Recall	0.87	0.88	0.77	0.82
		Precision	0.89	0.61	0.63	0.58
	Mistral	**F1-score**	0.91	0.72	0.77	0.76
		Recall	0.9	0.99	0.87	0.9
		Precision	0.92	0.57	0.7	0.66
Dad Jokes	LLaMA-2	**F1-score**	0.71	0.93	0.68	0.69
		Recall	0.91	0.93	0.89	0.66
		Precision	0.58	0.93	0.56	0.72
	Mistral	**F1-score**	0.74	0.94	0.66	0.66
		Recall	0.9	0.95	0.62	0.56
		Precision	0.63	0.93	0.71	0.8
Headlines	LLaMA-2	**F1-score**	0.45	0.6	0.9	0.59
		Recall	0.36	0.64	0.91	0.51
		Precision	0.64	0.56	0.9	0.71
	Mistral	**F1-score**	0.65	0.56	0.97	0.62
		Recall	0.77	0.75	0.98	0.38
		Precision	0.63	0.54	0.96	0.72
One Liners	LLaMA-2	**F1-score**	0.53	0.70	0.35	0.87
		Recall	0.43	0.89	0.25	0.88
		Precision	0.71	0.58	0.6	0.86
	Mistral	**F1-score**	0.65	0.56	0.97	0.62
		Recall	0.77	0.75	0.98	0.38
		Precision	0.63	0.54	0.96	0.72

Table 8: **Performance metrics for single-dataset training experiments.** The table presents the F1-score, Recall, and Precision for the LLaMA-2 and Mistral models across the four datasets. Results are averaged over four seeds.

Two Datasets	Model	Metric	Test Dataset			
			Amazon	Dad Jokes	Headlines	One Liners
Amazon + Dad Jokes	LLaMA-2	**F1-score**	0.82	0.89	0.72	0.63
		Recall	0.81	0.88	0.84	0.52
		Precision	0.84	0.91	0.63	0.8
	Mistral	**F1-score**	0.89	0.95	0.73	0.68
		Recall	0.88	0.95	0.74	0.57
		Precision	0.91	0.95	0.74	0.87
Amazon + Headlines	LLaMA-2	**F1-score**	0.82	0.60	0.9	0.62
		Recall	0.8	0.58	0.89	0.52
		Precision	0.84	0.63	0.9	0.77
	Mistral	**F1-score**	0.89	0.73	0.96	0.59
		Recall	0.86	0.89	0.97	0.49
		Precision	0.92	0.62	0.96	0.77
Dad Jokes + Headlines	LLaMA-2	**F1-score**	0.66	0.89	0.92	0.68
		Recall	0.71	0.89	0.92	0.62
		Precision	0.61	0.9	0.91	0.76
	Mistral	**F1-score**	0.68	0.9	0.95	0.7
		Recall	0.62	0.9	0.96	0.72
		Precision	0.75	0.91	0.93	0.69
Dad Jokes + One Liners	LLaMA-2	**F1-score**	0.73	0.9	0.61	0.91
		Recall	0.73	0.92	0.55	0.93
		Precision	0.75	0.88	0.69	0.9
	Mistral	**F1-score**	0.74	0.95	0.7	0.94
		Recall	0.96	0.96	0.7	0.95
		Precision	0.61	0.95	0.7	0.94
Headlines + One Liners	LLaMA-2	**F1-score**	0.58	0.67	0.93	0.92
		Recall	0.52	0.91	0.92	0.92
		Precision	0.68	0.53	0.93	0.92
	Mistral	**F1-score**	0.73	0.68	0.97	0.94
		Recall	0.87	0.99	0.96	0.95
		Precision	0.63	0.51	0.97	0.94
One Liners +Amazon	LLaMA-2	**F1-score**	0.86	0.69	0.65	0.91
		Recall	0.83	0.98	0.64	0.91
		Precision	0.88	0.53	0.66	0.92
	Mistral	**F1-score**	0.9	0.68	0.73	0.94
		Recall	0.89	1	0.71	0.95
		Precision	0.92	0.52	0.76	0.94

Table 9: **Performance metrics for pair-dataset training experiments.** The table presents the F1-score, Recall, and Precision for the LLaMA-2 and Mistral models across the four datasets. Results are averaged over four seeds.

Train Dataset	Model	Metric	Test Dataset			
			Amazon	Dad Jokes	Headlines	One Liners
Held Out Amazon	LLaMA-2	**F1-score**	0.71	0.88	0.89	0.88
		Recall	0.74	0.88	0.89	0.87
		Precision	0.67	0.88	0.9	0.89
	Mistral	**F1-score**	0.73	0.94	0.96	0.94
		Recall	0.94	0.95	0.96	0.93
		Precision	0.6	0.94	0.96	0.94
Held Out Dad Jokes	LLaMA-2	**F1-score**	0.84	0.69	0.9	0.87
		Recall	0.82	0.96	0.9	0.86
		Precision	0.85	0.54	0.9	0.88
	Mistral	**F1-score**	0.9	0.69	0.96	0.94
		Recall	0.88	0.99	0.97	0.94
		Precision	0.92	0.53	0.96	0.93
Held Out Headlines	LLaMA-2	**F1-score**	0.86	0.89	0.64	0.88
		Recall	0.84	0.89	0.56	0.86
		Precision	0.88	0.89	0.74	0.9
	Mistral	**F1-score**	0.89	0.95	0.72	0.93
		Recall	0.88	0.95	0.68	0.94
		Precision	0.91	0.95	0.76	0.93
Held Out One Liners	LLaMA-2	**F1-score**	0.83	0.89	0.91	0.66
		Recall	0.8	0.87	0.91	0.59
		Precision	0.87	0.91	0.91	0.74
	Mistral	**F1-score**	0.9	0.96	0.96	0.7
		Recall	0.88	0.95	0.97	0.61
		Precision	0.92	0.96	0.95	0.82

