CHum 2026

The 2nd Workshop on Computational Humor

Proceedings of the Workshop

July 3, 2026

The CHum organizers gratefully acknowledge the support from the following sponsors.

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-431-6

Preface

The first Computational Humor (CHum) Workshop that we organized took place at COLING 2025 in Abu Dhabi, UAE (chum2025.github.io). It featured eleven selected papers and three invited keynotes that we thought represented the landscape of computational humor in the age of the "AI" turmoil that is still very much going on, but also pointed to future developments beyond prompting humor or humor analysis out of LLMs.

Then, as now, we take the position that in order to advance humor-contained interaction/generation/detection in computational systems, there is a need to understand where the systems fail and where they succeed. Relying solely on examples of impressively generated humorous texts is insufficient, as such outputs may stem from models' capabilities in memorization and pattern matching rather than the acquisition of structured knowledge that demonstrates humor competence.

Because the aim of the workshop is to foster work on modeling the processes of humor with current methods in computational linguistics and natural language processing, papers to be accepted for presentation at the workshop had to be either based on current computational methods and informed by the existing humor research or advance the state of the art in humor research by applying current computational models. There are arguments that humor relies on the background knowledge that may not be available from a single modality, such as text, and as such is not machine-learnable from that modality in a straightforward manner. With this in mind, a principal goal of this workshop is to unite researchers who can together probe the limits of various meaning representations—symbolic, neural, hybrid, or unified—for humor processing.

Reliable and versatile humor-aware algorithms could have many uses. Commonly-cited examples, such as joke generation and the translation of humorous texts, tend to be application-oriented engineering tasks. But from another angle, computational humor, particularly if done in a symbolic (knowledge-based) fashion that captures the necessary and sufficient resources and algorithms, could help researchers in the social sciences and humanities better understand how human humor works. Along all these lines, knowledge-based work has progressed since the late nineties, from purely template-based to enhanced bag-of-word approaches to proposals for ontology-based systems. While this work has not yet left the arena of academic toy systems, there has been a slow but steady progression in the linguistic and computational complexity of the methods, as well as a broadening of application areas to include augmentative communication and computer-assisted translation.

As hinted at above, human humor is the result of a complex cognitive process that requires the experiencer to identify and act on culturally significant cues in the source material. Moreover, it depends on surprise and "violations of predictions", making it difficult to capture in trained models that are at their core predictable and only pseudo-random. Nonetheless, recent work in conventional text domains on addressing the knowledge acquisition bottleneck, on highly fluent text generation, and on attention mechanisms and interpretable models, seem ripe for application to computational humor in supervised settings. And neurosymbolic hybrid approaches and unification approaches have recently entered the stage of computational humor research, where interesting new results should be expected in the near future.

This year, the second CHum (chumweb.org), again organized by the Semantic AI & Creativity Lab at East Texas A&M (etamu.edu/saicl) and the Computational Linguistics at Manitoba lab (clam.cs.umanitoba.ca), but now also sponsored by the Applied Knowledge Representation and Natural Language Understanding (AKRaNLU) Lab at Purdue University (polytechnic.purdue.edu/facilities/akranlu), will feature a similar number of papers that fit these aims. We also invited a keynote by a psychologist, and will round out the workshop with a discussion panel on computational humor and the law.

Finally, we would like to thank everyone who submitted a paper to the workshop, as well as the members of our Program Committee for their timely and insightful reviews.

Tristan Miller, Ori Amir, Julia Rayz, Tiansi Dong, and Christian F. Hempelmann July 2026

Organizing Committee

Tristan Miller, University of Manitoba
Ori Amir, Fulbright University Vietnam
Julia Rayz, Purdue University
Tiansi Dong, Alan Turing Institute & University of Cambridge
Christian F. Hempelmann, East Texas A&M University

Program Committee

Adam Jatowt, University of Innsbruck
Anne-Gwenn Bosser, École Nationale d'Ingénieurs de Brest
Aram Sinnreich, American University
Elena Mikhalkova, Tyumen State University
Hongfei Lin, Dalian University of Technology
Joe Toplyn, Twenty Lane Media
Kory Mathewson, Google DeepMind
Larry Lefkowitz, Working Knowledge LLC
Liana Ermakova, University of Western Brittany
Liang Yang, Dalian University of Technology
Mirella Manfredi, Zurich University
Monika, JP Morgan
Nadezhda Ganzherli, Tyumen State University
Nathaniel Laywine, York University
Piotr Mirowski, Google DeepMind
Rada Mihalcea, University of Michigan
Sergio Spaccavento, Osservatorio Feldman
Sophie Jentzsch, German Aerospace Center (DLR)
Tatiana Ringenberg, Purdue University
Thomas Winters, KU Leuven
Tony Veale, University College Dublin
Vladislav Maraev, University of Gothenburg
Yi Zhang, Purdue University

Cognition Without Evolution: An Evolutionary Perspective on AI Humor

Gil Greengross
Aberystwyth University

Abstract: Recent advances in large language models have transformed research on computational humor and raised fundamental questions about the nature of humor itself. Human humor is widely viewed as an evolutionary adaptation that promotes social bonding, cooperation, status, and mate attraction. AI systems, however, lack the evolutionary history, emotions, and social motives that underlie these functions, yet they can generate, recognize, and evaluate humor. This talk explores AI humor through the lens of evolutionary psychology. I argue that AI provides a unique natural experiment for distinguishing between aspects of humor that depend on evolved human adaptations and those that reflect more general information-processing mechanisms. Drawing on research in computational humor, evolutionary psychology, and humor studies, I examine similarities and differences between biological and artificial systems, discuss evolved psychological adaptations that shape the way we assess and experience AI humor, and suggest ways in which computational approaches to humor can benefit from invoking an evolutionary perspective. The talk is exploratory and while it may not provide many answers, it is intended to identify new questions for interdisciplinary research.

Panel
Computational Humor and the Law

As humor-aware AI systems move from research prototypes into public-facing applications, they raise complex questions at the intersection of language, law, creativity, and governance. This multidisciplinary panel brings together digital rights advocate Brad Templeton, television writer and Witscript creator Joe Toplyn, and legal scholar Laura E. Little to examine the social, ethical, and legal implications of computational humor. Topics may include free speech and content moderation, intellectual property and comedic style, liability for AI-generated jokes, and the challenges of humor across cultures and legal jurisdictions. Drawing on perspectives from online community governance, professional comedy writing, and legal doctrine, the panel will explore how humor-aware AI can be developed responsibly while preserving creativity, expression, and cultural nuance. The session will conclude with reflections on guiding principles for future computational humor research.

Panelists:

Brad Templeton, Electronic Frontier Foundation and rec.humor.funny
Joe Toplyn, Twenty Lane Media, LLC
Laura E. Little, Temple Law School

Table of Contents

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

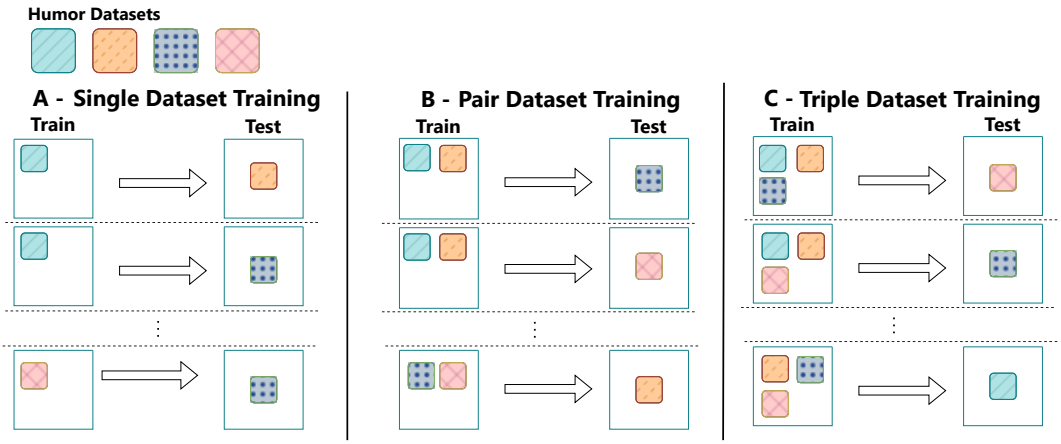

Figure 1: Overview of the experimental setups. **(A) Single Dataset Training:** Train on one dataset and test on each of the other datasets. **(B) Double Dataset Training:** Train on two of datasets and test on each of the remaining three. **(C) Triple Dataset Training:** Train on three datasets and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

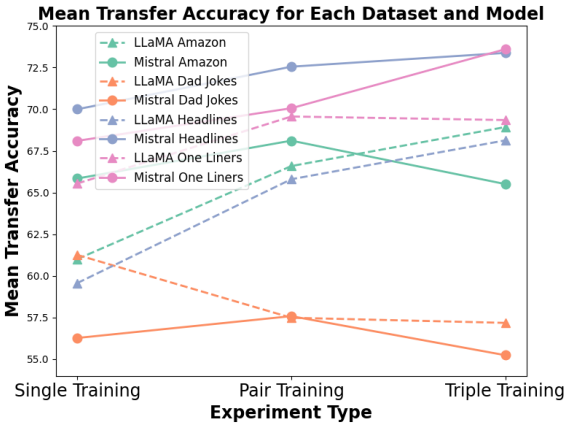

Figure 2: **[Increasing the training data diversity improves transfer.] Comparing transfer across the experiments.** Mistral results are shown with solid lines, LLaMA-2 with dotted lines. Colors represent different test datasets. In general, more diverse training data leads to better transfer than single-dataset 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.

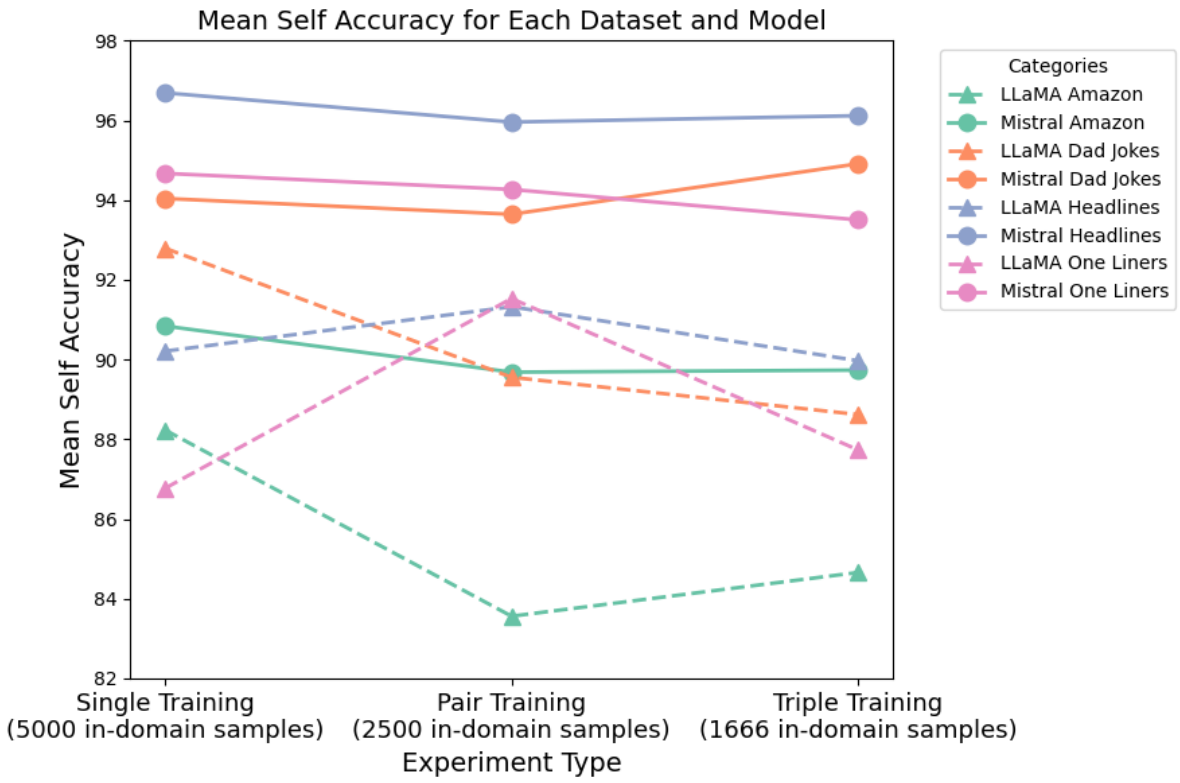

Figure 3: **Comparing self accuracy across the experiments.** Mistral results are shown with solid lines, LLaMA-2 with dotted lines. Colors represent different test datasets. The x-axis indicates the experiment type, along with the number of in-domain training samples, and the y-axis shows the mean accuracy 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.

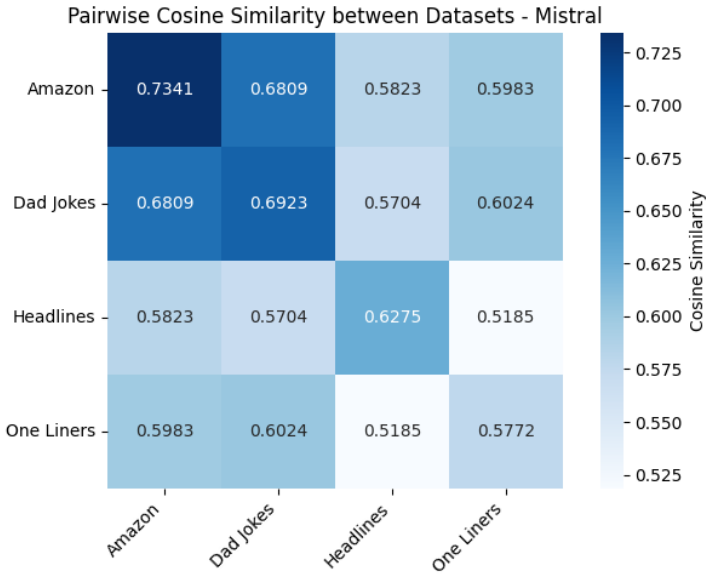

Figure 4: Mistral Embeddings Cosine Similarity Heatmap.

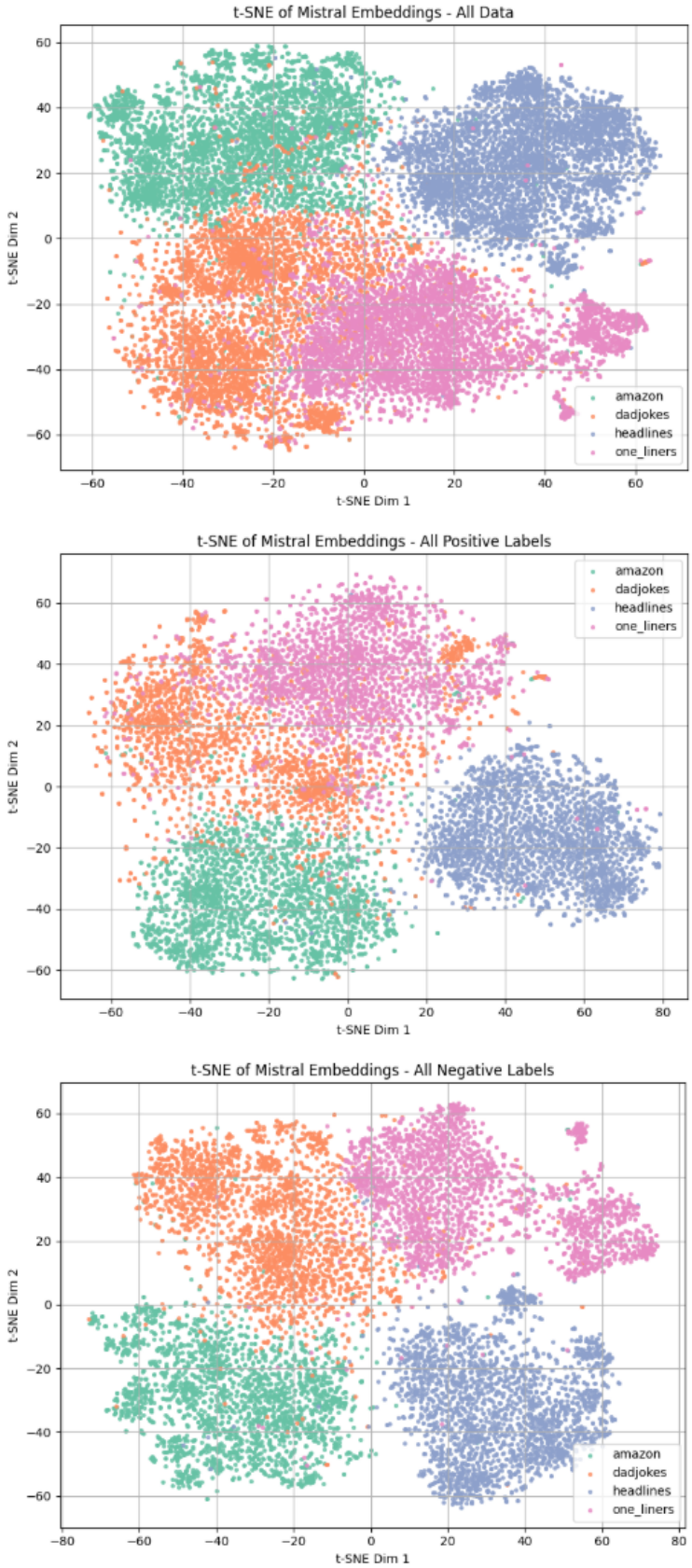

Figure 5: **t-SNE visualization of Mistral (pretrained) embeddings across the datasets.** The plots display embeddings for all data (top), positive samples only (middle), and negative samples only (bottom). Each color represents a specific dataset. The visualization demonstrates the relative distinction 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.

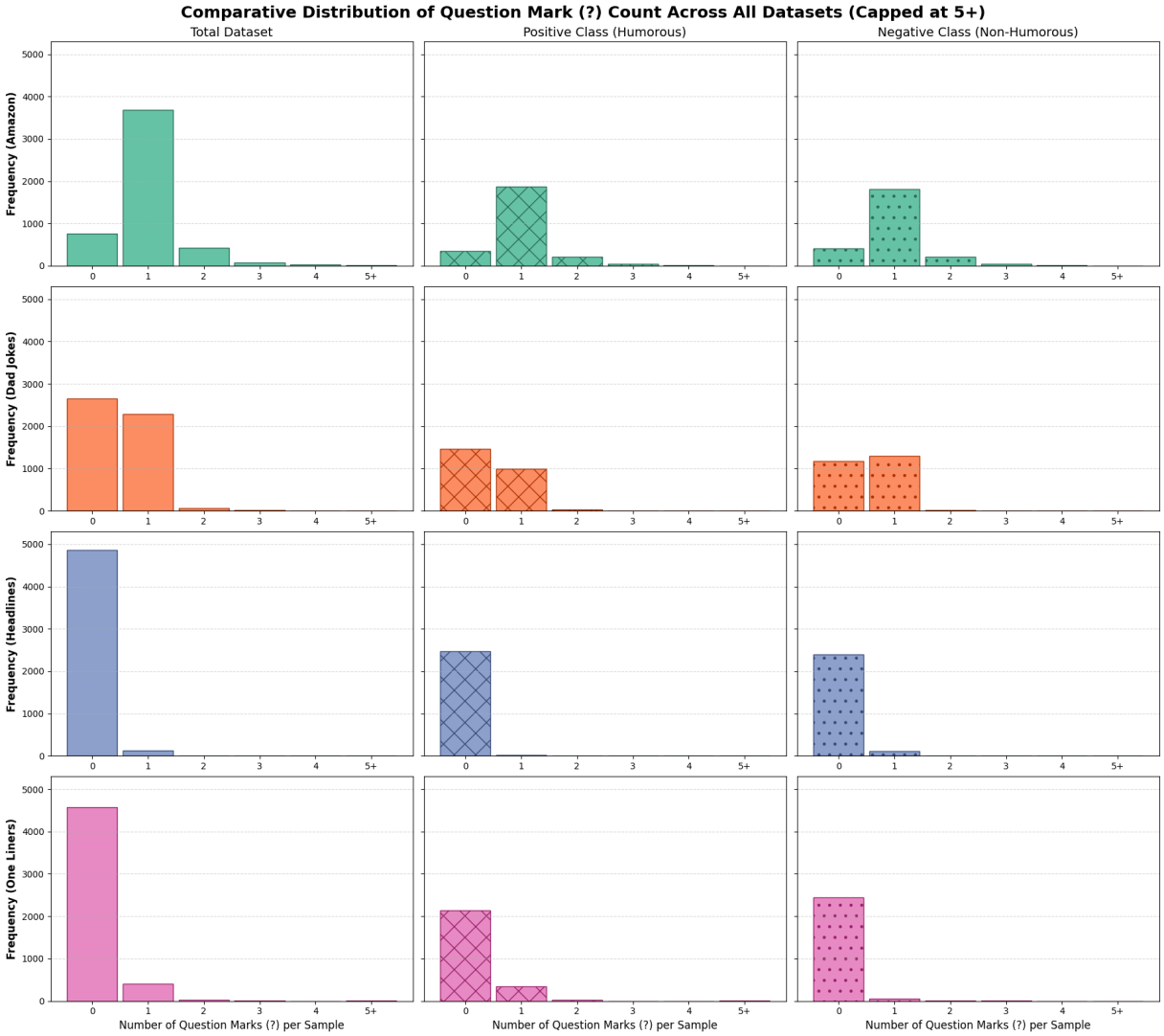

Figure 6: **Comparative distribution of question mark (?) counts across all datasets.** The charts illustrate the frequency of question marks per sample for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). Counts are capped at 5+ occurrences.