Table 10: **Performance metrics for triple-dataset training experiments.** The table presents the F1-score, Recall, and Precision for the LLaMA-2 and Mistral models across the four datasets. Results are averaged over four seeds.

Train Dataset	Model	Test Dataset (Std)			
		Amazon	Dad Jokes	Headlines	One Liners
Amazon	LLaMA-2	0.17	1.67	1.2	1.41
	Mistral	0.68	2.84	1.37	0.8
Dad Jokes	LLaMA-2	4.02	0.46	2.07	1.47
	Mistral	1.37	0.44	1.24	0.67
Headlines	LLaMA-2	4.7	0.64	0.45	1.04
	Mistral	2.42	1.14	1.24	0.67
One Liners	LLaMA-2	0.36	0.88	0.83	0.45
	Mistral	3.69	0.34	0.95	0.26
Amazon + Dad Jokes	LLaMA-2	0.5	0.76	1.1	1.17
	Mistral	0.33	0.45	0.71	2.37
Amazon + Headlines	LLaMA-2	0.34	0.68	0.25	1.16
	Mistral	0.13	3.08	0.49	1.68
Dad Jokes + Headlines	LLaMA-2	2.3	0.8	0.53	1.47
	Mistral	1.91	0.26	0.37	1.35
Dad Jokes + One Liners	LLaMA-2	1.1	0.52	0.92	0.37
	Mistral	5.7	0.36	0.71	0.16
Headlines + One Liners	LLaMA-2	1.5	1.68	0.49	0.24
	Mistral	3.58	0.51	0.29	0.13
One Liners + Amazon	LLaMA-2	0.36	2.06	1.84	0.23
	Mistral	0.41	0.28	0.8	0.37
Held Out Amazon	LLaMA-2	1.23	0.5	0.75	0.43
	Mistral	5.45	0.41	0.54	0.65
Held Out Dad Jokes	LLaMA-2	0.92	2.11	0.54	1.9
	Mistral	0.28	1.38	0.11	0.29
Held Out Headlines	LLaMA-2	0.41	0.7	1.38	0.29
	Mistral	0.29	0.17	0.56	0.21
Held Out One Liners	LLaMA-2	1.2	0.45	0.74	1.46
	Mistral	0.38	0.4	0.15	3.05

Table 11: **Standard deviations of model accuracy across all experimental setups.** The table presents the standard deviations of accuracy scores for both LLaMA-2 and Mistral. Rows represent the various Train Dataset configurations, including single-, pair-, and triple-dataset scenarios, while columns represent the Test Dataset

Train Dataset	Model	Test Dataset (95% CI)			
		Amazon	Dad Jokes	Headlines	One Liners
Amazon	LLaMA-2	±0.27	±2.66	±1.91	±2.24
	Mistral	±1.08	±4.52	±2.18	±1.27
Dad Jokes	LLaMA-2	±6.4	±0.73	±3.29	±2.34
	Mistral	±2.18	±0.7	±1.97	±1.07
Headlines	LLaMA-2	±0.75	±1.02	±0.72	±1.65
	Mistral	±3.85	±1.81	±0.29	±2.55
One Liners	LLaMA-2	±0.57	±1.4	±1.32	±0.72
	Mistral	±5.87	±0.54	±1.51	±0.41
Amazon + Dad Jokes	LLaMA-2	±0.8	±1.21	±1.75	±1.86
	Mistral	±0.53	±0.72	±1.13	±3.77
Amazon + Headlines	LLaMA-2	±0.54	±1.08	±0.4	±1.85
	Mistral	±0.21	±4.9	±0.78	±2.67
Dad Jokes + Headlines	LLaMA-2	±3.66	±1.27	±0.84	±2.34
	Mistral	±3.04	±0.41	±0.59	±2.15
Dad Jokes + One Liners	LLaMA-2	±1.75	±0.83	±1.46	±0.59
	Mistral	±9.07	±0.57	±1.13	±0.25
Headlines + One Liners	LLaMA-2	±2.39	±2.67	±0.78	±0.38
	Mistral	±5.7	±0.81	±0.46	±0.21
One Liners + Amazon	LLaMA-2	±0.57	±3.28	±2.93	±0.37
	Mistral	±0.65	±0.45	±1.27	±0.59
Held Out Amazon	LLaMA-2	±1.96	±0.8	±1.19	±0.68
	Mistral	±8.67	±0.65	±0.86	±1.03
Held Out Dad Jokes	LLaMA-2	±1.46	±3.36	±0.86	±3.02
	Mistral	±0.45	±2.2	±0.18	±0.46
Held Out Headlines	LLaMA-2	±0.65	±1.11	±2.2	±0.46
	Mistral	±0.46	±0.27	±0.89	±0.33
Held Out One Liners	LLaMA-2	±1.91	±0.72	±1.18	±2.32
	Mistral	±0.6	±0.64	±0.24	±4.85

Table 12: **Confidence Intervals (95%) for model accuracy across all experimental setups.** The table presents the 95% confidence intervals (CI) for accuracy scores calculated over four independent training seeds for both LLaMA-2 and Mistral. Rows represent the various Train Dataset configurations, including single-, pair-, and triple-dataset scenarios, while columns represent the Test Dataset.

Target Dataset	Model	Transfer Accuracy Change		
		Single → Double	Double → Triple	Single → Triple
Amazon	LLaMA-2	↑ 5.58	↑ 2.34	↑ 7.92
	Mistral	↑ 2.27	↓ −2.6	↓ −0.33
Dad Jokes	LLaMA-2	↓ −3.77	↓ −0.31	↓ −4.08
	Mistral	↑ 1.31	↓ −2.34	↓ −1.03
Headlines	LLaMA-2	↑ 6.23	↑ 2.33	↑ 8.56
	Mistral	↑ 2.55	↑ 0.83	↑ 3.38
One Liners	LLaMA-2	↑ 4.02	↓ −0.21	↑ 3.81
	Mistral	↑ 1.96	↑ 3.54	↑ 5.50
Average	LLaMA-2	↑ 3.02	↑ 1.04	↑ 4.05
	Mistral	↑ 2.02	↓ −0.14	↑ 1.88

Table 13: **Transfer Accuracy Differences Across Experiments.** Reported values reflect the change in mean transfer accuracy (in percentage points) between training setups for each model and target dataset. Accuracy is averaged over all relevant source datasets for each target. ↑ indicates improvement, and ↓ indicates degradation. For example, Mistral on Amazon improves from Single to Double training by 2.27 points, but decreases from Double to Triple training by 2.60 points. The "Average" row reports the mean change across all target datasets. Overall, LLaMA-2 tends to benefit more from increased training diversity than Mistral.

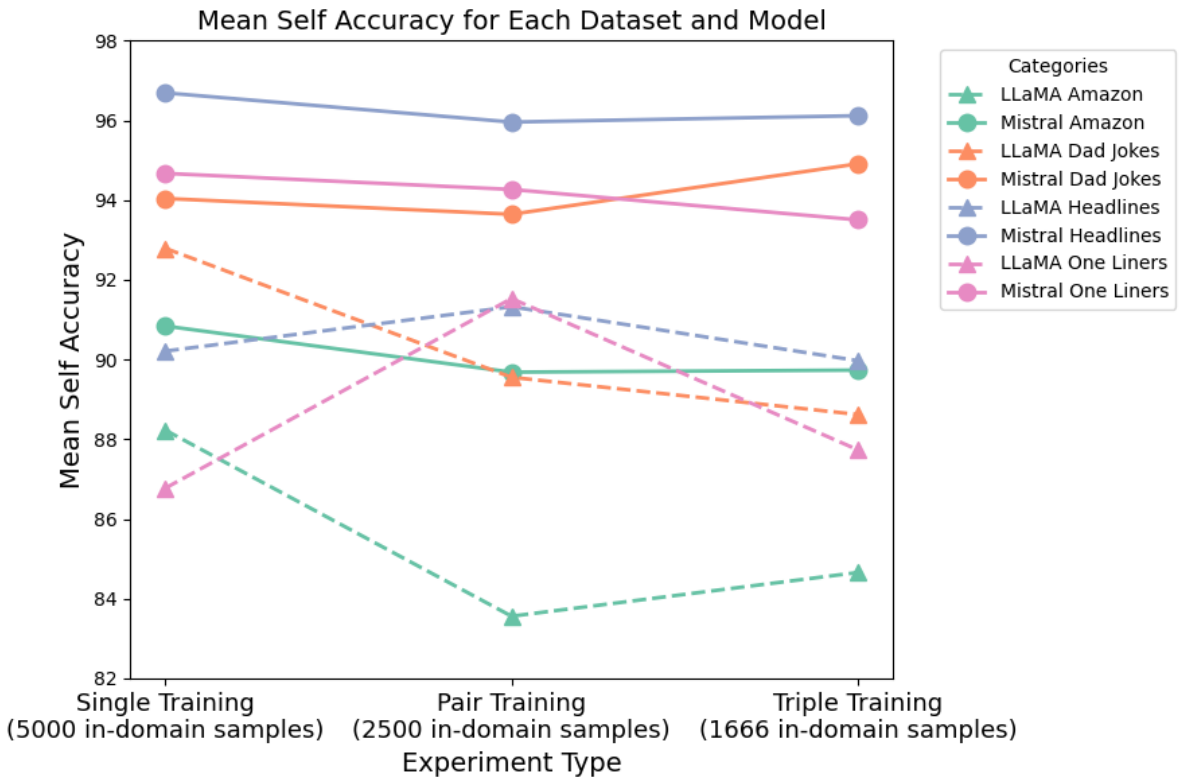

Figure 3: **Comparing self accuracy across the experiments.** Mistral results are shown with solid lines, LLaMA-2 with dotted lines. Colors represent different test datasets. The x-axis indicates the experiment type, along with the number of in-domain training samples, and the y-axis shows the mean accuracy on the in-domain dataset. Mistral exhibits robust performance even as in-domain data decreases. LLaMA-2 results are less stable but show only minor decreases in accuracy.