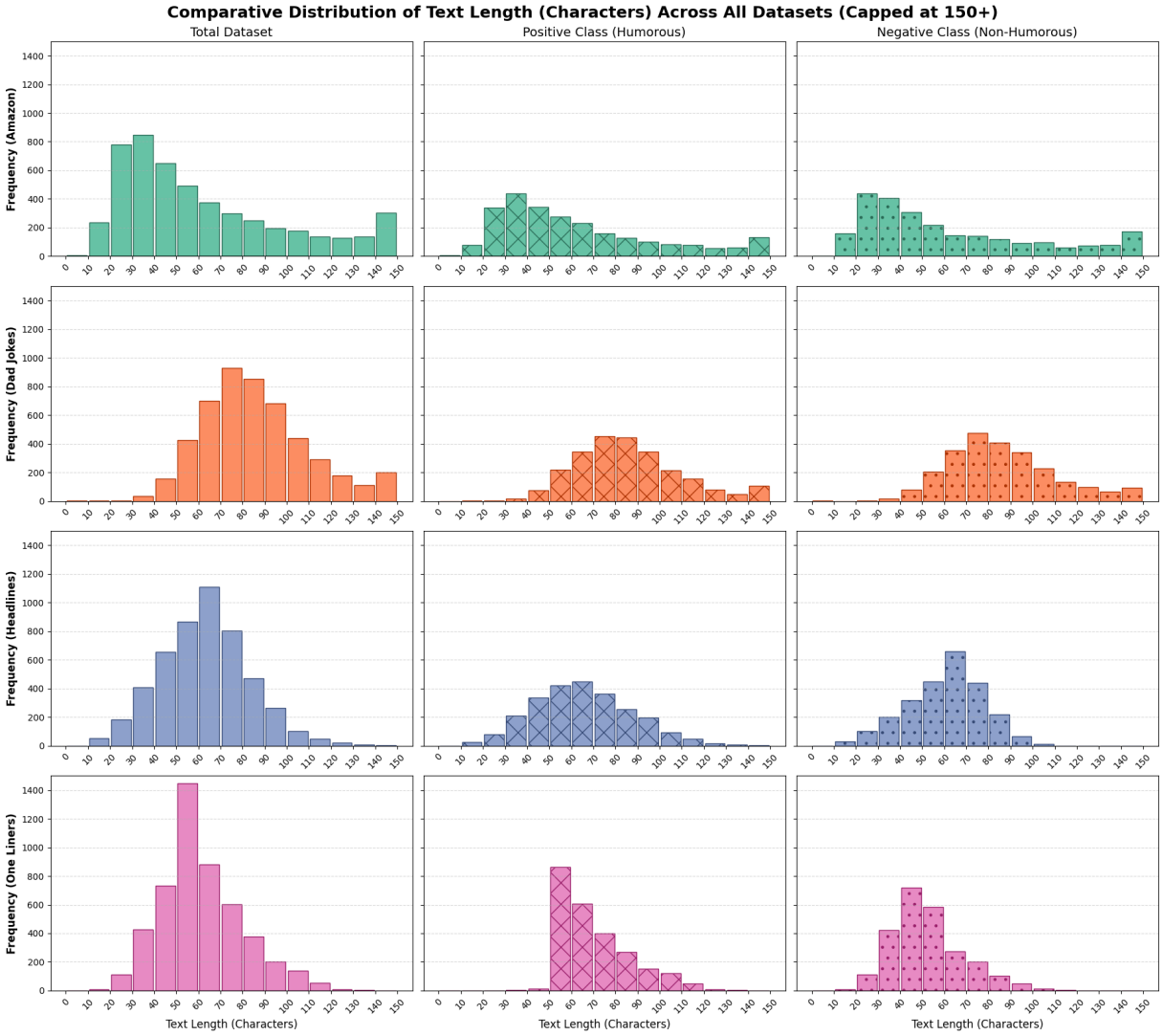

Figure 7: **Comparative distribution of text length (characters) across all datasets.** The charts illustrate the frequency of character counts per sample for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). Counts are binned and capped at 150+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

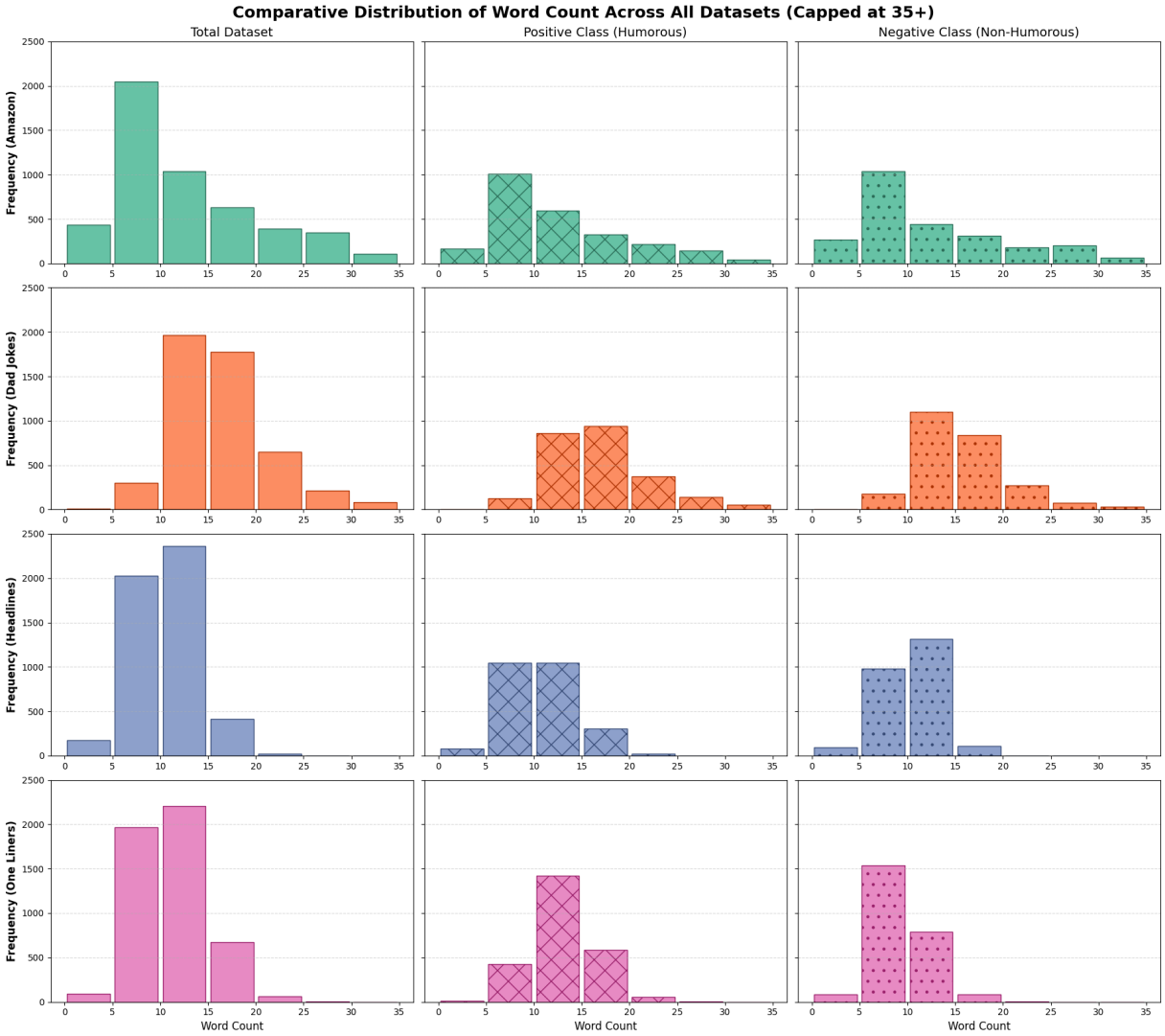

Figure 8: **Comparative distribution of word counts across all datasets.** The charts illustrate the frequency of word counts per sample for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). Counts are binned and capped at 35+ words.

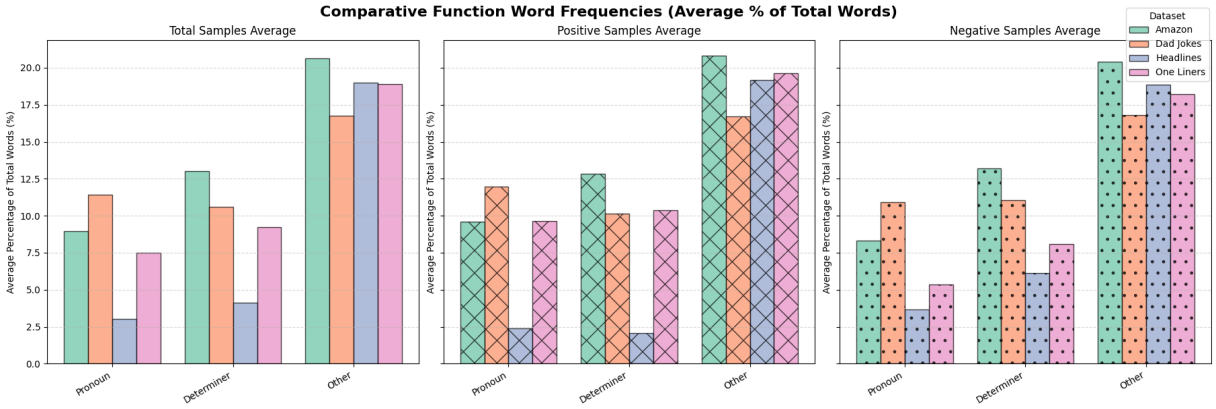

Figure 9: **Comparative function word frequencies across all datasets.** The charts illustrate the average percentage of total words for specific categories (Pronoun, Determiner, and Other) for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). 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.

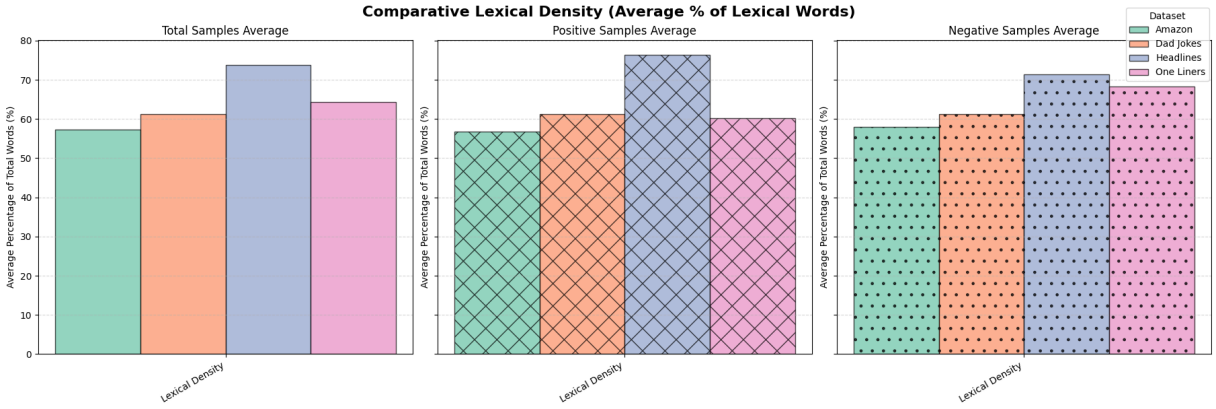

Figure 10: **Comparative lexical density across all datasets.** The charts illustrate the average percentage of lexical words for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (right). The y-axis represents the average percentage of total words per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

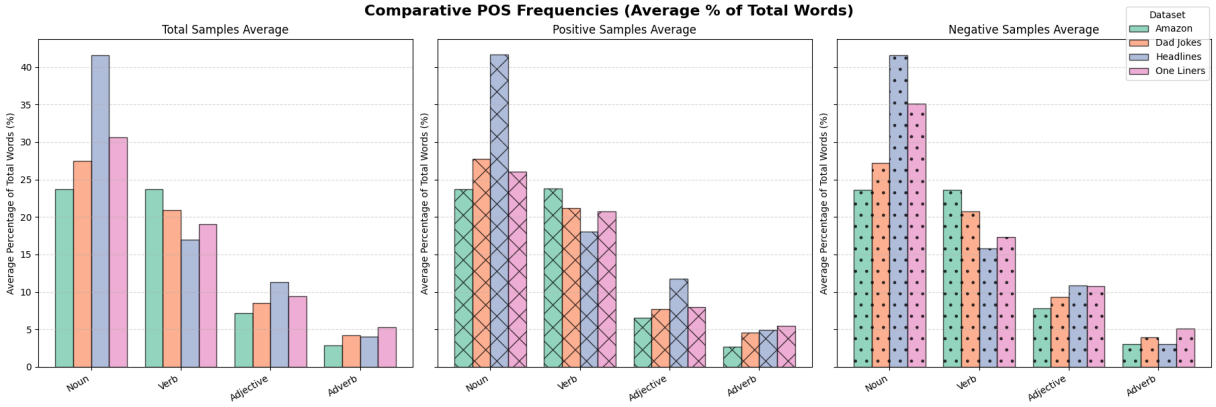

Figure 11: **Comparative Part-Of-Speech (POS) frequencies across all datasets.** The charts illustrate the average percentage of total words for major POS categories (Noun, Verb, Adjective, and Adverb) for the total dataset (left), the positive humorous class (center), and the negative non-humorous class (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.

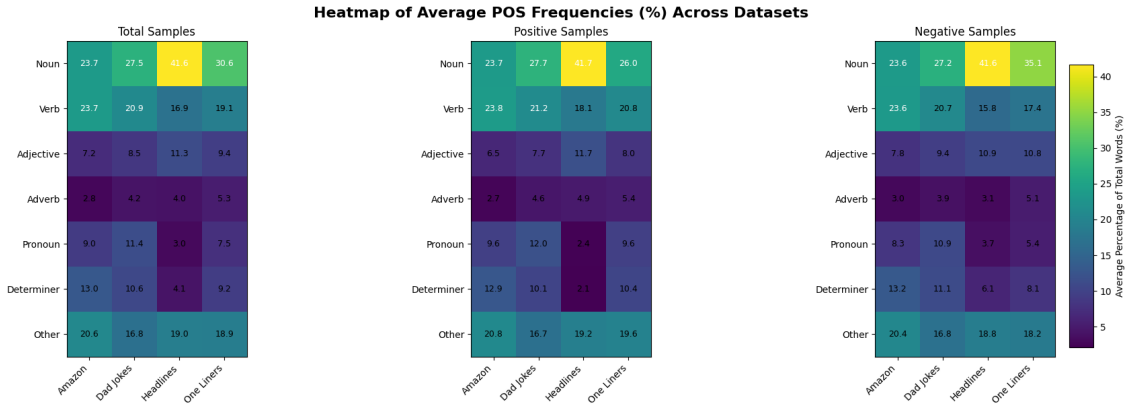

Figure 12: **Heatmap of average Part-Of-Speech (POS) frequencies across datasets.** The heatmaps illustrate the average percentage of total words for various POS categories across the four evaluation datasets for the total samples (left), positive humorous class (center), and negative non-humorous class (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.

Timing In stand-up Comedy: Text, Audio, Laughter, Kinesics (TIC-TALK): Pipeline and Database for the Multimodal Study of Comedic Timing

Yaelle Zribi[1] **Florian Cafiero**[1,2] **Vincent Lépinay**[3] **Chahan Vidal-Gorène**[1]

[1]Centre Jean-Mabillon, École nationale des chartes, PSL, Paris, France
[2]Laboratoire de Recherche, EPITA, Paris, France
[3]médialab, Sciences Po, Paris, France
{yaelle.zribi, florian.cafiero, chahan.vidal-gorene}@chartes.psl.eu
vincent.lepinay@sciencespo.fr

Abstract

Stand-up comedy, and humor in general, are often studied through their verbal content. Yet live performance relies just as much on embodied presence and audience feedback. We introduce TIC-TALK, a multimodal resource with 5,400+ temporally aligned topic segments capturing language, gesture, and audience response across 90 professionally filmed stand-up comedy specials (2015–2024). The pipeline combines BERTopic for 60 s thematic segmentation with dense sentence embeddings, Whisper-AT for 0.8 s laughter detection, a fine-tuned YOLOv8-cls shot classifier, and YOLOv8s-pose for raw keypoint extraction at 1 fps. Raw 17-joint skeletal coordinates are retained without prior clustering, enabling the computation of continuous kinematic signals—arm spread, kinetic energy, and trunk lean—that serve as proxies for performance dynamics. All streams are aligned by hierarchical temporal containment without resampling, and each topic segment stores its sentence-BERT embedding for downstream similarity and clustering tasks. As a concrete use case, we study laughter dynamics across 24 thematic topics: kinetic energy negatively predicts audience laughter rate ($r = -0.75$, $N = 24$), consistent with a stillness-before-punchline pattern; personal and bodily content elicits more laughter than geopolitical themes; and shot close-up proportion correlates positively with laughter ($r = +0.28$), consistent with reactive montage.

1 Introduction

Humor is one of the most complex forms of human communication, involving timing, embodiment, and interaction. Stand-up comedy, in particular, constitutes an ideal case study for modeling how verbal, acoustic, and visual cues align to produce shared affective meaning.

The performer's gestures, pauses, and interaction with audience laughter are essential components of meaning and rhythm. Modeling these elements jointly raises both technical and conceptual challenges: how can we represent and evaluate *comedic timing* computationally? We take *comedic timing* to denote short-lag coordination across text, gesture, and audience response in live delivery.

Before turning to the method itself, it is worth situating the kind of question such modeling opens up. It allows us to analyze, at scale, what happens on stage during stand-up performance, while also taking editing into account. This is not a trivial question for an art form whose conditions of success remain partly elusive. Laughter, whose mastery is a core skill of stand-up, has long been treated as a complex philosophical and aesthetic problem. Earlier philosophical accounts of laughter and the comic already point to dimensions that motivate our focus on timing, expectation, embodiment, and rupture. Kant famously defines laughter as arising from "the sudden transformation of a strained expectation into nothing" (Kant, 1914, Part I, sec. 54). Bergson, in turn, offers the influential image of "something mechanical encrusted on the living" (Bergson, 1911, ch. I). The very notion of timing raises the broader difficulty of defining time itself, famously formulated by Augustine: "What then is time? If no one asks me, I know: if I wish to explain it to one that asketh, I know not" (Augustine, Book XI, ch. XIV). This difficulty of definition, central to philosophy, is also the problem we engage with here through the proposed operationalization of comedic timing.

Turning to prior work in computational humor and digital humanities, multimodal modeling of live performance is still mostly missing. Computational humor has largely centered on *textual* humor and its assessment (Kalloniatis and Adamidis, 2025): detecting humorous passages (Weller and Seppi, 2019; Annamoradnejad and Zoghi, 2024; Yang et al., 2015; Cafiero and Puren, 2025; Hossain et al., 2020), ranking by funniness (Potash et al., 2017), and predicting offensiveness or controversy (Meaney et al., 2021).

Distant viewing offers a framework to tackle the conceptual and technical challenge to simulate human vision for artistic analysis (Arnold and Tilton, 2019). Affective and social computing provide tools for modeling prosody, gesture, and emotion, though primarily in human–machine interaction.

Movement analysis has been applied to iconography (Impett and Moretti, 2017), ergonomic gesture learning (Glushkova et al., 2023), and pose clustering in archival theater footage (Rau et al., 2023) but without linking gesture to speech or audience response.

Recent work explores how language models might

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 29–38

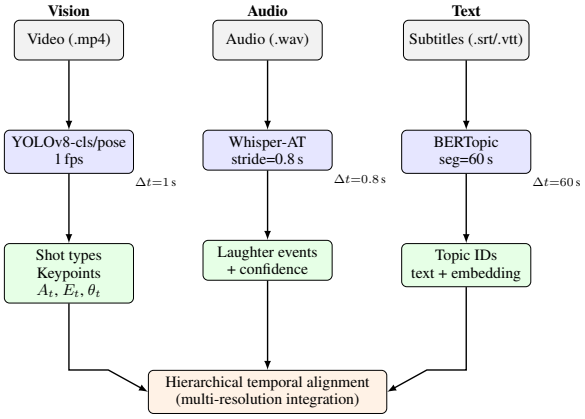

Figure 1: Multimodal processing pipeline. Each modality retains its native temporal resolution. Signals are merged by hierarchical temporal containment without resampling. Topic segments also store 384-dim sentence-BERT embeddings (`all-MiniLM-L6-v2`).

support stand-up writing, raising questions of prompt design, model bias, and cultural framing (Mirowski et al., 2024). The StandUp4AI dataset (Barriere et al., 2025) adds multilingual laughter labels, but focuses on audio.

Beyond text-based generation, live performance has been identified as a critical setting for evaluating computational humor under real-world temporal and social constraints. Mirowski et al. (2025) argue that improvised and staged comedy offer unique testbeds for studying the interaction between human performers, AI systems, and audiences, emphasizing that timing, embodiment, and audience feedback are inseparable from humor evaluation. While our focus differs from these experimental setups, we contribute a reproducible multimodal dataset of stand-up performances, designed to operationalize these performative dimensions (timing, gesture, and laughter) as measurable signals for computational analysis.

Recent work has similarly proposed an automatic method for the analysis of stand-up comedy, with particular attention to timing and performance dynamics. Using repeated recordings of the same routines performed on different nights by two comedians, Pope et al. (2026) analyze timing structures in live comedy through matched speech sequences across performances, demonstrating that the apparent spontaneity of stand-up relies on a complex craft of pacing and adaptation to audience context. Their study shares our interest in timing, performance structure, and audience response, but differs in both scale and design: it focuses on prosody, speech, and laughter placement in unedited live performances, whereas we analyze multimodal coordination within a large corpus of professionally recorded and edited stand-up comedy specials, examining themes, laughter placement, as well as the visual and bodily dimensions of performance through shot composition and pose-derived movement signals. Taken together, these contrasting approaches underscore how central yet elusive timing remains in the analysis of stand-up comedy.

In this paper, we present TIC-TALK: a multimodal corpus of 90 Netflix stand-up specials and a documented processing pipeline aligning text, audio, and vision without resampling. We report available performance indicators for the shot classifier and descriptive coverage metrics for the other streams, and include corpus-level cross-modal findings—notably a negative relationship between kinetic energy and laughter rate across topics—that validate the temporal alignment.

Our main contributions are:

1. A reusable multimodal **corpus** of 90 professionally edited stand-up specials with temporally aligned text, audio, and vision streams, including raw pose keypoints and per-segment sentence-BERT embeddings;

2. A transparent, reproducible **processing pipeline** (BERTopic, sentence-BERT, Whisper-AT, YOLOv8-cls, YOLOv8s-pose) with documented training choices and performance indicators;

3. **Three kinematic signals** (A_t, E_t, θ_t) derived from raw skeletal coordinates; and a **cross-modal use case** (Section 3.3) on laughter dynamics across 24 thematic topics, demonstrating: (i) kinetic energy is the strongest kinematic predictor of laughter ($r = -0.75$, $N = 24$), consistent with a stillness-before-punchline pattern; (ii) personal/bodily themes elicit systematically more laughter than geopolitical content; (iii) belly laughs are virtually absent at topic granularity, motivating event-level annotation; and (iv) shot close-up proportion correlates weakly with laughter rate ($r = +0.28$), consistent with reactive montage;

4. A **short-horizon laughter onset prediction benchmark** (Section 3.4): given multimodal context up to t, predict whether a new laughter event will begin in $[t, t+2)$; ablation over five feature sets (N=285,916 anchors, 90 shows, show-level train/val/test split) shows that temporal laughter history dominates prediction (AUROC = 0.643), that multimodal fusion achieves the best performance (AUROC = 0.647, AUPRC = 0.277 vs. random baseline 0.170), and that shot and pose features contribute marginal but consistent gains.

We first detail the processing pipeline per modality and the temporal alignment strategy (Section 2), then describe the corpus as a reusable resource and its data structure (Section 3). Section 3.3 presents a descriptive use case—laughter dynamics across thematic topics; Section 3.4 presents a predictive use case—short-horizon laughter onset prediction. Limitations and biases are discussed in Section 4.

2 Processing Pipeline and Temporal Alignment

The dataset integrates four modality streams—text, audio, and vision (pose and shot)—each processed independently at its native temporal resolution before alignment into a unified hierarchical representation (Figure 1).

2.1 Audio: Laughter Detection

Laughter events were detected using **Whisper-AT**, a pretrained AudioSet-based audio tagging model (Gong et al., 2023). Inference was performed at a 0.8 s stride in a high-recall configuration. The model outputs class probabilities for multiple laughter types; contiguous positive windows were merged into continuous events. Each event is represented by start and end times in seconds, a label (`type`), and a confidence score. These events constitute high-resolution temporal anchors for audience response.

2.2 Text: Topic Segmentation

2.2.1 Data and time-based segmentation

We start from subtitle transcripts in `.srt` format, parsed with their start/end timestamps and normalized by removing markup and formatting codes, standardizing apostrophes, and collapsing whitespace. We then construct contiguous time blocks by concatenating consecutive subtitle lines until a target duration is reached; a new block starts at the next subtitle line whose start time exceeds the current block limit. We remove stopwords using the union of a standard English stopword list and a curated set targeting fillers/discourse markers.

2.2.2 Sentence embeddings

Each block is embedded with a sentence-transformer encoder (`all-MiniLM-L6-v2`). Embeddings are computed in batches and L2-normalized: for each block i we obtain an embedding $\mathbf{e}_i \in \mathbb{R}^d$ and set

$$\tilde{\mathbf{e}}_i = \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2}.$$

These normalized vectors are used both for training the topic model and for later topic assignment to 60-second blocks.

2.2.3 Topic modeling with BERTopic

We learn topics using BERTopic. We first apply UMAP to the normalized embeddings using cosine distance, with $n_{\text{neighbors}} = 15$, $n_{\text{components}} = 5$, and $min\_dist = 0.1$. The reduced representations are clustered with HDBSCAN ($min\_cluster\_size = 15$, $min\_samples = 5$). Topic representations are obtained from a unigram–bigram count vectorizer. The model is capped at 40 topics and we retain the top 10 words per topic for interpretation.

2.2.4 Model selection over training block size

Topic quality depends on the training granularity. We select the training block size from $\{120, 150, 180, 210, 240\}$ seconds. For each candidate size, we train a BERTopic model and compute three diagnostics based on the resulting topic assignments:

- the *number of discovered topics* K,
- the *largest-topic share* $s_{\max}$
- a *normalized topic entropy* H_{norm}.

Let n_k be the number of blocks assigned to topic $k \in \{1, \ldots, K\}$ and $N = \sum_{k=1}^{K} n_k$. Define $p_k = n_k/N$. We compute

$$s_{\max} = \max_{k \in \{1,\ldots,K\}} p_k$$

and

$$H_{\text{norm}} = \frac{-\sum_{k=1}^{K} p_k \log p_k}{\log K}.$$

We enforce two validity constraints to avoid degenerate solutions: $K \geq 10$ and $s_{\max} \leq 0.35$. Among valid candidates, we select the model maximizing a composite score:

$$S = H_{\text{norm}} + C_{\text{NPMI}} - 2\, s_{\max},$$

where C_{NPMI} is a topic coherence measure computed from the model's top words.

Topic coherence (NPMI). We compute coherence using normalized pointwise mutual information (NPMI) over a tokenized version of the preprocessed blocks (subsampling documents when necessary for efficiency). For a topic t, let W_t be its top-M words (here $M = 10$). For each pair $(w_i, w_j) \in W_t \times W_t$ with $i < j$, let $P(w)$ be the probability that a word appears in a document and $P(w_i, w_j)$ the probability that both appear in the same document (estimated from document co-occurrence counts). NPMI for a pair is

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}.$$

Topic coherence is the average over word pairs, and corpus-level coherence averages over topics:

$$C_{\text{NPMI}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left(\frac{2}{M(M-1)} \sum_{i<j} \text{NPMI}(w_i, w_j) \right),$$

where $\mathcal{T}$ is the set of non-outlier topics.

2.2.5 Topic assignment to 60-second segments

After selecting the training block size, we retrain the topic model on all blocks at that granularity. We then construct 60-second blocks for each show and assign a topic label to each block using the trained BERTopic model. Since HDBSCAN can label uncertain points as outliers (-1), we apply a three-step post-processing

procedure to reduce outliers (1) BERTopic outlier reduction using an embedding-based strategy followed by a c-TF–IDF-based strategy, (2) reassignment of remaining outliers by nearest topic centroid in embedding space, and (3) within-show temporal gap filling: if a single outlier is flanked by two identical non-outlier topics, it is replaced by that topic.

For the centroid step, we compute a centroid for each non-outlier topic k by averaging the normalized training embeddings assigned to k and re-normalizing:

$$\mathbf{c}_k = \frac{\frac{1}{|I_k|}\sum_{i\in I_k}\tilde{\mathbf{e}}_i}{\left\|\frac{1}{|I_k|}\sum_{i\in I_k}\tilde{\mathbf{e}}_i\right\|_2},$$

where I_k is the set of training blocks assigned to topic k. For an outlier segment with embedding $\tilde{\mathbf{e}}$, we compute cosine similarities $\tilde{\mathbf{e}}^\top \mathbf{c}_k$ and reassign it to $\arg\max_k \tilde{\mathbf{e}}^\top \mathbf{c}_k$ when the maximum similarity is at least 0.30.

This procedure yields a topic sequence at 60-second resolution for each show, together with interpretable topic descriptors derived from c-TF–IDF.

2.3 Vision: Model Training and Feature Extraction

The visual pipeline combines two complementary YOLOv8 models operating hierarchically: (1) a **YOLOv8-cls** network fine-tuned for shot classification, and (2) a pre-trained **YOLOv8s-pose** estimator used to extract body keypoints from full-body frames only (Maji et al., 2022). This design ensures both efficient processing and consistent framing for pose analysis.

Shot classification. Fine-tuned to recognize six shot types: *full shot, medium close-up, medium long shot, medium shot, other angles,* and *other*. The model was trained on 594 manually annotated frames and validated on 128 held-out samples (100 epochs, batch size 32, learning rate 1×10^{-3}). Validation yielded an average F1 = 0.91, with most confusion between adjacent framings (e.g., chest ↔ waist). Predicted shot labels later serve both as contextual information and as filters for pose extraction, keeping only full-body and frontal views.

Pose estimation. Filtered frames are processed by the pre-trained **YOLOv8s-pose** model, producing 17 body keypoints (COCO skeleton) per detected performer at 1 fps. Raw pixel-normalized coordinates for all 17 joints are stored without prior discretization, preserving the full geometric information for downstream analysis. Joints with no detection confidence are recorded as $(0, 0)$ and excluded from derived computations via a validity filter.

Kinematic signals derived from raw keypoints. Raw keypoints enable the computation of three continuous scalar signals per frame, each capturing a distinct dimension of performance dynamics. Let $\mathbf{p}_j(t)$ denote the 2D coordinates of joint j at time t, and let $J(t)$ be the set of valid joints at t.

Arm spread measures the lateral extension of the performer's gesture relative to shoulder width:

$$A_t = \frac{\|\mathbf{p}_{\mathrm{W}_1}(t) - \mathbf{p}_{\mathrm{W}_2}(t)\|}{\|\mathbf{p}_{\mathrm{S}_1}(t) - \mathbf{p}_{\mathrm{S}_2}(t)\|},$$

where $\mathrm{W}_1, \mathrm{W}_2$ are the wrists and $\mathrm{S}_1, \mathrm{S}_2$ the shoulders. A_t=1 corresponds to a neutral stance; A_t>2 indicates open or emphatic gestures.

Kinetic energy quantifies total body movement between consecutive frames, normalized by performer height (bounding-box height h):

$$E_t = \frac{1}{h}\sum_{j\in J(t)\cap J(t-1)} \|\mathbf{p}_j(t) - \mathbf{p}_j(t-1)\|.$$

This serves as a proxy for performance intensity, capturing transitions between high-agitation delivery and still, high-confidence pauses.

Trunk lean encodes the signed angle of the torso axis relative to vertical:

$$\theta_t = \arctan\left(\frac{x_{\text{hip}}(t) - x_{\text{sho}}(t)}{y_{\text{hip}}(t) - y_{\text{sho}}(t)}\right) \times \frac{180}{\pi},$$

where $x_{\text{sho}}, y_{\text{sho}}$ and $x_{\text{hip}}, y_{\text{hip}}$ are the midpoints of the shoulder and hip pairs respectively. Lateral leans may, for instance, be characteristic of character-mimicry and asides, providing a posture-level complement to motion energy.

All three signals are smoothed with a sliding window of 30 s to suppress frame-level noise before analysis.

2.4 Hierarchical Temporal Alignment

Modalities operate on distinct temporal resolutions:

$$\Delta t_{\text{laugh}} = 0.8\,\text{s}, \quad \Delta t_{\text{pose}} = 1\,\text{s},$$

$$\Delta t_{\text{shot}} = 1\,\text{s}, \quad \Delta t_{\text{topic}} = 60\,\text{s}.$$

To preserve native granularity, temporal alignment is performed through hierarchical containment rather than resampling.

Let each modality m produce a sequence of temporal events or segments. Topic segments serve as the *anchor* level: each topic block $b_j = [s_j, e_j)$ defines a temporal window into which higher-frequency events are assigned by strict containment:

$$e \text{ is assigned to } b_j \iff t_e \in [s_j,\, e_j),$$

where t_e is the event timestamp. This applies uniformly to all nested streams: pose keyframes (Δt=1 s), shot labels (Δt=1 s), and laughter events (Δt=0.8 s). Kinematic signals (A_t, E_t, θ_t) are derived from the raw keypoints within each block after alignment. Each topic segment also stores the 384-dimensional sentence-BERT embedding $\tilde{\mathbf{e}}_j$ computed during topic modeling (Section 2.2.5), enabling direct use for similarity retrieval or cross-show clustering without re-encoding.

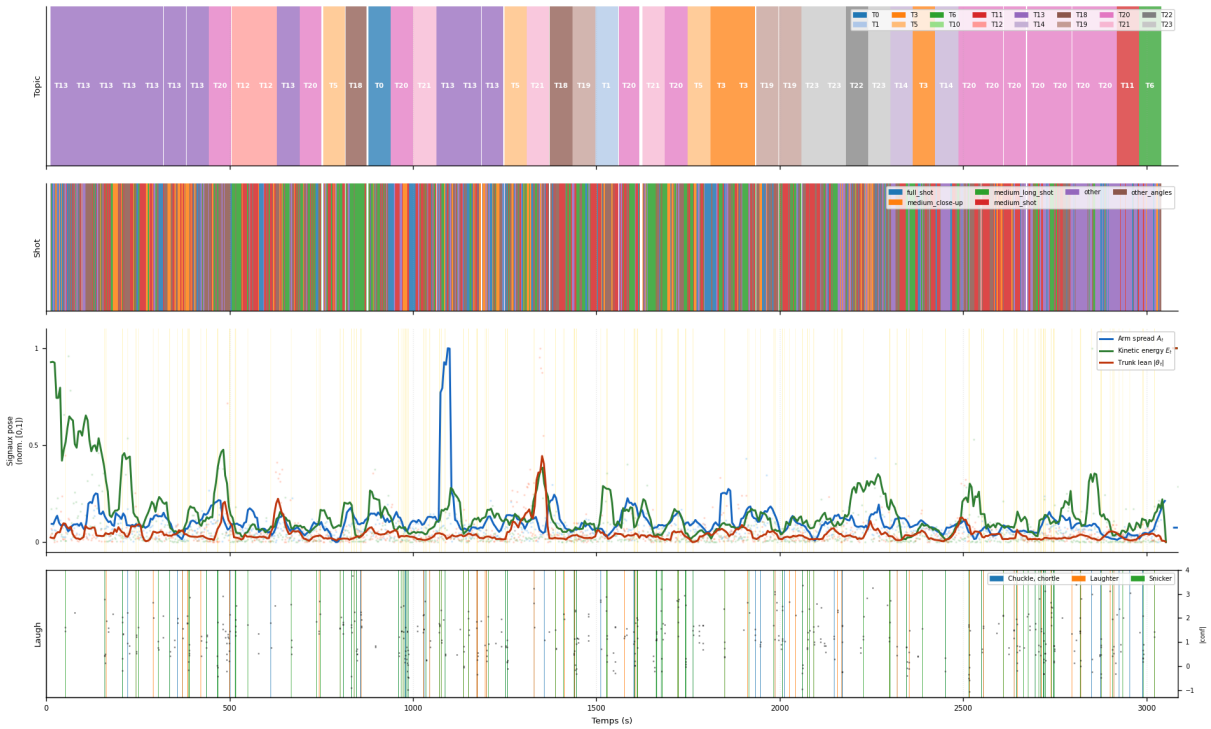

Figure 2: Example of aligned multimodal timelines for a show. Panels show from top to bottom: topic segments (BERTopic, 60 s blocks), shot-type predictions (1 Hz), three kinematic signals derived from raw pose keypoints (arm spread A_t, kinetic energy E_t, trunk lean θ_t, each normalized to $[0, 1]$; gold shading = laughter windows), and laughter events with confidence. The dense kinematic activity and prevalence of full-body shots reflect André's chaotic performance style; the alignment pipeline captures this without resampling across modalities.

3 Corpus Output and Statistics

Following the pipeline in Section 2, we release only *derived, time-aligned annotations* for 90 stand-up performances. No audio, image or video is distributed.

3.1 Delivered outputs

- **Topics**: 60 s segments with topic id, aggregated text, and sentence-BERT embedding.
- **Laughter**: contiguous events with start/end times, type label, and confidence score.
- **Shots**: one label per frame (1 Hz) from YOLOv8-cls.
- **Poses**: raw 17-joint (x, y) coordinates at 1 fps with bounding-box dimensions; no prior clustering.

3.2 Summary statistics

Average runtime per show is 63 min (total ≈94 h). The unified dataset contains 5,416 topic segments across 90 shows, each storing a 384-dim embedding. The visual stream covers 322,973 frames at 1 fps; 22% are full-body frames yielding raw keypoint sequences for pose analysis. Text yields ≈3,100 one-minute segments with non-outlier topic assignments.

3.3 Cross-modal Analysis: Laughter Dynamics as a Use Case

We ask whether thematic content, kinematic profile, and shot composition co-vary with audience laughter across the 24 BERTopic topics and 5,416 aligned blocks.

Method. For each topic block, we compute (i) the laughter rate r_ℓ, defined as the share of block duration covered by detected laughter events (mean coverage: 17.8%, ≈1.2 events/10 s); (ii) per-frame kinematic features (kinetic energy E_t, arm spread A_t, trunk lean θ_t); and (iii) shot-type proportions (full-body, close-up, medium). These are averaged per topic, weighted by block count, to obtain topic-level profiles. Pearson correlations are then computed between each feature and the mean laughter rate $\bar{r}_\ell$ across the 24 topics. Figure 3 presents the complete 24 topics × 10 features matrix as a hierarchical clustermap; Table 1 reports the six most contrasted topics.