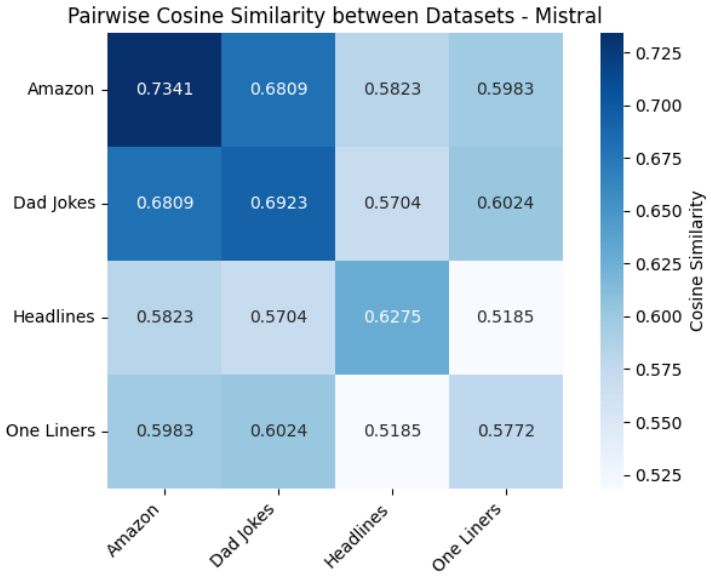

Figure 4: Mistral Embeddings Cosine Similarity Heatmap.

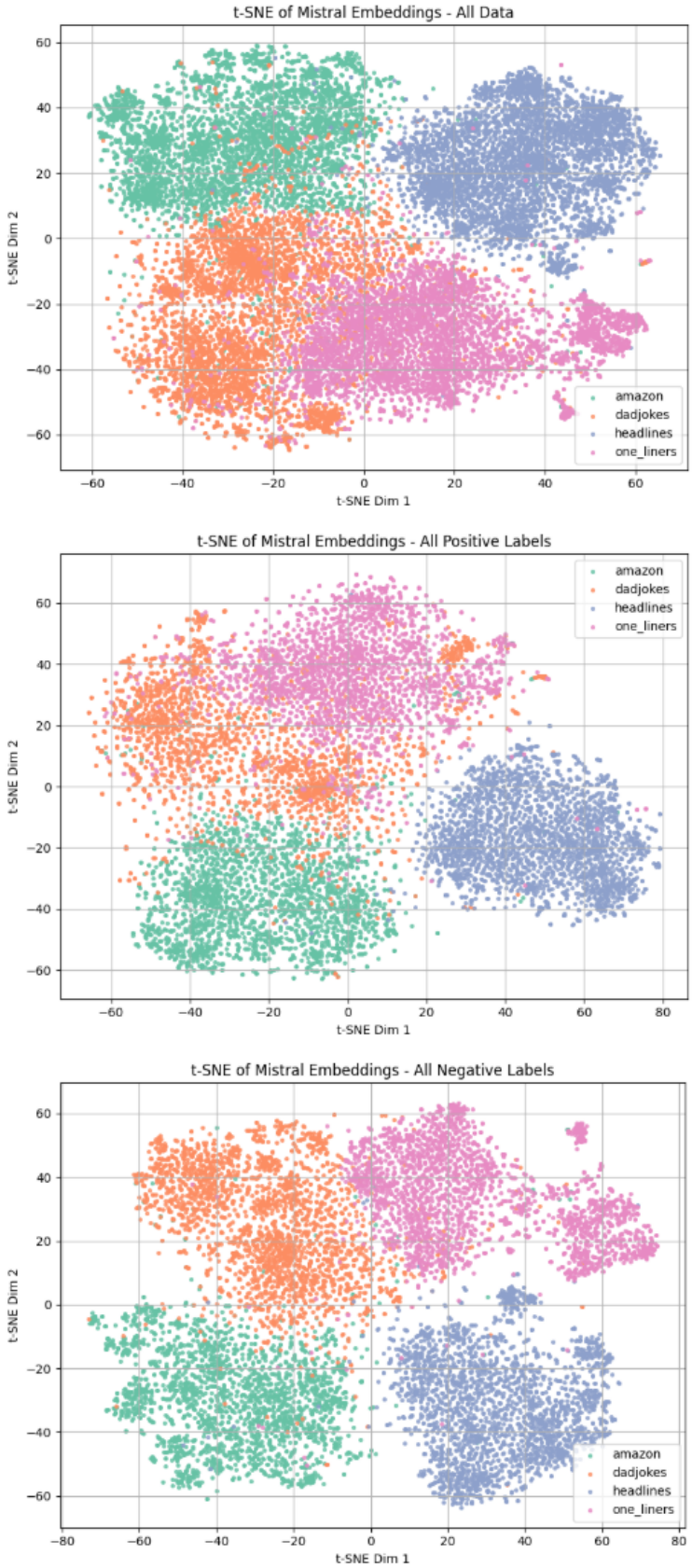

Figure 5: **t-SNE visualization of Mistral (pretrained) embeddings across the datasets.** The plots display embeddings for all data (top), positive samples only (middle), and negative samples only (bottom). Each color represents a specific dataset. The visualization demonstrates the relative distinction between the datasets, supporting our domain classification results. Notably, the overlap between Dad Jokes and One Liners aligns with our discussion regarding their shared structural properties and humor types.