Finding 1 — Kinetic energy negatively predicts laughter rate ($r = -0.75$). The strongest cross-modal signal across the 24 topics — after the quasi-tautological event count — is a *negative* correlation between mean kinetic energy $\bar{E}_t$ and laughter rate: Pearson $r = -0.75$, $N = 24$. Topics with the highest laughter rates all exhibit markedly low kinetic energy: T15 (body/dress, $\bar{E}_t = 1.13$, $\bar{r}_\ell = 0.253$), T22

Listing 1: Excerpt from the unified dataset (V2 structure).

```
{
  "ID_1": {
    "metadata": {
      "show_id": "SHOW_001",
      "n_blocks": 62,
      "embedding_dim": 384,
      "keypoint_joints": ["nose", "left_shoulder"
          , "...", "right_ankle"]
    },
    "timeline": [{
      "block_id": 58,
      "start": 3480.0, "end": 3540.0,
      "topic_id": 6,
      "text": "marriage gender roles ...",
      "embedding": [0.021, -0.143, "..."],
      "laugh_events": [
        {"start": 3482.4, "end": 3485.6,
         "type": "laughter", "confidence": 0.92}
      ],
      "pose_keypoints": [{
        "time": 3483.0, "has_detection": true,
        "bbox": {"xmin": 412, "ymin": 28,
                 "xmax": 895, "ymax": 716},
        "keypoints": {
          "left_shoulder": [634.2, 182.5],
          "left_wrist": [710.8, 480.3], "...": []
        }
      }],
      "shot_events": [
        {"time": 3483.0, "label": "full_shot",
         "class_id": 3, "score": 0.97}
      ]
    }]
  }
}
```

(baby/abortion, $\bar{E}_t = 1.05$, $\bar{r}_\ell = 0.248$), T10 (food, $\bar{E}_t = 1.19$, $\bar{r}_\ell = 0.230$), all below the corpus mean of 1.31. The low-laughter end is anchored by T11 (city/tonight/jewish, $\bar{E}_t = 1.65$, $\bar{r}_\ell = 0.121$) and the artefactual T6 ($\bar{E}_t = 2.24$, $\bar{r}_\ell = 0.051$).

This pattern is consistent with a *stillness-before-punchline* hypothesis: during high-laughter delivery, performers may reduce body movement and stabilize their posture to concentrate audience attention on the verbal content.

Finding 2 — A thematic hierarchy of funniness. Topic content stratifies audience laughter systematically. The four topics with the highest laughter rates ($\bar{r}_\ell > 0.20$) are all personal and bodily in register: physical appearance (T15, $\bar{r}_\ell = 0.253$), reproductive transgression (T22, $\bar{r}_\ell = 0.248$; also the highest *has_laughter* rate across the corpus at 0.869), everyday life (T10, $\bar{r}_\ell = 0.230$), and romantic relationships (T17, $\bar{r}_\ell = 0.222$). By contrast, geopolitical and identity-framing topics generate substantially less laughter: T3 (trump/india, $\bar{r}_\ell = 0.141$) and T1 (iceland, $\bar{r}_\ell = 0.085$, 50 blocks concentrated in a single show). This replicates content-level funniness gradients documented in text-only humor classification (Yang et al., 2015; Annamoradnejad and Zoghi, 2024), anchoring them in live audience response at corpus scale.

Finding 3 — Belly laughs are quasi-absent at topic granularity. The deepest laughter category (*belly laugh*, AudioSet class 20) is effectively absent across all 24 topics: only T22 ($\hat{r}_{\text{belly}} = 0.0051$) and T17 ($\hat{r}_{\text{belly}} = 0.0061$) register non-zero values. Two non-exclusive explanations apply: (i) the Whisper-AT classifier may be conservative on this class (it represents fewer than 2 events per 74,000+ inference windows across the corpus); (ii) belly laughs are genuinely triggered by specific delivery moments rather than by sustained thematic content, making them invisible at 60 s block granularity. Either interpretation suggests that capturing deep, distinctive laughter requires event-level annotation at a finer temporal resolution.

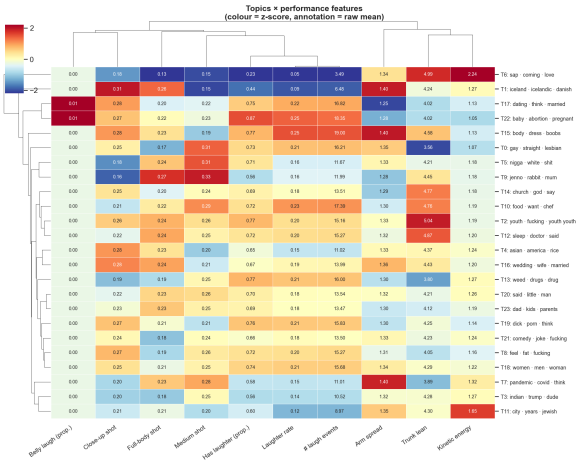

Figure 3: Hierarchical clustermap of 24 BERTopic topics × 10 performance features (z-scored row-wise). Colour encodes standardized deviation from the per-feature mean. The dominant pattern separates a low-$\bar{E}_t$ / high-$\bar{r}_\ell$ cluster (top: T15, T22, T10 — personal/bodily topics) from a high-$\bar{E}_t$ / low-$\bar{r}_\ell$ cluster (bottom: T6, T11). T6 (sap/mae/hello) is a structural artefact corresponding to subtitle encoding markers and on-stage entry sequences; it should be excluded from content-level comparisons.

Topic	**n**	$\bar{r}_\ell$	$\bar{E}_t$	$\bar{A}_t$
T15: body / dress / boobs	97	.253	1.13	1.40
T22: baby / abortion	198	.248	1.05	1.28
T10: food / want / chef	210	.230	1.19	1.30
T3: indian / trump / india	197	.141	1.27	1.32
T1: iceland / icelandic	50	.085	1.27	1.40
T6†: sap / mae / hello	81	.051	**2.24**	1.34

Table 1: Six most contrasted topics by mean laughter rate $\bar{r}_\ell$ (share of block duration covered by laughter events). $\bar{E}_t$: mean kinetic energy (normalized joint displacement between consecutive frames). $\bar{A}_t$: mean arm spread (wrist-to-wrist / shoulder-to-shoulder ratio). †T6 is a structural artefact; see Section 4.

Finding 4 — Shot composition and reactive montage ($r = +0.28$). Close-up shot proportion shows a weak positive correlation with laughter rate ($r = +0.28$, $N = 24$). High-laughter topics T15 and T22 both display above-average close-up ratios (0.278 and 0.265

respectively), consistent with *reactive montage*: directors increase close-up coverage during high-laughter passages, foregrounding the performer's facial expression at punchline delivery. A notable exception is T1 (iceland/icelandic), which exhibits the highest close-up ratio in the corpus (0.315) despite a low laughter rate ($\bar{r}_\ell = 0.085$), indicating that this association is partially confounded by performer- and show-level filming conventions.

3.4 Short-horizon Laughter Onset Prediction

Given the multimodal stream up to time t, can we predict whether a new laughter event will begin in the next $\delta = 2$ s?

Task and experimental setup. For each show, we sample anchor points at 1 s steps, excluding moments inside an ongoing laughter event. This yields 285,916 anchors across 90 shows; the positive rate is 17.0% (a new onset occurs in the next 2 s). Shows are split at the group level (GroupShuffleSplit): 62 shows for training, 14 for validation, 14 for test—no show appears in more than one split. A HistGradientBoostingClassifier is trained with balanced sample weights; the decision threshold is tuned on validation by maximizing F1.

Three feature groups are defined. **History** (10 scalars): from laughter events in the past 10 s—event count, rate, coverage, maximum event duration, mean and maximum confidence, coverage in the last 2 s and 5 s, time since last onset, time since last end. **Text** (64 scalars): the current topic block's 384-dim sentence-BERT embedding, reduced to 64 dimensions via PCA fitted on the training set only. **Vision** (20 scalars): shot proportion histogram over 6 classes, shot change rate, mean shot confidence, and 12 pose scalars (arm spread mean/std/max/trend, trunk lean mean/std, kinetic energy mean/std/max/trend, detection rate)—all aggregated over the past 10 s window.

System	AUROC	AUPRC	F1	Prec	Rec
history-only	0.643	0.275	0.336	0.223	0.682
text-only	0.554	0.197	0.297	0.177	0.926
vision-only	0.538	0.187	0.291	0.178	0.806
text + vision	0.577	0.210	0.300	0.182	0.867
text + vision + history	**0.647**	**0.277**	**0.342**	**0.248**	0.553
Random (AUPRC baseline)	—	0.170	—	—	—

Table 2: Short-horizon laughter onset prediction: ablation over feature groups. Positive rate = 0.170. Test set: 45,894 anchors from 14 held-out shows. AUPRC of a random classifier equals the positive rate. Threshold tuned on validation for F1/precision/recall.

Findings. First, temporal laughter history is by far the strongest individual predictor (AUROC = 0.643), leaving only marginal room for the other modalities: the best multimodal system (text+vision+history) improves AUROC by only 0.004 over history alone. This reflects a *temporal auto-correlation* of audience laughter: a hot room stays hot, independently of what is being said or shown. Second, vision-only is the weakest unimodal system (0.538), but combining it with text (0.577) outperforms both text-only (0.554) and vision-only—a consistent, if small, multimodal synergy. Third, adding text and vision to history improves precision from 0.223 to 0.248 while moderately reducing recall—the full model is less trigger-happy, reducing false positives. Fourth, the overall performance level (AUPRC = 0.277 vs. random 0.170, a $1.6\times$ lift) indicates that laughter onset is predictable above chance from a 10 s window, but far from deterministic: the stochastic nature of audience response and the 60 s temporal granularity of topical context both limit the ceiling of short-horizon prediction.

3.5 Data Structure and Access

Annotations are serialized as a hierarchical JSON per show (Figure 1); each topic block stores its four aligned streams, enabling direct temporal queries without resampling.

3.6 Multimodal Visualization Examples

Figure 2 illustrates the aligned multimodal timeline for one show; analogous visualizations for all 90 specials are distributed with the corpus as an interpretability and consistency check for the alignment pipeline.

4 Discussion

The descriptive use case (Section 3.3) yields four findings. (1) The E_t–laughter anti-correlation ($r = -0.75$) is consistent with a *stillness-before-punchline* pattern, but remains correlational: filming conventions, performer mobility, and the artefactual T6 are plausible confounders; event-level replication (kinetic energy in the 5 s before vs. after laughter onset) would be a stronger test. (2) Personal/bodily topics outperform geopolitical ones on laughter rate, replicating text-only funniness gradients at the level of live audience response and suggesting thematic content is a first-order predictor independently of delivery style. (3) The near-absence of belly laughs at 60 s granularity motivates finer annotation: deep laughter is likely tied to specific delivery moments invisible at block level. (4) The shot composition signal ($r = +0.28$), coherent with reactive montage, is partially confounded by show-level filming conventions.

The predictive use case (Section 3.4) adds three observations. (5) Laughter auto-correlation (history-only AUROC = 0.643) accounts for most predictable variance, consistent with crowd contagion (Provine, 1992). (6) Text and vision contribute marginally to precision (0.248 vs. 0.223), limited by 60 s block granularity; sentence-level and frame-level features would likely yield stronger gains. (7) The modest ceiling (AUPRC = 0.277 vs. 0.170 random) reflects the inherent stochasticity of audience response, absent from our single-recording specials.

Together, these results show that hierarchical temporal alignment enables both descriptive and predictive

cross-modal analyses while exposing their respective limits. The resource's value lies in enabling comparable, interpretable measurements of multimodal synchrony at corpus scale rather than absolute claims about funniness. Because all shows are professionally edited Netflix specials, platform conventions are embedded in the signal; comparisons are most reliable within similarly produced shows. Next steps include event-level annotation, prosodic features, and intra-comedian analyses to disentangle performer style from content.

5 Limitations

A first limitation concerns the corpus. This study focuses on Netflix stand-up comedy specials. Such specials do not represent stand-up comedy in its totality, but they correspond to a highly visible and professionally recognized form of the genre. Being selected for a platform such as Netflix is itself a strong marker of symbolic recognition within the field. Understanding more precisely what this selection entails would require further work on comedians' career trajectories, platform curation, and the industrial criteria through which such specials are produced and distributed.

A second limitation concerns what this type of quantitative analysis can tell us about comic intent and success. To what extent quantitative and computational analysis can help identify "what works" in stand-up comedy remains an open question. There is arguably a culture of informal quantitative analysis among comedians and comedy audiences, sometimes drawing comparisons with other performing arts such as dance or music. Yet, as with these art forms, technical description and training cannot fully account for artistic singularity, comic intuition, or poetic inclination.

A third limitation concerns the level at which jokes are represented. In this article, topic modeling is used to capture broad thematic patterns across the corpus, but the internal structure of jokes is not explicitly operationalized. This does not mean that we reduce a joke to its theme. Rather, thematic segmentation provides a workable level of description for a large-scale study focused on timing, stage dynamics, and cross-modal alignment. Finer-grained units such as setups, punchlines or callbacks would require a different annotation scheme and a more local temporal resolution. Our approach therefore treats topic as one useful layer of comic material, without claiming to model the full rhetorical or dramaturgical structure of jokes.

A fourth limitation concerns the nature of filmed stand-up specials as audiovisual objects. In this article, editing is treated as part of the object under study, which is not only stand-up as a theatrical and embodied practice, but stand-up video. A stand-up comedy special is not a neutral record of a stand-up performance: its montage constitutes another layer of artistic work. A more specific analysis of editing practices in stand-up specials would be needed to distinguish more clearly between stage performance and audiovisual construction, and would also inform another aspect of how humor is crafted for the screen. Marty Callner, often credited with pioneering the modern stand-up comedy special, described his approach by saying: "I learned the comedy directs me" (Zinoman, 2022). We did not treat editing as mere noise, but rather as part of the interplay between body, camera choices, thematic content, and live audience response.

6 Conclusion

This lightweight and modular pipeline provides an effective, reusable, and scalable framework for modeling the multimodal structure of stand-up comedy — a form whose apparent simplicity belies the artistry and craftsmanship of its performers. By operationalizing core dimensions such as gesture, timing, and audience response — a conceptual and technical challenge — and thus offering a model of stand-up performance, it supports both systematic analysis and critical reflection on what remains beyond computation.

Future work. While our pipeline was applied across a wide variety of performers and stand-up comedy specials, future experiments could focus on multiple video recordings by the same comedian, in order to track stylistic evolution over time and test the modularity of performance elements, a relevant hypothesis for stand-up, where long-form shows are often assembled from recombined short routines. Such a project would also open broader questions about stand-up archival practices, including both self-archiving by comedians and institutional archiving by clubs, venues, or platforms. Capturing original data would make it possible to reduce some of the audio and visual noise inherent in existing recordings, and to focus more directly on live performance rather than on professionally edited audiovisual objects. It would also open new avenues toward the study of more local, ephemeral, or amateur practices, anchored in specific socio-geographic contexts. In this work, we did not use our method to test a pre-formulated theory of which techniques elicit the most laughter. However, our model and pipeline could support such an approach in future work, especially if combined with an artist-centered perspective: stand-up comedians are often theorists of their own craft, and their practical expertise would be valuable for interpreting computational patterns of timing, delivery, and audience response.

7 Code and Data Availability

All code used for processing and analysis is available at the following repository: https://github.com/yaelle-z/TIC-TALK.

The derived dataset will be made available on Hugging Face at: https://huggingface.co/datasets/ENC-PSL/TIC-TALK.

8 Acknowledgements

We are thankful to participants at the AV in DH workshop (George Mason University), at the Sciences Po Medialab seminar and at the Bridging Computational Humanities and Computational Social Science Workshop (Ecole nationale des chartes - PSL) for their insightful remarks. We particularly thank Sylvaine Guyot and Jean-Philippe Cointet for their guidance. We also thank Dan Tirel for technical advice. This study was conducted as part of the DH master's program at École nationale des chartes–PSL and with the support of the PSL Research University's Major Research Program CultureLab, implemented by the ANR (reference ANR-10-IDEX-0001). Claude AI was used to assist with condensing and proofreading the paper. Any remaining mistakes are our own.

9 Bibliographical References

References

Issa Annamoradnejad and Gohar Zoghi. 2024. Colbert: Using BERT sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249:123685.

Taylor Arnold and Lauren Tilton. 2019. Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*, 34(Supplement_1):i3–i16.

Augustine. *Confessions*. Book XI, Chapter XIV. Online edition, Georgetown University.

Valentin Barriere, Nahuel Gomez, Leo Hemamou, Sofia Callejas, and Brian Ravenet. 2025. Standup4ai: A new multilingual dataset for humor detection in stand-up comedy videos. *arXiv preprint arXiv:2505.18903*.

Henri Bergson. 1911. *Laughter: An Essay on the Meaning of the Comic*. The Macmillan Company, New York.

Florian Cafiero and Marie Puren. 2025. A riddle in a haystack: Llm detection of intricate wordplays in colette and willy's novels for authorship attribution. In *Digital Humanities 2025*, Lisbon, Portugal.

Alina Glushkova, Dimitrios Makrygiannis, and Sotiris Manitsaris. 2023. Interactive sensorimotor guidance for learning motor skills of a glass blower. In Unknown, editor, *Culture and Computing*, volume 13933 of *Lecture Notes in Computer Science*, pages 29–43. Springer.

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Proceedings of INTERSPEECH 2023*, pages 2798–2802, Dublin, Ireland. ISCA.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. Semeval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Leonardo Impett and Franco Moretti. 2017. Totentanz: Operationalizing aby warburg's pathosformeln. *New Left Review*, (107):68–97.

Antonios Kalloniatis and Panagiotis Adamidis. 2025. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58:43. Article 43.

Immanuel Kant. 1914. *The Critique of Judgement*, 2 edition. Macmillan, London. Part I, Section 54.

Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. 2022. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2637–2646.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Piotr Mirowski, Kory Mathewson, and Boyd Branch. 2025. The theater stage as laboratory: Review of real-time comedy LLM systems for live performance. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 88–95, Online. Association for Computational Linguistics.

Piotr W. Mirowski, Juliette Love, Kory W. Mathewson, and Shakir Mohamed. 2024. A robot walks into a bar: Can language models serve as creativity support tools for comedy? an evaluation of LLMs' humour alignment with comedians. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pages 1622–1636, Rio de Janeiro, Brazil. Association for Computing Machinery. See also arXiv:2405.20956 for open-access preprint.

Vanessa C Pope, Rebecca Stewart, and Elaine Chew. 2026. Timing structures in live comedy: A matched-sequence approach to mapping performance dynamics. *PNAS Nexus*, 5(1):pgaf394.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Semeval-2017 task 6: #hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

Robert R. Provine. 1992. Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles. *Bulletin of the Psychonomic Society*, 30(1):1–4.

Michael J. Rau, Peter Broadwell, Simon Wiles, and Vijoy Abraham. 2023. Ai-assisted performance analysis: Deep learning for live and archival theater. In *Digital Humanities 2023: Book of Abstracts*, Graz, Austria. Centre for Information Modelling — Austrian Centre for Digital Humanities, University

of Graz. ADHO Digital Humanities Conference (DH2023), 10–14 July 2023.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Jason Zinoman. 2022. He might be the most influential director you've never heard of. *The New York Times*.

Arabic Humor as a Diagnostic Probe for Large Language Models

Wajdi Zaghouani
Communication Program
Northwestern University in Qatar
Doha, Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Arabic humor provides a challenging diagnostic test for large language models because interpreting jokes often requires pragmatic inference, sociolinguistic awareness, and culturally grounded knowledge that standard NLP benchmarks do not evaluate. Arabic is particularly suitable for probing these abilities given its diglossic structure and dialect diversity, where humor frequently arises from register contrast, dialect-specific vocabulary, and shared cultural references. We propose a three-layer taxonomy of Arabic humor mechanisms covering pragmatic, semantic, and sociolinguistic phenomena, illustrated through thirteen curated examples spanning Egyptian, Levantine, Gulf, Tunisian, and Iraqi Arabic. Building on this taxonomy, we introduce a diagnostic evaluation framework using contrastive minimal pairs, a multi-dimensional scoring rubric, and a cultural presupposition ontology. A small proof-of-concept probing study with GPT-4o, Gemini 2.0 Flash, and Claude Sonnet 4.5 reveals recurring failure patterns in sarcasm interpretation, register contrast reasoning, dialectal vocabulary coverage, and cultural grounding. We position this work as a diagnostic framework and pilot, not a mature benchmark, and outline a path toward larger annotated resources.

1 Introduction

A good diagnostic test for large language model (LLM) competence should probe capabilities that standard benchmarks leave untested and produce failures that are informative about specific, identifiable gaps. We argue that Arabic humor satisfies both criteria. Interpreting a joke in Arabic requires pragmatic inference, diglossic register reasoning, and culturally grounded knowledge (Raskin, 1985), three dimensions that reasoning, knowledge, and safety benchmarks do not directly evaluate. When a model fails on Arabic humor, the failure reveals whether it lacks pragmatic inference, dialect-register awareness, or cultural background knowledge.

LLMs achieve impressive results across NLP benchmarks, yet humor remains a persistent challenge. Jentzsch and Kersting (2023) found that over 90% of ChatGPT-generated jokes were drawn from just 25 templates, revealing template reproduction rather than genuine understanding. Sarcasm detection similarly shows LLMs underperforming fine-tuned models when meaning depends on pragmatic context (Abu Farha et al., 2022).

Arabic makes the diagnostic sharper. Its diglossic structure (Ferguson, 1959) creates humor from the contrast between Modern Standard Arabic (MSA) and colloquial dialectal punchlines, a mechanism absent from English-centric humor research. Despite major Arabic NLP resources for dialect identification and sarcasm detection (Abu Farha and Magdy, 2020; Abdul-Mageed et al., 2021; Bouamor et al., 2018), humor mechanisms beyond binary sarcasm remain understudied. Arabic LLMs exhibit "pragmatic literalism" and cultural context gaps even in MSA settings (Al-Olimat and Alshareef, 2026); our work extends this into the more demanding domain of dialectal humor.

We position this paper deliberately as a diagnostic framework and pilot study rather than a mature benchmark. The thirteen examples and the small probing study should be read as a proof of concept showing how Arabic humor can surface specific LLM failure modes, not as a finalised resource for model comparison. Our contributions are: (1) a three-layer taxonomy of Arabic humor mechanisms cov-

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 39–50

ering pragmatic, semantic, and sociolinguistic phenomena; (2) thirteen curated examples across five dialect regions with explicit mechanism assignment criteria; (3) a probing study on GPT-4o, Gemini 2.0 Flash, and Claude Sonnet 4.5 demonstrating four failure categories; (4) a worked example of contrastive minimal pairs; and (5) a concrete evaluation protocol for future corpus development.

2 Related Work

2.1 Theories of Humor and Computational Approaches

Raskin (1985) introduced the Semantic Script Theory of Humor (SSTH): a text is humorous when compatible with two opposing semantic scripts and a trigger switches interpretation between them. Attardo and Raskin (1991) extended this into the General Theory of Verbal Humor (GTVH), adding knowledge resources for narrative strategy, logical mechanism, target, situation, and language. Our taxonomy organises mechanisms by the type of reasoning required to recover the script switch and is one operationalisation of GTVH's logical-mechanism and language resources for an NLP setting, not a claim that these three layers exhaust humor theory.

On the computational side, Mihalcea and Strapparava (2005) established humor recognition as a tractable NLP task; Miller et al. (2017) provided the first shared evaluation for pun detection in English, demonstrating that semantic ambiguity could be modelled at scale; Amin and Burghardt (2020) survey humor generation and identify the lack of theoretically grounded evaluation as a central weakness; and Jentzsch and Kersting (2023) confirmed that LLMs still rely on surface templates rather than genuine humor understanding, with over 90% of ChatGPT-generated jokes drawn from just 25 recurring patterns.

2.2 Arabic Pragmatic Benchmarks and Sarcasm Detection

Al-Olimat and Alshareef (2026) developed the Arabic Linguistic and Pragmatic Suite (ALPS), showing that Arabic LLMs exhibit systematic pragmatic literalism and fail to recover culturally specific meaning even in MSA. This directly motivates extending diagnostic evaluation to dialectal humor, where pragmatic and cultural demands compound. Abu Farha and Magdy (2020) constructed ArSarcasm by reannotating sentiment data; their BiLSTM baseline reached only F1 = 0.46, illustrating the difficulty of even binary sarcasm classification. The iSarcasmEval shared task (Abu Farha et al., 2022) confirmed the challenge persists with MARBERT (Abdul-Mageed et al., 2021), because dialectal variation interacts with ironic meaning in ways that pretraining does not address.

2.3 Dialect Identification and Multi-Task Arabic NLP

NADI (Abdul-Mageed et al., 2020) and MADAR (Bouamor et al., 2018) established dialect identification as tractable, but identifying a dialect label is a prerequisite for humor interpretation, not an end in itself: a system that labels an utterance as Egyptian Arabic but cannot reason about the pragmatic significance of a variety switch has not solved the relevant problem. Kaseb and Farouk (2022) introduced SAIDS, showing that informing sentiment analysis with dialect and sarcasm predictions improves overall sentiment performance, supporting our proposal for humor annotation that captures dialect identity, mechanism, and cultural presupposition simultaneously.

2.4 Arabic Diglossia and Code-Switching

Ferguson (1959) defined diglossia as the coexistence of a formal High (H) and informal Low (L) variety, with Arabic as the prototypical case. Code-switching humor is documented across many language pairs (Winata et al., 2023), but what is structurally distinctive in Arabic is that the H/L distinction is formalised through diglossia and available to all speakers as a shared, unmarked communicative resource. This salience makes register switching a default humor mechanism in Arabic rather than an individual identity marker.

3 Humor in Arabic: A Taxonomy

Drawing on SSTH and GTVH (Raskin, 1985; Attardo and Raskin, 1991) and on the examples in Section 4, we propose a taxonomy organising Arabic humor mechanisms into three

computationally relevant layers, defined by the type of reasoning required to interpret the humor: a pragmatic layer requiring inference about speaker intent, a semantic layer requiring resolution of lexical or conceptual ambiguity, and a sociolinguistic layer requiring shared awareness of variety norms and cultural presuppositions. Table 1 maps each mechanism to its primary linguistic trigger and computational challenge.

Theoretical grounding of mechanism names. The mechanism names in our taxonomy are not arbitrary. *Sarcasm*, *irony*, and *self-deprecation* are standard categories in the humor research literature and in Arabic-language NLP work on sarcasm (Abu Farha and Magdy, 2020; Abu Farha et al., 2022). *Expectation inversion* and *wordplay* correspond to what SSTH calls script opposition triggered by lexical or narrative pivots, where the punchline forces reinterpretation of the setup. *Metaphorical humor* and *hyperbole* are recognised cross-domain mapping and implausibility mechanisms within GTVH's logical-mechanism resource. *Dialectal lexical humor*, *register switching*, and *cultural references* correspond to what GTVH groups under the language and situation knowledge resources, refined here to reflect Arabic-specific phenomena documented in the dialect identification (Abdul-Mageed et al., 2020; Bouamor et al., 2018) and code-switching (Winata et al., 2023) literatures. This mapping makes the taxonomy traceable to established humor theory rather than emergent from the example set alone.

Mechanism assignment criteria. Each example is assigned a primary mechanism according to the dominant trigger. When multiple triggers are present (as in Example 7, where register switching co-occurs with a cultural reference), we assign the mechanism that is *necessary and sufficient* for the humorous effect: removing it while preserving propositional content destroys the humor, as Section 5 illustrates. A second trigger that amplifies but does not constitute the humor is noted as secondary. Future annotators working with broader corpora should flag items that resist single-layer assignment for adjudication.

Mechanism	**Linguistic Trigger**	**Computational Challenge**
Sarcasm / Irony	Polarity reversal	Pragmatic inference beyond literal polarity
Self-deprecation	Expectation gap	Narrative expectation tracking
Expectation inversion	Genre frame violation	Discourse-type world knowledge
Wordplay	Lexical ambiguity	Ambiguity resolution
Metaphorical humor	Cross-domain mapping	Conceptual metaphor detection
Hyperbole	Implausible quantity	Implausibility detection
Dialectal lexical	Dialect-specific cue	Dialect vocabulary coverage
Register switching	MSA to dialect shift	Sociolinguistic register modeling
Cultural references	Shared presupposition	Cultural background knowledge

Table 1: Taxonomy of Arabic humor mechanisms with primary linguistic triggers and computational challenges.

The three layers. *Pragmatic humor* arises when intended meaning differs from literal content and encompasses sarcasm and ironic praise, self-deprecating humor, and expectation inversion. *Semantic humor* exploits lexical or conceptual ambiguity and covers wordplay and lexical reframing, metaphorical humor (cross-domain mappings), and hyperbole. *Sociolinguistic humor* draws on shared awareness of language-use norms and includes dialectal lexical humor (variety-specific vocabulary), register switching (MSA setup, colloquial punchline), and cultural reference humor (food, religious practices, seasonal events, regional customs).

4 Analysis of Arabic Humor Examples

We analyse thirteen humorous expressions drawn from publicly circulating Arabic social media content (Twitter/X posts and widely shared memes from 2023 and 2024). Selection required that each example instantiate a distinct primary mechanism, be interpretable without speaker identity information, and that examples span at least five dialect regions. Each example was verified by the author and one additional Arabic-speaking colleague fa-

miliar with the relevant dialect.

4.1 Pragmatic Humor Examples

Example 1, sarcastic praise (Egyptian):
ما شاء الله يا عبقري...كسرت الكوب بعد ما قلتلك تمسكه كويس
"Mashallah, you are a genius…you broke the cup after I told you to hold it properly."
Primary mechanism: Sarcasm. *Masha'allah*, conventionally an expression of admiration, is deployed sarcastically after a careless error. Lexical polarity alone assigns positive sentiment, missing the context-dependent inversion. This example carries a secondary cultural dimension (the religious formula's broad pragmatic range in Arabic), but the primary trigger is polarity inversion rather than register switching.

Example 2, expectation inversion (Tunisian):
عملت ريجيم جمعة...نقصت 200 غرام من صبري
"I did a week's diet…I lost 200 grams of my patience."
Primary mechanism: Expectation inversion. The setup primes physical weight loss, then the punchline substitutes an emotional quantity, violating the semantic type constraint between mass entity and psychological state. All three models scored at or near zero on this example, making it the clearest instance of pragmatic inference failure in the dataset.

Example 3, ironic self-evaluation (Levantine):
اشتريت كتاب تطوير الذات...ولسه نفس الشخص
"I bought a self-improvement book…I am still the same person."
Primary mechanism: Sarcasm. The humor exploits the contrast between self-help genre promises and the complete absence of change, requiring pragmatic inference about discourse-type conventions rather than any textual cue of irony.

4.2 Semantic Humor Examples

Example 4, lexical reframing (Egyptian):
أنا مش بخيل...أنا اقتصادي
"I am not stingy…I am economical."
Primary mechanism: Lexical reframing. The negative term *bakhīl* is replaced with the positive *iqtisādī*. The humor lies in the transparent self-justification: the denial implicitly confirms the original characterisation.

Example 5, conceptual metaphor (Egyptian):
شغّل دماغك شوية
الدماغ: تم إيقاف الخدمة مؤقتاً
"Use your brain a little."
"Brain: service temporarily unavailable."
Primary mechanism: Metaphorical humor. The brain-as-tool metaphor is extended into the telecommunications service-notification register, producing cross-domain incongruity.

Example 6, hyperbole (Gulf):
الحر اليوم كأنه الشمس نازلة تشرب قهوة معنا
"The heat today is as if the sun came down to drink coffee with us."
Primary mechanism: Hyperbole. Extreme heat is personified through a domestic hospitality scenario specific to Gulf cultural norms.

4.3 Sociolinguistic Humor: Register Switching

Register switching is one of the most structurally distinctive Arabic humor sources: a formal MSA setup is deflated by a colloquial punchline, requiring models to treat language variety as a meaning-bearing dimension rather than a classification label (Abdul-Mageed et al., 2020; Bouamor et al., 2018). Current Arabic NLP systems can detect variety switches but are not equipped to reason about their pragmatic implications. The pilot results in Section 6 confirm this gap: Examples 7 and 8 produce the sharpest inter-model divergence in the dataset.

Example 7, MSA setup with Egyptian dialect punchline:
هل تفضّل تناول وجبة صحيّة؟
لا يا عم هاتلي كشري وخلاص
Setup (MSA): "Would you prefer to consume a healthy meal?"
Punchline (Egyptian): "No man, just bring me koshary and that's it."
Primary mechanism: Register switching. *Secondary:* cultural reference (koshary as Egyptian working-class street food). The register switch is necessary and sufficient; the food reference amplifies but does not constitute the humor.

Example 8, MSA setup with Levan-

tine punchline:
أرجو أن تتحلّى بالصبر
الصبر خلص من زمان
Setup (MSA): "Please adorn yourself with patience."
Punchline (Levantine): "Patience ran out long ago."
Primary mechanism: Register switching. A classical MSA rhetorical construction is undercut by colloquial Arabic treating patience as a depleted commodity.

Example 9, cultural reference (Gulf):
قالوا نعمل دايت
قلت بعد رمضان
"They said let's diet."
"I said: after Ramadan."
Primary mechanism: Cultural reference. "After Ramadan" signals indefinite deferral only if the listener knows Ramadan is immediately followed by Eid feasting, making the stated timing a humorous non-commitment rather than a concrete plan. All three models scored 1.3 here, the most consistent result among sociolinguistic examples, suggesting partial cultural knowledge of Ramadan but without full grounding in the deferral convention.

4.4 Additional Examples: Broadening Dialect Coverage

Example 10, sarcastic praise (Levantine):
يعني عنجد فكرة عبقرية...خلينا نأجل الشغل لآخر دقيقة
"Wow, truly a brilliant idea…let's postpone the work until the last minute."
Primary mechanism: Sarcasm. The positive framing introduces an obviously counterproductive plan, requiring contextual reasoning to invert surface polarity.

Example 11, expectation inversion (Egyptian):
ذاكرت طول الليل...عشان أنام في الامتحان مرتاح
"I studied all night…so I could sleep comfortably during the exam."
Primary mechanism: Expectation inversion. Academic performance is primed, then the punchline replaces the goal (passing) with its antithesis (comfortable sleeping).

Example 12, lexical reframing (Gulf):
أنا مو كسول...أنا موفّر للطاقة
"I am not lazy…I am energy efficient."
Primary mechanism: Lexical reframing. The negative *kasūl* is relabelled with the technical-environmental *muwaffir lil-ṭāqa*, an incongruous register shift that humorously rebrands a mundane failing.

Example 13, cultural reference (Iraqi):
قررت أبدأ حمية من اليوم...بس اليوم عزيمة
"I decided to start dieting today…but today we have a feast invitation."
Primary mechanism: Cultural reference. The joke depends on the norm of *'azīma*, a formal meal invitation that social obligation requires accepting. Without this knowledge the utterance reads as a simple scheduling conflict. This example scored lowest among cultural reference items (1.0 for GPT-4o and Claude, 1.3 for Gemini), with all models recognising the conflict between dieting and feasting but none articulating the specific social obligation that makes refusal culturally impossible.

Table 2 summarises all thirteen examples.

5 Contrastive Minimal Pairs: A Worked Illustration

Reviewers reasonably pointed out that minimal pairs are central to our framework but had not been worked out concretely in the original draft. Section 8 sets them out as a protocol for future corpus development. Here we present the principle in concrete form on one example so the rest of the paper can refer to a worked instance rather than an abstract proposal.

For Example 7 (the koshary register switch), three variants isolate the role of register switching:

Variant A: all-MSA, propositional content preserved.
هل تفضّل تناول وجبة صحيّة؟
لا، أفضّل تناول الكشري
Setup (MSA): "Would you prefer to consume a healthy meal?"
Punchline (MSA): "No, I prefer to consume koshary."
Effect: no register contrast, no humor. The cultural reference (koshary) is preserved, demonstrating that the cultural element alone is not sufficient.

Variant B: all-colloquial Egyptian.
عايز ماكلة صحية؟
لا يا عم هاتلي كشري وخلاص

Ex.	Arabic	English Translation	Dialect	Mechanism	Computational Challenge
1	ما شاء الله يا عبقري...كسرت الكوب	"Genius...broke the cup after being told"	Egyptian	Sarcastic praise	Polarity inversion
2	عملت ريجيم...نقصت 200 غرام من صبري	"Did a week's diet...lost 200 grams of patience"	Tunisian	Expectation inversion	Semantic type constraint
3	اشتريت كتاب تطوير الذات...ولسه نفس الشخص	"Bought self-help book...still same person"	Levantine	Ironic self-evaluation	Genre convention awareness
4	أنا مش بخيل...أنا اقتصادي	"Not stingy...economical"	Egyptian	Lexical reframing	Implicit admission via denial
5	شغّل دماغك...تم إيقاف الخدمة	"Use your brain / service unavailable"	Egyptian	Conceptual metaphor	Cross-domain metaphor
6	الحر كأن الشمس تشرب قهوة معنا	"Heat as if the sun came down for coffee"	Gulf	Hyperbole	Personification, implausibility
7	وجبة صحيّة؟...هاتلي كشري	"Healthy meal? / Just bring me koshary"	MSA + Egyptian	Register switching	Diglossia plus cultural (secondary)
8	تحلّى بالصبر...الصبر خلص	"Please be patient / patience ran out"	MSA + Levantine	Register switching	Register plus pragmatic inference
9	نعمل دايت...بعد رمضان	"Let's diet / after Ramadan"	Gulf	Cultural reference	Religious-cultural knowledge
10	فكرة عبقرية...نأجل الشغل لآخر دقيقة	"Brilliant idea...postpone work to last minute"	Levantine	Sarcastic praise	Contextual polarity inversion
11	ذاكرت طول الليل...عشان أنام في الامتحان	"Studied all night...to sleep in the exam"	Egyptian	Expectation inversion	Goal inversion tracking
12	أنا مو كسول...أنا موفّر للطاقة	"Not lazy...energy efficient"	Gulf	Lexical reframing	Register incongruity detection
13	أبدأ حمية...بس اليوم عزيمة	"Dieting today...but today there's a feast"	Iraqi	Cultural reference	Social obligation knowledge

Table 2: All analysed examples with Arabic source, English translation, dialect, primary humor mechanism, and computational challenge. Arabic text is abbreviated for the summary; full versions appear in Section 4. Example 7 carries a secondary cultural reference dimension noted in the text.

Setup (Egyptian): "Want a healthy meal?"
Punchline (Egyptian): "No man, just bring me koshary and that's it."
Effect: weaker humorous effect. The colloquial-to-colloquial pairing removes the register asymmetry; the punchline reads as ordinary preference rather than register-defying refusal.

Variant C: original (MSA setup, Egyptian punchline).
The register switch is restored, and the humorous effect returns in full.

This contrast supports the necessity and sufficiency criterion for mechanism assignment in Section 3: register switching, not the cultural reference, is the necessary trigger for the humor in Example 7. Section 8 generalises this protocol to sarcasm (where the trigger removal variant replaces the sarcastic frame with a direct statement) and to cultural references (where the loaded element is substituted with a culturally neutral equivalent). Validation of these contrasts via human funniness judgements is, as we acknowledge in the Limitations, a necessary next step before the framework can be used at scale.

6 Pilot Probing Study

To ground the four failure categories in observable model behaviour rather than intuition alone, we conducted a small-scale probing study using the thirteen examples from Section 4. We emphasise that thirteen items constitute a proof of concept rather than a sta-

tistically reliable evaluation; results are indicative and motivate larger-scale follow-up work, not direct model comparison or ranking.

We probed three widely used frontier LLMs representing different providers: GPT-4o (OpenAI, version `gpt-4o-2024-08-06`), Gemini (Google, version `gemini-2.0-flash`), and Claude (Anthropic, version `claude-sonnet-4-5`). All models were queried via their respective APIs using default decoding parameters (temperature = 1.0, top-p = 1.0) with no system prompt. Each model received the Arabic text of each example followed by the identical prompt: *"Explain why this Arabic utterance is humorous and identify the humor mechanism."* The high-temperature default was chosen to allow each model's typical generative behaviour to surface rather than to optimise for a single best response.

We deliberately focus on frontier LLMs rather than Arabic-specialised encoders such as MARBERTv2 or AraBERTv2 because our study targets open-ended explanation rather than classification. Encoder-only models are not equipped to produce free-text explanations of humor mechanisms and therefore cannot be evaluated on the same scoring dimensions; they remain an important complementary baseline for future classification-oriented tasks derived from this framework.

Scoring procedure. Responses were evaluated on three dimensions using a 0–2 scale. **Mechanism**: 0 if incorrect or absent, 1 if partially correct (for example, identifying irony without explaining the mechanism), 2 if correct with appropriate reasoning. **Register**: 0 if dialect or register is ignored, 1 if a variety difference is noted but not interpreted pragmatically, 2 if register contrast is linked to the humorous effect. **Cultural**: 0 if cultural presupposition is absent, 1 if partially recognised, 2 if clearly articulated. Two annotators (the author and one colleague familiar with Arabic NLP and humor analysis) independently evaluated all model outputs. Initial inter-annotator agreement, computed as Krippendorff's α over ordinal ratings, was $\alpha = 0.71$ (Mechanism), $\alpha = 0.68$ (Register), and $\alpha = 0.73$ (Cultural), indicating acceptable reliability before adjudication. Remaining disagreements were resolved through discussion. The reported per-example value is the mean of the three dimension scores after adjudication.