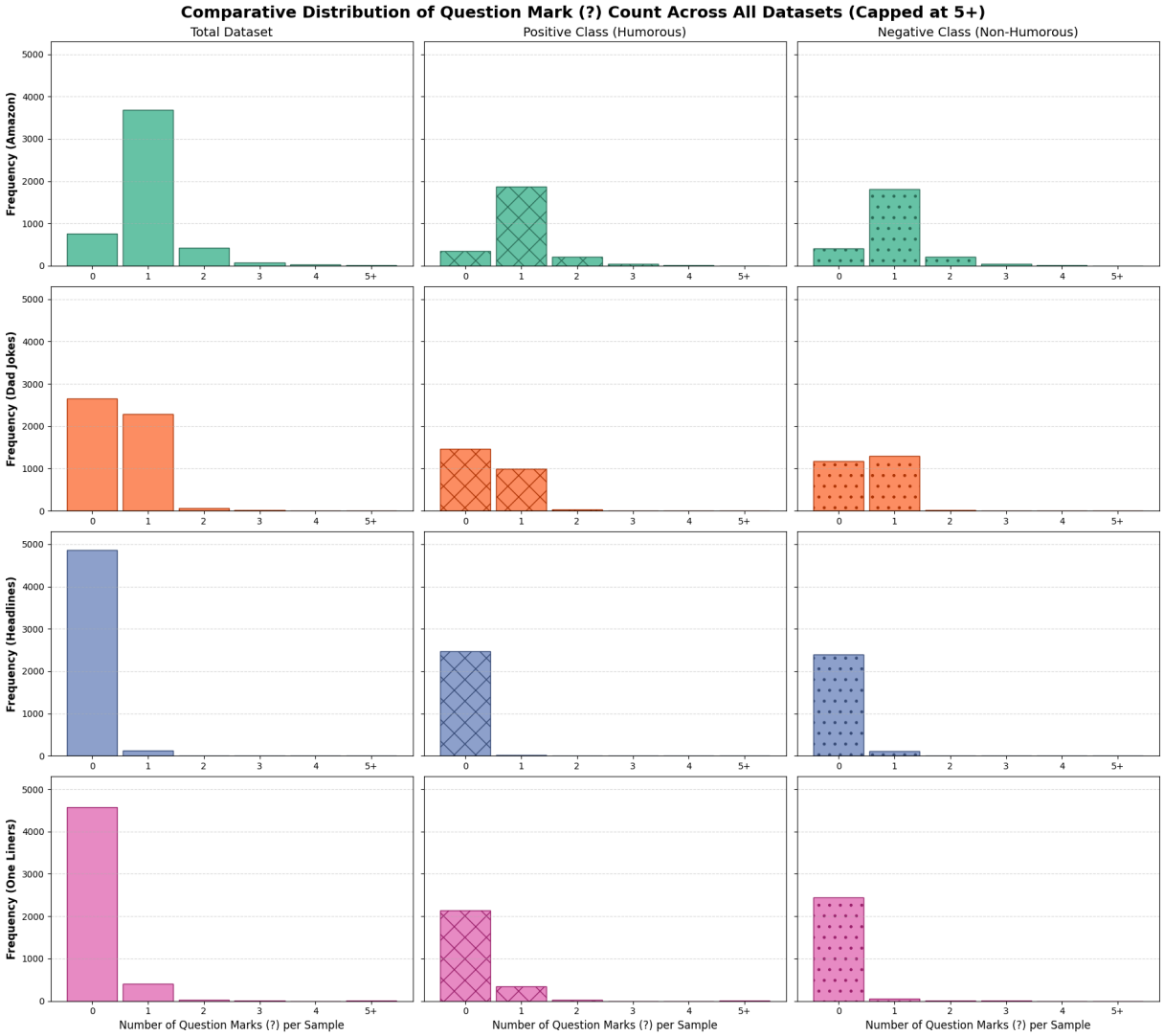

Figure 6: **Comparative distribution of question mark (?) counts across all datasets.** The charts illustrate the frequency of question marks per sample for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). Counts are capped at 5+ occurrences.