Ex.	Layer	GPT-4o	Gemini	Claude
1	Pragmatic	1.0	1.0	1.0
2	Pragmatic	0.3	0.0	0.0
3	Pragmatic	0.7	0.3	0.7
4	Semantic	0.7	0.7	0.7
5	Semantic	0.7	0.7	0.7
6	Semantic	0.7	0.7	0.0
7	Socioling.	1.7	1.7	0.7
8	Socioling.	0.0	0.0	1.3
9	Socioling.	1.3	1.3	1.3
10	Pragmatic	0.7	0.7	0.7
11	Pragmatic	0.7	0.7	0.7
12	Semantic	0.7	0.7	0.7
13	Socioling.	1.0	1.3	1.0

Table 3: Mean scores (0–2) across the three scoring dimensions (Mechanism, Register, Cultural) for the thirteen probed examples. Per-example maxima vary by layer: approximately 0.67 for semantic-only examples, 1.33 for examples requiring two dimensions, and 2.0 for examples requiring all three. Scores should be read relative to these per-example ceilings, not against a uniform 2.0 maximum.

Interpreting the scores. Because not every example exercises every dimension, the three-dimension mean has a particular interpretation. For a purely semantic example (Example 4, lexical reframing), the maximum achievable score is 2 on Mechanism and 0 on Register and Cultural, yielding a maximum mean of approximately 0.67. A score around 0.7 therefore reflects close-to-ideal performance, not mediocre performance. For a register-switching example (Example 7), a model that correctly identifies the mechanism and the register contrast but misses the cultural reference would receive (2, 2, 0), yielding a mean of approximately 1.3. The full 2.0 mean is reserved for examples whose interpretation requires all three dimensions. Results appear in Table 3.

Read relative to these per-example ceilings, semantic humor is handled close to its expected maximum: scores of 0.7 on Examples 4, 5, and 12 indicate that Mechanism was reliably captured while Register and Cultural remained at zero as expected. Pragmatic hu-

mor involving expectation inversion is interpreted inconsistently (Example 2 collapses to 0 or near 0 across all three models). Sociolinguistic humor presents the greatest challenge relative to its higher per-example ceiling: register-switching Examples 7 and 8 reveal that models sometimes detect dialectal variation but fail to connect the register contrast to the humorous mechanism, and the pattern of strengths reverses between the two examples across models. Cultural reference humor (9, 13) is sometimes recognised but explanations rely on generic social reasoning rather than explicit cultural grounding. These patterns align with ALPS (Al-Olimat and Alshareef, 2026) on pragmatic literalism and with SAIDS (Kaseb and Farouk, 2022) on dialect-sarcasm interactions in Arabic.

7 Implications for Large Language Models

The pilot results suggest four systematic categories of LLM failure, framed as hypotheses motivating larger studies rather than as conclusions established at the present scale.

7.1 Literal Interpretation of Sarcasm and Irony

Mashallah in Example 1 may be genuine admiration or deep sarcasm depending entirely on context; all three models scored 1.0, identifying irony but failing to explain the register-dependent mechanism. Example 2 scored at or near 0 across the three models, confirming that semantic type constraint violations go undetected. ArSarcasm (Abu Farha and Magdy, 2020) and iSarcasmEval (Abu Farha et al., 2022) both document this gap; even MARBERT (Abdul-Mageed et al., 2021) struggles because dialectal variation interacts with ironic meaning in ways pretraining does not address.

7.2 Dialect-Specific Vocabulary and Cultural Reference Gaps

The koshary reference (Example 7) requires knowing the food's pragmatic associations with working-class Egyptian culture; GPT-4o and Gemini scored 1.7, correctly identifying the register switch but offering only surface cultural explanation. Examples 9 and 13 scored 1.0 to 1.3, suggesting partial cultural knowledge, with explanations relying on generic social reasoning rather than the specific presuppositions involved.

7.3 Register Contrast Modeling

The register-switching examples reveal the sharpest inter-model divergence. On Example 7, GPT-4o and Gemini scored 1.7 while Claude scored 0.7. On Example 8 the pattern reversed: GPT-4o and Gemini scored 0.0 while Claude scored 1.3. This asymmetry suggests that dialectal training-data coverage shapes which register contrasts a model can reason about, and that the bottleneck is pragmatic reasoning over the shift rather than variety detection per se. Given the small sample, this is best read as a hypothesis to test on larger sets.

7.4 Pragmatic Inference Failures

Example 4 requires recognising that a denial can implicitly confirm an accusation; Examples 2 and 3 require tracking genre-specific expectations. Uniformly modest scores (0.3 to 0.7) align with documented LLM weaknesses in Arabic pragmatic inference (Al-Olimat and Alshareef, 2026). SAIDS (Kaseb and Farouk, 2022) shows that dialect and sarcasm signals improve downstream sentiment; a humor-focused multi-task architecture predicting humor type, register identity, and cultural presupposition simultaneously is a promising direction.

8 Discussion

8.1 Toward an Operationalised Evaluation Protocol

Building on the pilot study and the worked minimal-pair example in Section 5, we propose a concrete, operationalised evaluation protocol for future work.

Contrastive minimal pairs. For each mechanism, pair an original humorous utterance with a variant that removes the humor trigger while preserving propositional content. The koshary case in Section 5 illustrates the principle for register switching. For sarcasm, the trigger removal variant replaces the sarcastic frame with a direct statement (Example 1 becomes a literal complaint about a broken

cup). For cultural references, the loaded element is substituted with a culturally neutral equivalent (Example 9 becomes a generic deferral to an unspecified future date). Annotators should rate each pair for funniness so that an effect size can be reported per mechanism.

Scoring rubric. Large-scale evaluation should score Mechanism identification, Register recognition, and Cultural presupposition recovery (each 0 to 2), with worked examples per level to support inter-rater reliability and Krippendorff's $\alpha \geq 0.70$ before aggregation. As noted in Section 6, the mean across the three dimensions should be interpreted relative to a per-example ceiling determined by which dimensions are genuinely required, not against a uniform 2.0 maximum.

Cultural presupposition ontology. Presuppositions should be classified into five coarse domains: (1) food and hospitality, (2) religious practices and calendar, (3) social obligations and norms, (4) bureaucracy and institutions, and (5) generational or political references.

Predicted difficulty ordering. The pilot is consistent with the ordering sociolinguistic and cultural reference humor harder than pragmatic humor harder than semantic humor, when each is scored relative to its appropriate ceiling. Cultural reference examples (9, 13) and register-switching Example 8 fall short of their ceilings most consistently; semantic examples (4, 5, 12) reach close to their per-example maxima. Future corpora should report per-mechanism accuracy and per-mechanism ceiling-relative scores to track this ordering over time.

Annotation guidelines. Each corpus item should include: dialect label, primary and secondary mechanism categories, a non-humorous paraphrase preserving propositional content, and a cultural presupposition field. Multi-label cases where two mechanisms are jointly necessary should be flagged for adjudication.

Connections and extensions. ALPS (Al-Olimat and Alshareef, 2026) provides the template for diagnostic Arabic probing; our protocol extends it to dialectal humor, and SAIDS (Kaseb and Farouk, 2022) motivates joint evaluation of dialect, pragmatics, and humor type. Although the present study uses Arabic, the framework itself transfers naturally to other diglossic or dialect-rich language families. Future work should explore retrieval-augmented prompting with cultural knowledge snippets and multimodal extensions using Arabic ASR for prosodic register cues in spoken humor.

9 Conclusion

We argued and demonstrated that Arabic humor functions as a principled diagnostic probe for LLMs, surfacing gaps in pragmatic inference, diglossic register reasoning, and cultural grounding that standard benchmarks do not reach. A three-layer taxonomy across five dialect regions was illustrated through a small probing study on GPT-4o, Gemini 2.0 Flash, and Claude Sonnet 4.5. When scored relative to per-example ceilings, semantic humor reaches close to its maximum, while sociolinguistic humor produces the most variable results, with register-switching examples revealing sharp inter-model divergence tied to dialectal training-data coverage. We operationalised an evaluation protocol with contrastive minimal pairs (worked out concretely for one example), a three-dimensional scoring rubric, a cultural presupposition ontology, and annotation guidelines. We position this contribution as a diagnostic framework and pilot study, not a finished benchmark, and view the construction of a large, dialectally balanced Arabic humor corpus as the natural next step.

Limitations

This study has several limitations that should be borne in mind when interpreting its contributions.

First, the probing study in Section 6, while covering all thirteen examples across three frontier LLMs, remains limited in scale. Thirteen items constitute a proof-of-concept demonstration rather than a statistically reliable empirical evaluation. The mean scores in Table 3 should be interpreted as indicative rather than definitive, and the Krippendorff α values reported (0.68 to 0.73) fall close to the conventional 0.70 threshold, meaning the scoring dimensions require further validation

on a larger item set before they can be used for high-stakes model comparison. We also note that the stability of open-ended explanation evaluation across runs has not been measured here; future work should report variance across multiple samples per item and ideally across multiple prompt formulations.

Second, the pilot does not include Arabic-specialised encoder models such as MARBERTv2 or AraBERTv2. These models are not capable of producing free-text explanations of humor mechanisms and therefore cannot be directly evaluated under the same three-dimensional rubric. However, they represent important baselines for classification-oriented tasks, and future work should compare frontier LLM explanation quality against encoder-based classification performance to establish whether the gaps observed here are specific to general-purpose LLMs or persist for Arabic-pretrained systems.

Third, the necessity and sufficiency criterion for mechanism assignment, while operationally useful and illustrated by the worked minimal-pair example in Section 5, has not been validated through human funniness judgments at scale. Ideally, minimal-pair ablation studies would collect funniness ratings from native speakers of the relevant dialect before and after trigger removal, reporting effect sizes to demonstrate that removing the assigned trigger reliably suppresses the humorous effect. This validation is a necessary step before the taxonomy can be used to train annotation models or derive evaluation metrics.

Fourth, the thirteen examples were selected to illustrate the taxonomy rather than to constitute a statistically representative sample of Arabic humor. No formal inter-annotator agreement measurement was conducted for mechanism assignment across the full set. A large-scale corpus covering multiple dialects, humor types, and cultural reference categories, annotated by multiple native speakers with explicit reliability measures, is necessary for quantitative analysis of mechanism distribution and for training or evaluating computational models.

Fifth, while the example set spans five dialect regions, the Arab world encompasses many other varieties including Moroccan Darija, Sudanese, Yemeni, and Libyan Arabic. Moroccan Darija in particular, which draws heavily on Amazigh and French alongside Arabic, may exhibit humor dynamics that differ substantially from the MSA and colloquial contrast we foreground here.

Sixth, our treatment of cultural reference humor is necessarily incomplete. Cultural knowledge underlying humor is dynamic, generational, and community-specific. The examples represent a snapshot of humor circulating in 2023 and 2024, and the cultural presuppositions they rely on may shift over time.

Seventh, the study focuses exclusively on written text. Spoken Arabic humor involves prosodic and intonational cues that interact with register switching in ways that text alone cannot capture, as noted in the Discussion. The framework should be extended to spoken modalities as Arabic ASR resources for dialectal speech continue to develop.

Finally, the broader applicability of the framework to other languages, though plausible (and noted in Section 8), has not been tested empirically. The contrastive minimal-pair protocol and three-dimensional rubric are language-independent in design, but their utility for, say, Swiss German and Standard German diglossia, or for Mandarin and topolect contrasts, remains an open question.

Ethical Considerations

This research is primarily analytical and does not involve the collection of personal data, the construction of a new annotated dataset, or the deployment of a computational system. The humorous examples are drawn from publicly circulating informal discourse on social media platforms and do not identify specific individuals or communities in ways that could cause harm.

Humor in Arabic, as in any language, exists on a spectrum that includes not only benign wit but also content that may be offensive, stereotypical, or exclusionary. Our taxonomy focuses on humor mechanisms rather than humor targets, and we deliberately selected examples that illustrate linguistic and computational properties without targeting particular ethnic, religious, gender, or socioeconomic groups. Future work building computational systems for Arabic humor processing should

include explicit annotation guidelines for offensive and harmful humor, with careful attention to how such categories are defined across different Arabic-speaking communities.

Computational systems capable of detecting and interpreting humor in Arabic social media content could be misused, including for surveillance of political dissent, overreaching content moderation, or automated analysis of communications in ways that violate user expectations of privacy. Researchers and practitioners building on the framework proposed here should consider these downstream risks explicitly during system design and evaluation.

Finally, humor is deeply culturally situated, and interpretations of what is funny in a given context reflect the perspectives and lived experiences of individuals embedded in those cultural contexts. Annotated datasets for Arabic humor should include annotation from native speakers of the specific dialect variety in question rather than relying on annotators with general Arabic proficiency. Annotation tasks should be designed to surface disagreements rather than resolve them prematurely, since variation in humor interpretation across age groups, genders, regions, and communities is itself a scientifically meaningful signal rather than noise to be eliminated.

Acknowledgment

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar Development and Innovation Council (QRDI).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

Hussein S. Al-Olimat and Ahmad Alshareef. 2026. ALPS: A diagnostic challenge set for Arabic linguistic and pragmatic reasoning. *Preprint*, arXiv:2602.17054.

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 4(3–4):293–347.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

Abdelrahman Kaseb and Mona Farouk. 2022. SAIDS: A novel approach for sentiment analysis informed of dialect and sarcasm. In *Proceedings*

of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 22–30, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

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.

Victor Raskin. 1985. *Semantic Mechanisms of Humor*. D. Reidel Publishing Company, Dordrecht.

Genta Indra Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Cards Against LLMs: Benchmarking Humor Alignment in Large Language Models

Yousra Fettach[1], Guillaume Bied[1], Hannu Toivonen[2], Tijl De Bie[1]

[1] Ghent University, Belgium
[2]University of Helsinki, Finland

Abstract

Humor is one of the most culturally embedded and socially significant dimensions of human communication, yet it remains largely unexplored as a dimension of Large Language Model (LLM) alignment. In this study, five frontier language models play the same *Cards Against Humanity* games (CAH) as human players. The models select the funniest response from a slate of ten candidate cards across 9,894 rounds. While all models exceed the random baseline, alignment with human preference remains modest. More striking is that models agree with each other substantially more often than they agree with humans. We show that this preference is partly explained by systematic position biases and content preferences, raising the question whether LLM humor judgment reflects genuine preference or structural artifacts of inference and alignment.

1 Introduction

Large language models (LLMs) have made significant strides in coding (Chen et al., 2021), writing (Brown et al., 2020), and creative tasks such as storytelling and poetry generation (Chakrabarty et al., 2022). This is largely attributed to training on vast human-generated corpora, emergent capabilities unlocked by scale (Wei et al., 2022), and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). As these models approach human-level performance on structured tasks, attention has turned to whether they can match humans on dimensions of communication that are less formal and more culturally loaded.

Humor embodies one such dimension. It operates at the intersection of language, cognition, and social context (Martin and Ford, 2018). It also relies on the violation and resolution of expectations (Attardo, 1997), shared cultural knowledge, and sensitivity to timing, tone, and audience (Hay, 2001). These properties make it simultaneously deeply human and notoriously difficult to formalize (Mihalcea and Strapparava, 2005).

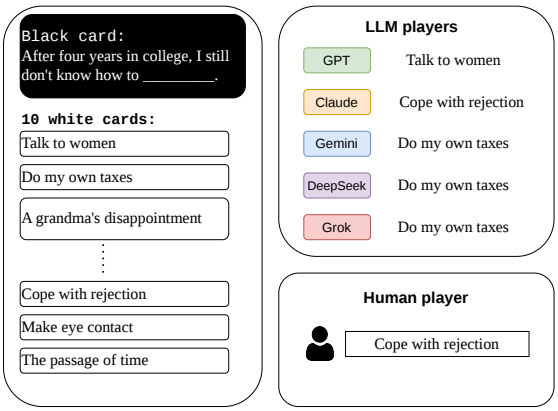

Figure 1: Framework overview. Given a black card prompt and a slate of 10 white card candidates, five frontier LLMs and a human player independently select the card they deem funniest.

This complexity makes humor a uniquely revealing test for LLMs. What a model finds funny reflects the cultural references it recognizes, the social boundaries it is willing to cross, and the moral and political values it has absorbed through training and alignment. Humor is therefore not only a capability to be scored but more of a window into the worldview of the model itself, and one of the final frontiers for AI alignment (Zhou et al., 2025). Probing this requires a setting that elicits genuine humor judgment while remaining structured enough for systematic comparison.

We study the alignment of LLMs with human humor through a controlled, game-based evaluation, operationally measuring it by comparing LLM-derived to human-derived preference judgments. *Cards Against Humanity* (CAH) offers a natural testbed for this inquiry. As a fill-in-the-blank party game built around subversive, context-sensitive punchlines, it requires players to select the funniest response to a prompt from a fixed set of options.

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 51–64

This makes it specific enough for systematic evaluation but also rich enough to expose meaningful variation in humor perception. We show our experimental setup in Figure 1. The following is a summary of our work and contributions [1]:

1. We present a framework in which five frontier models, GPT-5.2, Gemini 3 Flash, Claude Opus 4.5, Grok 4, and DeepSeek-V3.2, play the same CAH games that humans have played.

2. We assess human-LLM humor alignment and find that all LLMs evaluated only achieve modest accuracy.

3. We study the agreement between LLMs and their self-consistency across repeated runs. We demonstrate that LLMs agree with each other substantially more than they agree with humans, suggesting the emergence of stable but human-misaligned humor profiles.

4. We provide an analysis of LLM humor selection, identifying significant position bias and specific content preferences across all five LLMs, and show that these two factors jointly explain a substantial share of inter-model agreement.

2 Related Work

LLMs and Humour Understanding Humor poses a fundamental challenge for NLP, requiring models to integrate world knowledge, pragmatics, and cultural context in order to read – or write – between lines. Early computational studies on humor understanding tried to recognize humor by classifying texts to humorous and non-humorous ones (Mihalcea and Strapparava, 2005). More recently, the focus has shifted from detecting humor towards rating humor, e.g., asking which of several options is the funniest. For instance, Hessel et al. (2023) introduced benchmarks derived from the New Yorker Caption Contest and asked LLMs to rank caption funniness as well as to match captions to cartoons. The results show LLMs falling substantially short of human-level judgment in this humor ranking task. This gap was further underscored by Zhang et al. (2024): they assembled 250 million crowdsourced human ratings of the New Yorker's cartoon captions and observed systematic divergences between language model and human rankings of funniness of the captions. Directly related to our setting, Ofer and Shahaf (2022) introduced a dataset of 300,000 online *Cards Against Humanity* games and trained traditional machine learning models to predict which cards humans select as the winning jokes. Their models primarily focused on the punchline card only and achieved around 20% accuracy in card choice prediction.

LLM Alignment LLMs are aligned with human preferences via RLHF (Ouyang et al., 2022), yet this alignment can fail on subjective tasks (Ying et al., 2025) and skews toward specific demographic groups in ways that resist correction even under explicit steering (Santurkar et al., 2023). While Argyle et al. (2023) demonstrate that LLMs can approximate subpopulation response distributions under appropriate conditioning, humor, as a maximally subjective, culturally loaded domain that alignment procedures may not target directly, remains an untested frontier for such demographic mapping.

3 Methodology

We formalize humor alignment as a discrete preference selection task. Let $\mathcal{M} = \{m_1, \ldots, m_k\}$ denote a set of k LLMs and $\mathcal{G} = \{g_1, \ldots, g_n\}$ a set of n game rounds. Each round $g_i \triangleq (b_i, W_i)$ consists of a black card prompt b_i and a set of candidate white card responses $W_i = \{w_i^1, \ldots, w_i^{|W_i|}\}$. Each model m_j acts as a humor judge, selecting the white card it deems the funniest response:

$$f_{m_j}(g_i) = w_i^* \in W_i \tag{1}$$

To assess preference stability and any position bias, each round is repeated across R replicates, each time following a random permutation π_r of the white card order, denoted $f_{m_j}^{(r)}(g_i)$ for replicates $r = 1, \ldots, R$.

3.1 Datasets

CAH Lab Gameplay Dataset This dataset[2] consists of games played on the online *Cards Against Humanity* Labs platform[3]. Participants engaged with the game voluntarily for entertainment purposes and were not recruited as annotators or crowdworkers for this study. In each round, a

[1] Code available at: https://github.com/aida-ugent/cards_against_llms

[2] Data is available upon request to CAH Lab at mail@cardsagainsthumanity.com.