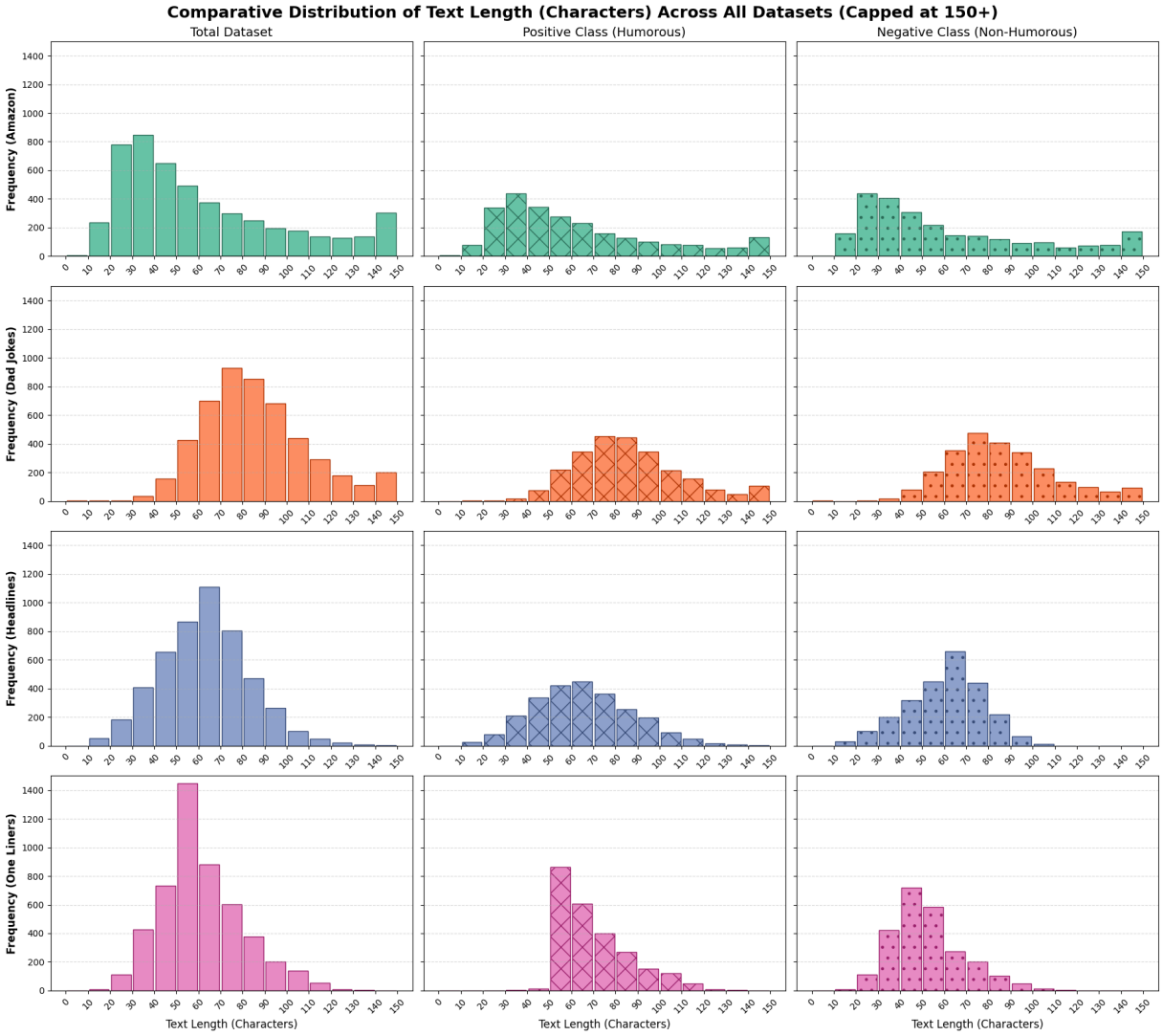

Figure 7: **Comparative distribution of text length (characters) across all datasets.** The charts illustrate the frequency of character counts per sample for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). Counts are binned and capped at 150+ characters.

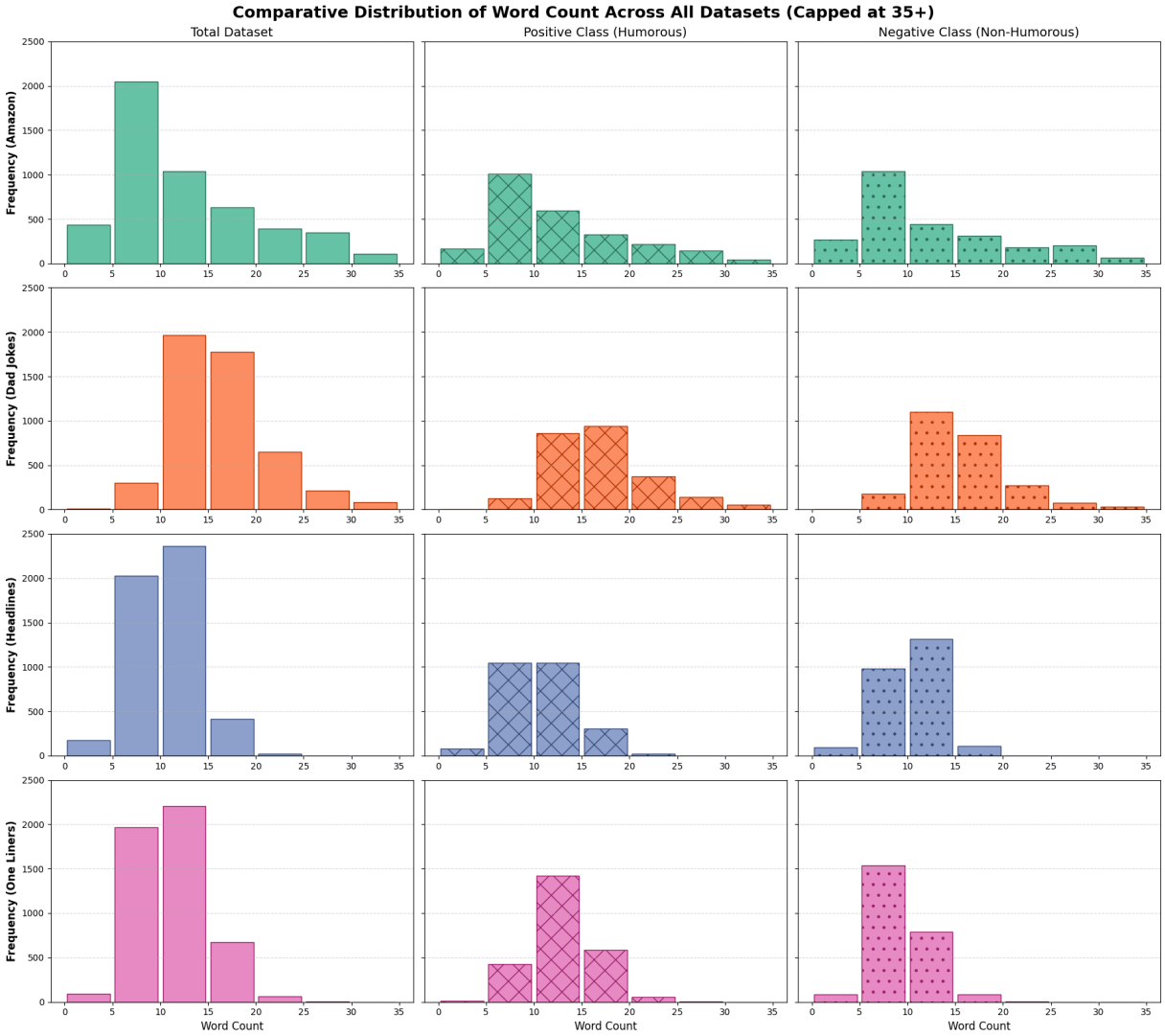

Figure 8: **Comparative distribution of word counts across all datasets.** The charts illustrate the frequency of word counts per sample for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). Counts are binned and capped at 35+ words.

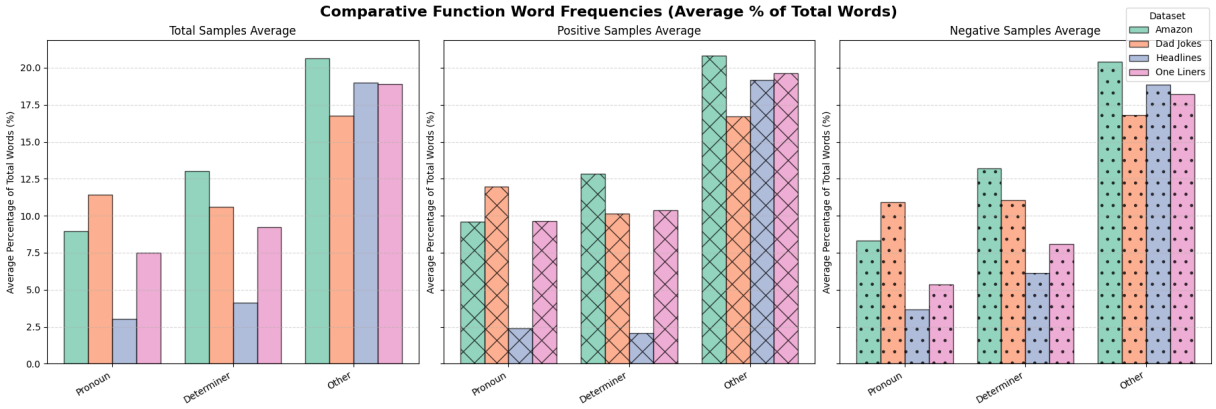

Figure 9: **Comparative function word frequencies across all datasets.** The charts illustrate the average percentage of total words for specific categories (Pronoun, Determiner, and Other) for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). The y-axis represents the average percentage of total words per sample. Colors and patterns distinguish the four datasets: Amazon, Dad Jokes, Headlines, and One Liners.