[3] https://lab.cardsagainsthumanity.com

player is presented with a prompt card (black card) and ten candidate punchline cards (white cards), and is required to select the punchline they find funniest. The logs were collected between November 2023 and April 2025. More information about the original dataset in Appendix A.1.

CAH Lab Demographic Answers Dataset As part of the platform's optional survey, players can provide demographic information. The dataset contains responses from players, covering six demographic attributes: gender, sexual orientation, race, political ideology, country, and U.S. state. Each entry consists of a player identifier, a demographic category, and the corresponding self-reported answer. The dataset enables analysis of demographic variation in gameplay behavior while preserving anonymity.

Card Topic Labeling To better interpret players' and LLMs' answer preferences, we annotate all white cards with 1 to 4 topics among 15 possibilities (e.g. "Anatomy, bodily fluids, gross-out physical humor", "Sexual content: innuendo, explicit acts, relationships"). The list of possible topics was empirically derived from the distributions of white cards. Cards were annotated using an "LLM-as-a-judge" scheme using Mixtral 8x7B. More details on the topic selection strategy and prompting scheme, and the full list of topics, are presented in both Appendix A.2 and A.3

3.2 Round Selection

Each record in the CAH Gameplay dataset captures an online game round where the player picks a white card they think is the funniest given the black card prompt. This includes the black card prompt, the full set of candidate white card responses presented to the player in the round, round completion time, and the identity of the winning card and the identity of the player. To ensure data quality and meaningful human deliberation, we applied three filtering criteria: rounds completed in fewer than 10 seconds were excluded (as likely reflecting inattentive or reflexive judgments), as were rounds exceeding 120 seconds (which may involve distraction or disconnection rather than active play) and rounds explicitly marked as skipped.

To manage experimental cost while maintaining statistical coverage, we randomly sampled approximately 5,000 rounds from the filtered corpus (4.6% of the filtered rounds), obtaining 4,947 rounds. We preserved the original distribution of black card types and content categories of the full dataset. More details about this can be found in Appendix A.4.

3.3 Experimental Setup

We leverage the CAH Lab Gameplay dataset as our human baseline, enabling direct comparison between model selections and the cards that human players actually found funniest under naturalistic conditions. Each sampled round was presented to 5 LLMs: GPT-5.2, Gemini 3 Flash, Claude Opus 4.5, Grok 4, and DeepSeek-V3.2 as a structured humor selection task. The models were given the black card prompt and a numbered list of ten candidate white card responses, and were instructed to select the single funniest card by responding with its number followed by the exact card text. For rounds requiring two-card combinations, models were additionally instructed to select the funniest card for one blank slot, with `target_slot` indicating which blank was being filled. More details on the prompting are given in Appendix B.1. We also note that the real life CAH game is played where each round, a "Card Czar" reads a black question card, and others submit their funniest white card. The Czar picks the best answer to win a point. Since our dataset doesn't allow for such setup and our purpose is to study humor alignment, we focus on comparing LLM choices to the human players choices.

Each round was administered across two replicates ($R = 2$), with white cards shuffled into a different random order in each replicate to mitigate position bias in model selections. Responses that could not be resolved to a valid card were recorded as abstentions. Rounds in which any model abstained were flagged as invalid and excluded from agreement analyses, ensuring that all metrics are computed over a consistent set of complete observations across all five models. More details about the models abstentions can be found in Appendix B.2.

3.4 Humor Benchmarking

We study humor alignment along three complementary axes: human–LLM alignment, LLM–LLM alignment, and decision-level analysis. In the latter, we look at the LLM behavior across position bias and topic choice.

Human–LLM Alignment We measure the degree to which the humor judgment of model m_j

matches human preference using the accuracy rate:

$$\text{ACC}(m_j) = \frac{1}{2}\sum_{r=1}^{2}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left[f_{m_j}^{(r)}(g_i) = w_i^{\dagger}\right], \tag{2}$$

where $w_i^{\dagger}$ denotes the white card picked by the human.

Heterogeneity analysis If different socio-demographic groups have different senses of humor, a low average accuracy rate could mask a situation where a model's sense of humor is strongly aligned with some groups' while strongly misaligned with others'. To address this issue and better understand model alignment, we report accuracy rates for different socio-demographic groups, aggregated at the player level across rounds and replicates using the CAH Lab Demographics Answers Dataset. For this analysis, confidence intervals are bootstrapped at the player level.

LLM Agreement We measure LLM behavior along two dimensions. First, internal consistency quantifies how reliably an LLM reproduces its own judgments across replicates with randomized white card order.

Second, inter-model pairwise agreement measures the proportion of rounds in which two distinct models select the same response across replicates. We define inter-model pairwise agreement between replicates r and r' as:

$$\text{AGR}^{r,r'}(m_j, m_l) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left[f_{m_j}^{(r)}(g_i) = f_{m_l}^{(r')}(g_i)\right]. \tag{3}$$

We obtain a single measure of inter-model pairwise agreement by averaging $\text{AGR}^{1,2}$ and $\text{AGR}^{2,1}$.

Together, the two metrics reveal whether models are self-consistent and whether they agree with each other.

Explaining LLM Humor Behavior We investigate LLM humor selection along three dimensions.

Position bias We test whether each model's pick distribution across the ten slate positions deviates from uniform using a chi-square goodness-of-fit test:

$$\chi^2(m_j) = \sum_{p=1}^{|W_i|}\frac{(O_p - E_p)^2}{E_p} \tag{4}$$

where O_p is the number of times model m_j picks position p across all rounds, and $E_p = n/|W_i|$ is the expected count under uniformity, with n the total number of picks by model m_j.

Topic bias While humor in our setup is intrinsically linked to the match between a white and a black card, LLMs' selections might simply be driven by the topics of the black cards, regardless of context. We conduct an analysis of the distribution of topics present in the different LLMs' answers. To do so, we display a heatmap of the share of each of the 15 white card topics among LLMs' answers. We also display these shares in human picks and among all possible hands of white cards models could pick from for comparison.

Joint position and topic bias analysis Finally, we seek to quantify whether the combination of position bias and card topics is sufficient to accurately predict the different LLMs' answers, or if their sense of humor involves more complex mechanisms. To do so, we fit for each model j a conditional logit model, modeling the probability $p_j(k,i)$ of model j picking the card at position k in round i as

$$p_j(k,i) = \frac{\exp(x_{ik}^T\beta_j)}{\sum_{k=1}^{|W_i|}\exp(x_{ik}^T\beta_j)}$$

where x_{ik} describes the k-th card at round i in terms of topic flags and (one hot encoded) position, and β_j encodes model j's valuation of topics and card position. The models are fitted on 80% of rounds. We report the round-level accuracy of the learned models on a test set, composed of the remaining 20% of rounds.

4 Experiments

We evaluated the five LLMs on the humor preference selection task. We report results across 4,947 unique rounds with two replicates each, i.e., 9,894 records. Of these, 282 were excluded because at least one model abstained or returned an unparseable response in that round; more details are given in Appendix B.2. This left 9,612 records where all five models produced a valid card selection. Unless otherwise noted, all analyses use only valid rounds. For brevity, we refer to each model by its family name throughout the following sections (e.g., GPT, Gemini, Claude, DeepSeek, Grok) rather than the exact model version used.

4.1 Human–LLM Alignment

First, we measure human alignment as the proportion of rounds in which a model selects the same white card as the human-designated winner out of ten candidate cards. While the CAH Lab Gameplay dataset does not allow us to measure

inter-annotator agreement, the performance of several baselines can help contextualize these findings. Random card choice would achieve an accuracy of 10%, whereas picking cards based on popularity would achieve 19.11% accuracy, and an ensemble of boosted trees learned from human player choices 19.77%. More details on the construction of these baselines is provided in Appendix D. The small gap between the accuracies of the popularity baseline and of the boosted tree ensemble is coherent with the findings of Ofer and Shahaf (2022).

All five models exceed the random baseline of 10%, as shown in Figure 2. Claude achieves the highest overall alignment, followed by Grok and Gemini, while DeepSeek and GPT trail behind. However, absolute alignment rates remain low, ranging between 13% and 18%. This suggests that although LLMs capture some aspect of human humor preference, the task remains largely unsolved. Performance is consistent across replicates, indicating that the low alignment reflects a genuine limitation rather than model instability.

Heterogeneity analysis We turn to the investigation of differences in model alignment across sociodemographic groups covered by the CAH Lab Demographic Answers Dataset. Note that the rounds for which LLM picks were obtained cover 824 players, as 50% of rounds have missing player IDs and players can play multiple rounds. Figure 4 displays accuracy rates by demographic subgroups at the player level on this population. While we do not have enough statistical power to detect fine-grained demographic differences, the results do not suggest the existence of substantial heterogeneity. In other words, we find no evidence that the low-to-moderate human–LLM alignment found on average is driven by differences in alignment across sociodemographic lines.

4.2 LLM Agreement

Next, we examine whether models have their own senses of humor, or even converge on a shared 'LLM sense of humor', through two complementary lenses: inter-model agreement, and response consistency across replicates. Figure 3 shows pairwise agreements between LLMs (rows and columns), while the diagonal indicates intra-model consistency, computed as agreement between replicates of the same model.

Intra-model consistency (cells in the diagonal) is much higher than LLM-human alignment across all models. Grok shows the strongest self-consistency (63.3%), followed by Gemini (59.9%) and Claude (59.8%), while GPT is the least consistent (49.5%).

Inter-model agreements (other cells than the diagonal) range from 21.4% to 44.9%. Thus, remarkably, they also substantially exceed the human-LLM alignment rates of 13–18% reported in Figure 2.

Model	χ^2	*p*-value	Dominant Pos.
GPT	356	<0.001	—
Gemini	282	<0.001	—
Claude	678	<0.001	Early positions
DeepSeek	1851	<0.001	Position 3
Grok	658	<0.001	Position 10

Table 1: Chi-square test of uniformity for position bias per model (df=9). Higher χ^2 indicates stronger deviation from a uniform pick distribution across the 10 slate positions.

4.3 Explaining LLM Humour Behavior

Next we analyze two types of possible biases in LLMs in this task: position bias and topic bias.

4.3.1 Position bias

All five models exhibit significant deviation from a uniform pick distribution as displayed in Table 1. This aligns with finding from previous literature (Pezeshkpour and Hruschka, 2024). DeepSeek shows the strongest position bias, driven by a pronounced concentration of picks at Position 3. Claude and Grok display similarly strong bias, with Grok favoring the last position and Claude showing moderate spread with a mild early-position preference. GPT and Gemini exhibit the weakest bias, though still highly significant, with picks distributed more evenly across the slate. We show this in Appendix C.

4.3.2 Topic bias

We turn to the description of the topics present in the white cards chosen by LLMs, displayed in Figure 5. In coherence with pairwise agreement patterns, we find commonalities between the topics involved in Gemini, DeepSeek, Claude and Grok's responses. In particular, these models' card picks often rely on bodily humor (31% to 40% of answers, compared to 21% for humans) and sexual themes (29% to 38%, against 24% for humans). On the other hand, GPT makes less use of cards with these topics (24% and 15% respectively), and stands out in terms of the use of "miscellaneous"-themed cards (27% of answers, against 16% to

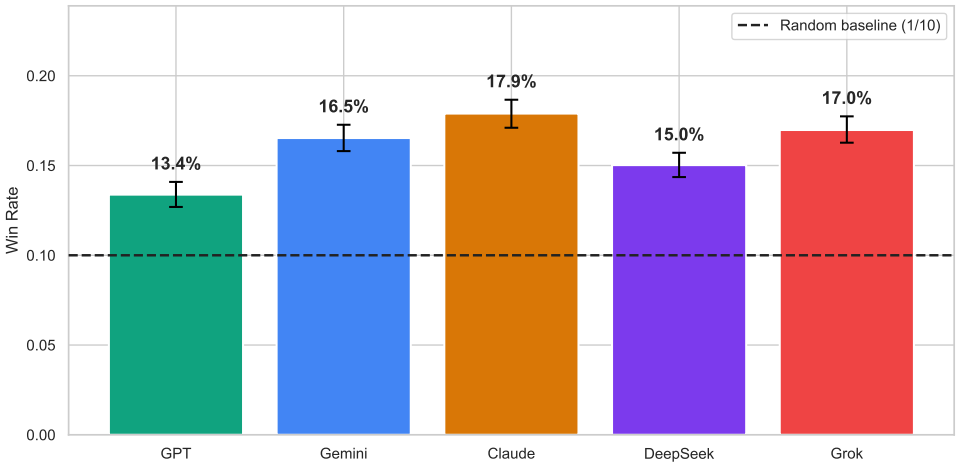

Figure 2: Human-LLM Alignment for all 5 models, with bootstrapped 95% confidence intervals. The dashed line indicates the random baseline (1/10), reflecting the expected win rate of a model selecting cards uniformly at random from a slate of 10.

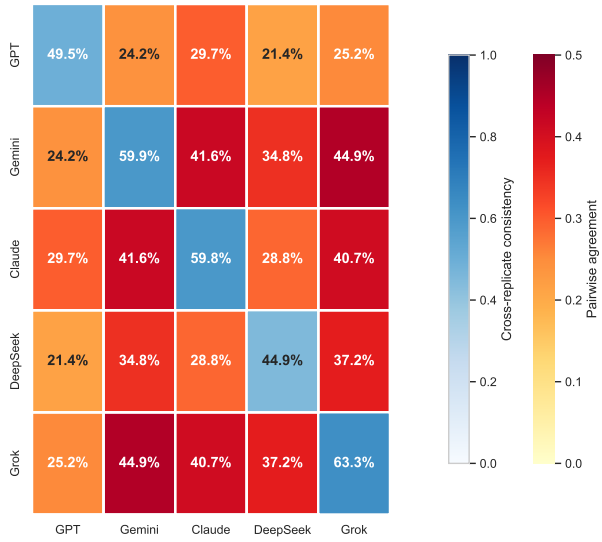

Figure 3: Pairwise agreement rate between models, measured as the proportion of rounds in which two models select the same white card. Off-diagonal cells show inter-model agreement (yellow-red scale); diagonal cells show intra-model consistency across the two replicates (blue scale).

18% for other models and humans). It is also worth noting that all models' responses involve fewer answers related to "politics/society" and "identity/demographic" topics (6-8% and 3-5% of answers respectively) than humans (14% and 10%).

4.3.3 Joint position and topic bias analysis

Finally, we assess to what extent the combination of position and topic biases predicts LLMs' responses. We fit conditional logistic models with card position and topic as covariates to predict LLM behavior based on the list of white answer cards only, ignoring the black card (Ofer and Shahaf, 2022). Table 2 displays the share of rounds in the test set for which logistic models correctly predict LLMs' answers. Note that an upper bound for such a prediction quality metric is given by the LLMs' own self-consistency, whereas if both topics and position biases were irrelevant, the metric would be 0.1. The surrogate models for Grok, DeepSeek and Gemini correctly characterize chosen white cards for more than a third of rounds (0.361, 0.351 and 0.339). The surrogate conditional logistic models for GPT and Claude achieve markedly lower round-level accuracies, at 0.171 and 0.244. Altogether, these results suggest that a non-trivial share of model responses can be explained by position bias and white card content preference, but also suggests the existence of more complex mechanisms at play in LLM humor.

Model	Round-level accuracy
GPT	0.171
Gemini	0.339
Claude	0.244
DeepSeek	0.351
Grok	0.361

Table 2: Share of rounds for which surrogate logistic models correctly predict LLMs' answers.

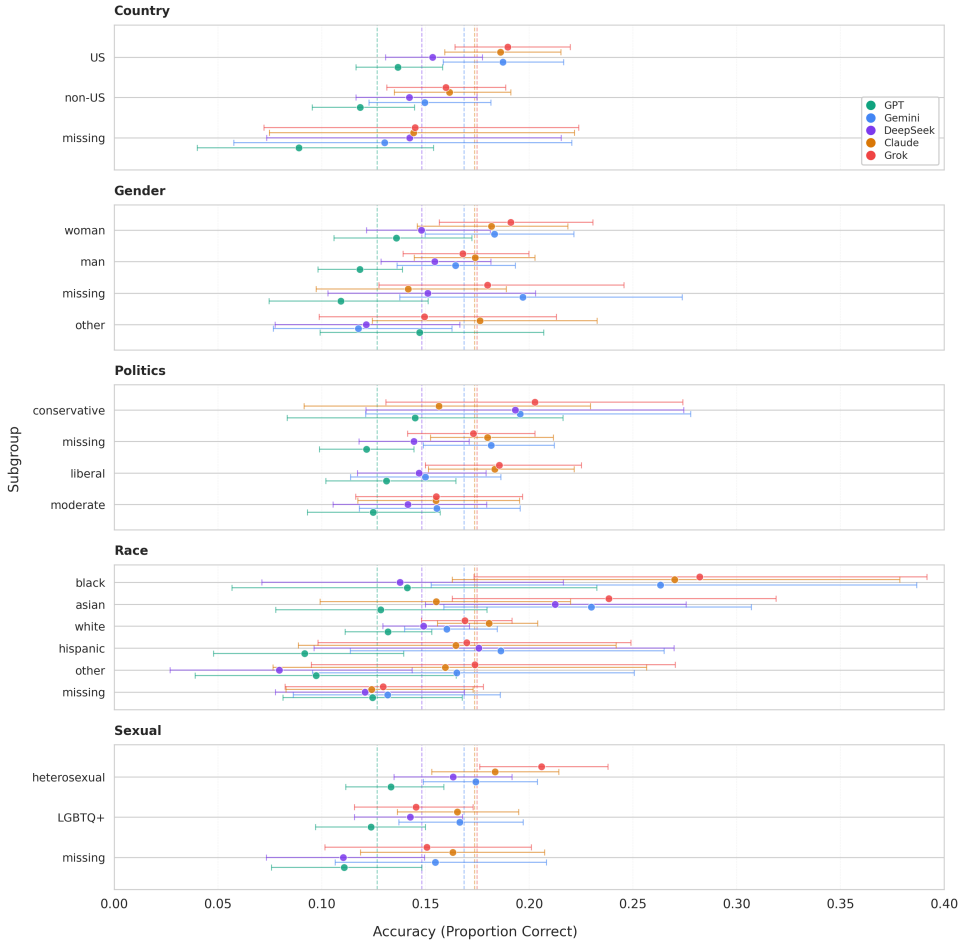

Figure 4: Accuracy rates by demographic subgroup, aggregated at the player level across replicates and rounds. The displayed 95% confidence intervals are bootstrapped at the player level. Vertical bars indicate player-level mean accuracy of models on the population of the 824 players with non-missing IDs.

5 Discussion

Altogether, our analysis of the *Cards Against Humanity* gameplay of LLMs points towards the existence of i) an accuracy ceiling in current human-LLM humor alignment, and ii) the convergence of frontier LLMs towards a somewhat shared, but human-misaligned, notion of humor.

Our findings suggest LLMs **capture some meaningful aspects of human humor preferences**: accuracy rates of LLMs significantly outperform a random baseline in terms of human-LLM agreement. Nevertheless, these accuracy rates are only moderate (13-18%), whereas simple baselines, as presented in Section 4.1, suggest accuracies around at least 19%-20% can be achieved. This average finding does not seem to be driven by underlying considerable differences in alignment to different demographic subgroups.

Moreover, we also find that LLMs **agree with each other far more than they agree with humans**. This may indicate that LLMs have converged to some extent towards a shared but human-misaligned notion of humor where bodily and sexual topics are emphasized more, and demographic topics less, than by humans.

We hypothesize that LLMs rely partly on shallow heuristics rather than deeper pragmatic and contextual reasoning that drives human humor selection. Indeed, we find models to exhibit significant position bias and commonalities in topic choices in their answers, as shown by surrogate model accuracies up to 36%.