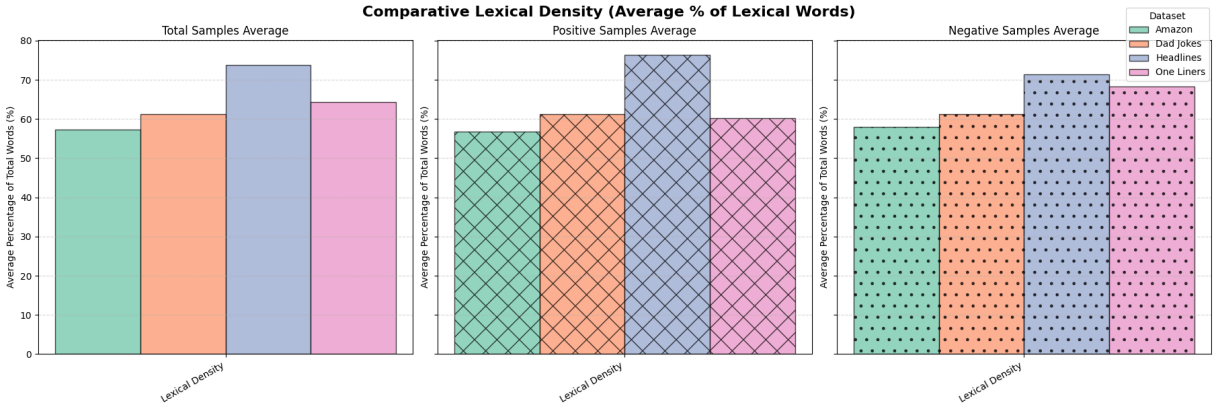

Figure 10: **Comparative lexical density across all datasets.** The charts illustrate the average percentage of lexical words for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). The y-axis represents the average percentage of total words per sample. Colors and patterns distinguish the four datasets: Amazon, Dad Jokes, Headlines, and One Liners.

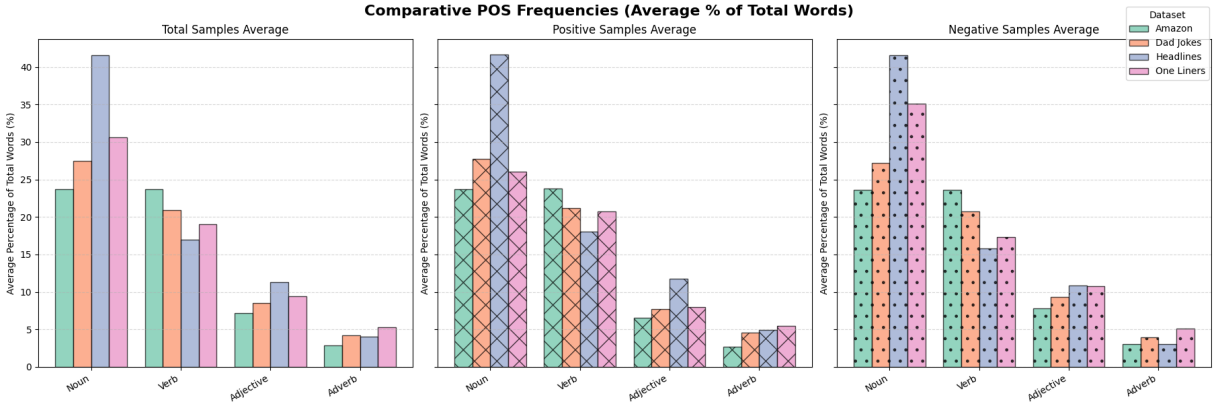

Figure 11: **Comparative Part-Of-Speech (POS) frequencies across all datasets.** The charts illustrate the average percentage of total words for major POS categories (Noun, Verb, Adjective, and Adverb) for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). The y-axis represents the average percentage of total words per sample. Colors and patterns distinguish the four datasets: Amazon, Dad Jokes, Headlines, and One Liners.

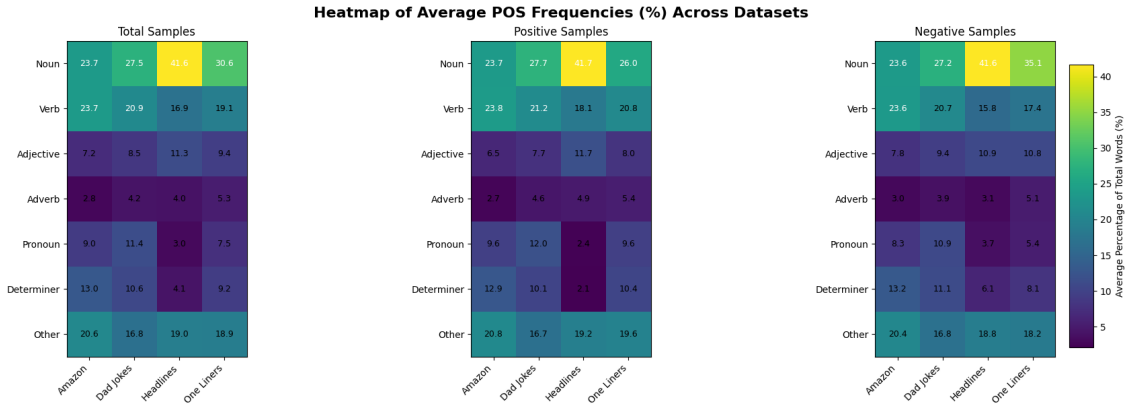

Figure 12: **Heatmap of average Part-Of-Speech (POS) frequencies across datasets.** The heatmaps illustrate the average percentage of total words for various POS categories across the four evaluation datasets for the total samples (left), positive humorous class (center), and negative non-humorous class (right). The color intensity and numerical values represent the average percentage of total words per sample. Rows correspond to POS categories, while columns represent the datasets: Amazon, Dad Jokes, Headlines, and One Liners.