The current accuracy ceiling in human-LLM humor alignment also likely reflects both the inherent subjectivity of humor and the fact that CAH provides a space for, and even rewards, transgressive incongruity. This mode of humor **may conflict with the safety-oriented fine-tuning of current LLMs**: shared patterns in instruction tuning or RLHF that reward certain types of responses over

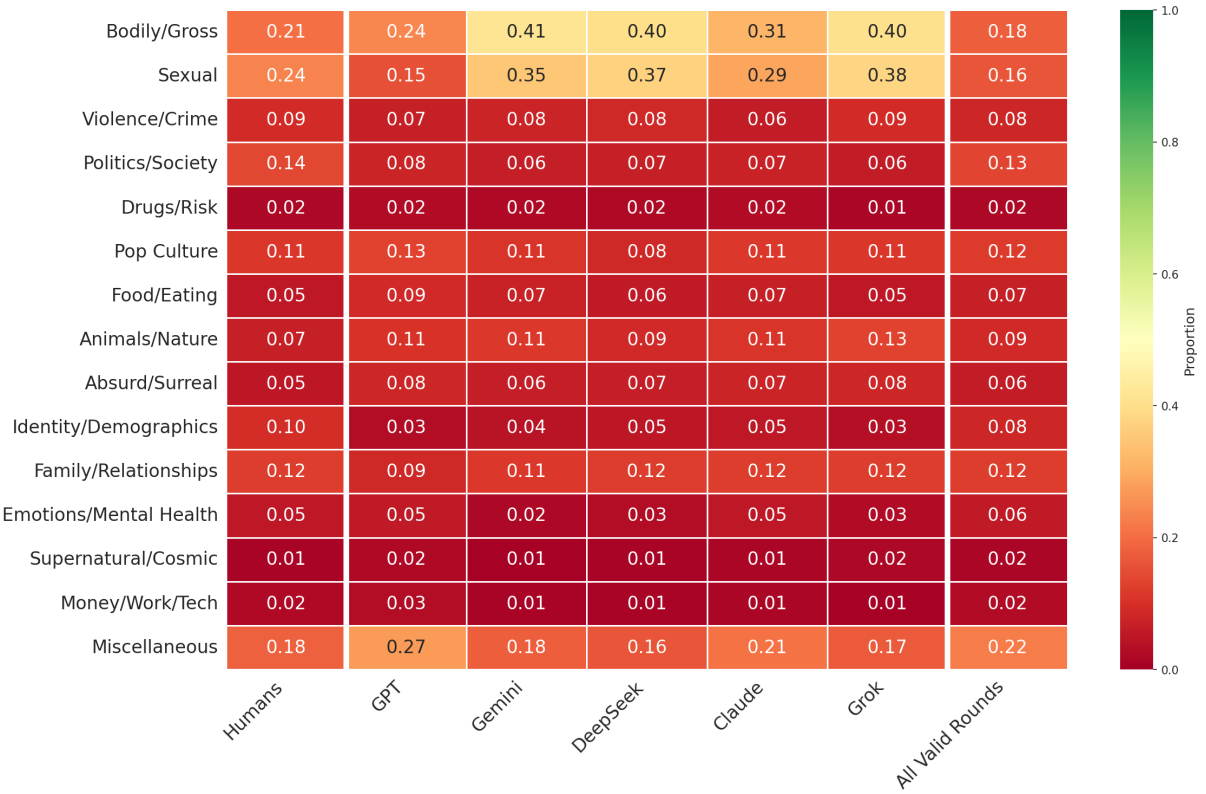

Figure 5: Shares of human (column 1) and LLM (columns 2-5) white card picks involving different topics (a card can have several topic tags). Column 6 provides the shares of topics among all available cards for reference.

the transgressive and absurdist style that makes CAH funny to humans. Consistent with this hypothesis, the analysis of topic choices reveals LLMs to choose white cards about "politics/society" and "identity/demographics" topics less frequently than humans, or compared to the distribution of such cards among possible answers.

6 Conclusion and Future Work

We presented a large-scale study of humor alignment between frontier language models and humans, using *Cards Against Humanity* as a naturalistic preference selection testbed. Across 9,894 rounds and five models, we find LLMs' alignment with human humor judgments to be modest. Strikingly, models agree with each other substantially more than they agree with humans. Further analysis reveals that this divergence is partly explained by systematic position biases and content preferences, raising important questions about the extent to which LLM behavior on tasks involving humor reflects genuine understanding or structural artifacts of inference.

Several directions emerge naturally from this study. A larger experimental scale — more rounds, replicates, and a more culturally diverse set of models — would yield more robust estimates and enable cross-cultural comparison. Repeating the experiment with milder humor registers, such as CAH: Family Edition, would help disentangle a general human–LLM humor gap from a specific mismatch with transgressive incongruity. Additionally, extending the framework to other humor formats such as memes or satirical writing would test the generalizability of our findings. Finally, our observation that models agree with each other more than with humans raises the question of whether specific human subgroups, defined by humor style rather than demographics, show substantially higher LLM alignment.

7 Limitations

- **Experimental scale.** While our study covers 9,894 rounds across two replicates, a larger number of rounds and replicates would yield more stable estimates and stronger statistical conclusions. Expanding the experimental scale was constrained by the API costs associated with querying five frontier models across multiple runs.

- **Single-player ground truth.** The CAH Lab dataset provides the choice of a single player for each round, which we use as human reference. However, the data does not allow us to estimate inter-rater agreement, which would help contextualize the modest LLM accuracy rates, and distinguish model failure from inherent task subjectivity.

- **Temperature Setting.** All models were queried at a fixed temperature of 0.8, allowing for some response variability while maintaining comparability across models. Varying this parameter could yield different self-consistency profiles and is left for future investigation.

- **LLM-as-judge for topic labeling.** Our content analysis relies on LLM-generated labels to categorize white card themes, introducing potential noise and systematic bias. We manually verified label coherence on a subset of cards, finding the labels to be generally consistent and meaningful. However, a more systematic validation methodology such as inter-annotator agreement with human raters would strengthen the reliability of this analysis and is left as future work.

- **Player population.** The dataset reflects the preferences of a specific population of CAH players, who are predominantly Western and self-selected. Broader coverage of player demographics and cultural backgrounds would be needed to generalize our human alignment findings.

- **Western model bias.** Four of the five models — GPT, Gemini, Claude, and Grok — are developed primarily in a Western context, and their humor profiles may reflect cultural biases embedded in their training data and alignment procedures. DeepSeek is the only exception, though a single non-Western model is insufficient to draw cross-cultural conclusions.

Acknowledgments

We would like to thank the reviewers for their comments. We would also like to thank our colleagues Maarten Buyl for helping with early-stage conception, MaryBeth Defrance, Edith Heiter and Jefrey Lijffijt for their valuable comments and feedback. This work was carried out while Hannu Toivonen was an International Francqui Professor at VUB, KU Leuven, UAntwerp, UGent and UCLouvain and a Senior Fellow at BrIAS. The research leading to these results was funded/co-funded by the European Union (ERC, VIGILIA, 101142229), the Special Research Fund (BOF) of Ghent University (BOF20/IBF/117), the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme, and the FWO (project no. G073924N). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. For the purpose of Open Access the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

Anthropic. 2025. System card: Claude Opus 4.5. Technical report, Anthropic.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Salvatore Attardo. 1997. The semantic foundations of cognitive theories of humor. *Humor*, 10(4):395–420.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6848–6863.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Tulsee Doshi. 2025. Gemini 3 flash: Frontier intelligence built for speed. Google Blog.

Jennifer Hay. 2001. The pragmatics of humor support. *Humor: International Journal of Humor Research*, 14(1).

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

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

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 531–538.

Dan Ofer and Dafna Shahaf. 2022. Cards against ai: Predicting humor in a fill-in-the-blank party game. *arXiv preprint arXiv:2210.13016*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

xAI. 2025. Grok 4 model card. Technical report, xAI.

Shuangshuang Ying, Yunwen Li, Xingwei Qu, Xin Li, Sheng Jin, Minghao Liu, Zhoufutu Wen, Xeron Du, Tianyu Zheng, Yichi Zhang, and 1 others. 2025. Beyond correctness: Evaluating subjective writing preferences across cultures. *arXiv preprint arXiv:2510.14616*.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan L Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, and 1 others. 2024. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *Advances in Neural Information Processing Systems*, 37:125264–125286.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T Rogers, Lalit K Jain, Robert D Nowak, Bob Mankoff, and Jifan Zhang. 2025. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in llms. *arXiv preprint arXiv:2502.20356*.

A Datasets Details

A.1 CAH Gameplay dataset

The raw dataset had 148,497 past games (rounds). There are 501 unique black prompt cards and 2074 white punchline cards, resulting in 1,484,970 possible unique jokes (where a joke is the result of filling in the blank of the prompt card with a punchline).

A.2 Topic list selection

To obtain the list of 15 topics used for the analysis, we prompted a version of GPT-5 accessed through Microsoft Copilot to provide a taxonomy of topics based on the following prompt: *"The following short pieces of text are "white cards" in the game "Card against humanity". Help me find a taxonomy of topics they cover, so I can label them downstream."*, followed by white cards. This prompt was run twice, with two different samples of 100 white cards, yielding two different taxonomies which were then harmonized.

Table 3 provides the resulting list of topics (with their short-hand labels, as used in Figure 5).

Table 3: Topic Labels and Definitions

Label	Definition
Bodily/Gross	Anatomy, bodily fluids, gross-out physical humor
Sexual	Sexual content: innuendo, explicit acts, relationships
Violence/Crime	Physical harm, mortality, criminal acts, threats
Politics/Society	Government, activism, social norms, cultural commentary
Drugs/Risk	Substance use, addiction, reckless actions
Pop Culture	Celebrities, movies, memes, brands, viral trends
Food/Eating	Meals, ingredients, dining, consumption
Animals/Nature	Wildlife, pets, ecosystems, biological refs
Absurd/Surreal	Illogical juxtapositions, nonsense, anti-humor
Identity/Demographics	Race, gender, age, disability, sexuality, nationality
Family/Relationships	Parenting, friendships, domestic life, mundane interactions
Emotions/Mental Health	Anxiety, joy, depression, coping, psychological framing
Supernatural/Cosmic	Ghosts, aliens, magic, existential cosmic themes
Money/Work/Tech	Jobs, finance, digital life, institutional critique
Miscellaneous	Concrete items/concepts not captured above

A.3 Topic annotation prompt

White cards were annotated with topics using the following prompt, passed to Mixtral 8x7B.

```
”’ You are an expert annotator labeling "white cards" from Cards Against Humanity for humor research.
### TASK Analyze the input card and return a JSON object matching the schema below.

### OUTPUT SCHEMA (STRICT)
{
"topics": ["slug1", "slug2"], // 1-4 items from TOPICS list
}

### TOPICS (use slugs exactly; select 1-4)
1. bodily_functions_gross_out // Anatomy, bodily fluids, gross-out physical humor 2. sexual_themes // Sexual content: innuendo, explicit acts, relationships
3. violence_crime_death_threat // Physical harm, mortality, criminal acts, threats
4. politics_ideology_society_culture // Government, activism, social norms, cultural commentary
5. drugs_alcohol_risky_behavior // Substance use, addiction, reckless actions
6. pop_culture_media_consumerism // Celebrities, movies, memes, brands, viral trends
7. food_eating_consumables // Meals, ingredients, dining, consumption
8. animals_nature_creatures // Wildlife, pets, ecosystems, biological refs
9. absurdism_surreal_nonsensical // Illogical juxtapositions, nonsense, anti-humor
10. identity_demographics_traits // Race, gender, age, disability, sexuality, nationality
11. family_relationships_everyday // Parenting, friendships, domestic life, mundane interactions
12. emotional_states_mental_health // Anxiety, joy, depression, coping, psychological framing
13. supernatural_cosmic_paranormal // Ghosts, aliens, magic, existential cosmic themes
14. money_work_technology_modern // Jobs, finance, digital life, institutional critique
15. random_objects_miscellaneous // Concrete items/concepts not captured above

### RULES
1. Use slugs exactly as written (underscores, lowercase, no spaces).
2. Output ONLY valid JSON. No markdown, no preamble, no extra text.
”’
USER_PROMPT_TEMPLATE = ”’
### INPUT CARD
"{card_text}"

### OUTPUT ”’
```

All 2074 unique white cards in the gameplay dataset were successfully annotated with 1 to 3

topics. Coherence of the annotations was validated by the authors. We also note the topic annotation was done on our final dataset that we've gotten after the round selection.

A.4 Round Selection Details

After applying the filtering criteria mentioned in Section 3.2, we obtain a dataset of 107,562 unique rounds. This dataset contains 493 unique black cards and 2,072 unique white cards, resulting in 1,075,620 possible unique jokes. After randomly sampling rounds from this filtered dataset, we get 4,947 rounds. Among these rounds, 2,437 can be mapped to a player with valid identifier; there are 824 such players in the studied dataset.

B Models

We evaluate five conversational large language models (LLMs) in our experiments:

- `gpt-5.2` (OpenAI) (Singh et al., 2025).
- `gemini-3-flash-preview` (Google) (Doshi, 2025).
- `claude-opus-4-5` (Anthropic) (Anthropic, 2025).
- `grok-4` (xAI) (xAI, 2025).
- `DeepSeek-V3.2` (DeepSeek) (Liu et al., 2025).

This selection reflects the current landscape of frontier LLMs, covering both open-weight and proprietary systems trained on diverse corpora with differing alignment objectives. All models were queried via their respective APIs with a fixed temperature of 0.8, balancing response diversity with reproducibility, and a maximum output length of 128 tokens, sufficient for the card selection format required. To isolate humor preference from instruction-following variability, all models received an identical prompt template, described in Appendix B.1. All outputs were collected between the 17th of January and the the 10th of March 2026.

B.1 Model Prompting

All models received the same prompt template, instantiated from the black card and white card slate. No examples or chain-of-thought instructions were provided, so as to elicit each model's intrinsic humor judgment. Models were instructed to respond with only the number and exact text of their chosen card to facilitate automated parsing.

B.1.1 Single-Blank Prompt

For black cards with a single blank, models were asked to select the funniest card from their hand:

```
You are playing Cards Against Humanity. It's
just a game so it's okay to use profanity and
to be politically incorrect.
Choose the single funniest white card.

Black card:
{black_card}

Your hand:
1. {card_1}
2. {card_2}
...

Respond ONLY with:
<number>. <exact card text>
```

where `{black_card}` is the text of the black card with the blank represented as an underscore, and `{card_1}...{card_N}` are the white cards in the slate ($N = 10$ in all rounds).

B.1.2 Multi-Blank Prompt

For black cards with two blanks, a target slot was designated (BLANK #1 or BLANK #2) and models were asked to fill only that slot. This keeps the response format consistent across all rounds and allows positional effects to be analyzed independently. It does not capture combinatorial humor arising from the interaction between two cards.

```
You are playing Cards Against Humanity. It's
just a game so it's okay to use profanity and
to be politically incorrect.
Choose the funniest white card to fill BLANK
#{target_slot}.

Black card:
{black_card}

Your hand:
1. {card_1}
2. {card_2}
...

Respond ONLY with:
<number>. <exact card text>
```

where `{black_card}` is the text of the black card with blanks represented as underscores, `{target_slot}` is either 1 or 2 indicating which blank to fill, and `{card_1}...{card_N}` are the white cards in the slate ($N = 10$ in all rounds).

B.2 Models Abstentions

Not all models engaged with every round. Gemini exhibited a notably higher abstention rate, produc-

ing null or failed picks in 280 rounds (2.8% of its total records), likely reflecting its content moderation filters triggering on CAH's characteristically offensive prompts. All other models showed near-zero failure rates: GPT, Claude, and DeepSeek recorded no null picks whatsoever. Grok produced only 2 failed picks. One of the fails was caused by a connection error rather than a content refusal. These abstentions were excluded from downstream analyses to ensure that win rate comparisons reflect genuine card selection behavior rather than differential willingness to engage with the game's content.

C Explaining LLM Humor Behavior Details

We showcase the distribution of model picks across the slate positions in Figure 6. All five models deviate from the uniform baseline, confirming the presence of positional bias across the board.

D Baselines predicting human behavior

In the absence of proper inter-annotator agreement measures in the CAH Lab Gameplay dataset, we construct two simple baselines (a card popularity baseline, and boosted tree predictions of human choices) to better contextualize the accuracy rates achieved by different LLMs, building on previous work exploration of human humor prediction in CAH (Ofer and Shahaf, 2022). Both baselines are trained on a training set of 84,110 CAH rounds, obtained by removing from the CAH Gameplay dataset all LLM-annotated rounds, and data from any players who participated in these rounds (when player IDs were known).

The first baseline predicts white card choices based on their winrates in a training set (disregarding associated black cards). Despite its simplicity, this strategy was noted to be a strong baseline (Ofer and Shahaf, 2022), outperforming more elaborate modeling attempts in predicting human choices when card position information is not available, as is the case in the dataset.

The second baseline trains an XGBoost classifier on human choices to predict winning white cards in a given round. The learning task is formulated as a binary classification objective at the card-option level. Each row corresponds to a white card in a given round, with a binary label $y \in \{0, 1\}$ indicating if it was picked that round. Features describing the card-round combination are: i) a 384-dimensional embedding of the white card text; ii) a 384-dimensional embedding of the black card text; iii) the 15-topic annotations for the white card, leading to a total of 783 features. White and black card embeddings are obtained using a Sentence-Bert embedding of the card texts using *all-MiniLM-L6-v2*. To compute test accuracies, we simply compare the card with highest predicted pick probabilities in the round to the actual human choice. Hyperparameters are tuned via *scikit-learn*'s *RandomizedSearchCV* (50 iterations, 3-fold stratified CV) optimizing ROC-AUC; considered options are listed in Table 4. The final model is refit on the full training set.

Table 4: Hyperparameter Search Space for XGBoost Baseline

Hyperparameter	**Search Values**
n_estimators	100,150,200,250,300
max_depth	4,5,6,7,8
learning_rate	0.05,0.08,0.1,0.15,0.2
min_child_weight	3,5,7,10
scale_pos_weight	7,8,9,10,11
subsample	0.7,0.8,0.9,1.0
colsample_bytree	0.7,0.8,0.9,1.0

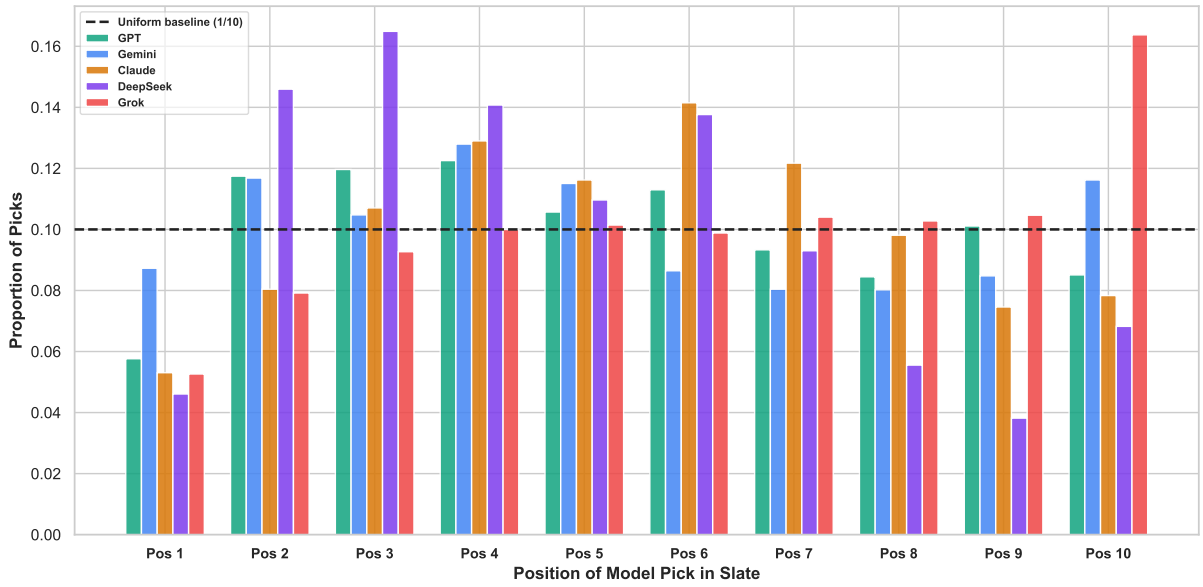

Figure 6: Distribution of model picks across slate positions (N=10). Each bar represents the proportion of rounds in which a model selected the card at a given position, aggregated across all rounds and replicates. Under the null hypothesis of no positional bias, picks would be uniformly distributed at 0.10 (dashed line).

The Roast of GPT4o: Experiments in Generating, Detecting and Evaluating Celebrity Roast Comedy

Jens Lemmens, Jérémy Genette, Walter Daelemans
University of Antwerp
Prinsstraat 13, 2000 Antwerp, Belgium
firstname.lastname@uantwerpen.be

Tony Veale
University College Dublin
Belfield, Dublin 4, Ireland
tony.veale@ucd.ie

Abstract

We present exploratory experiments in the comedic roasting capabilities of GPT4o. Specifically, @ComedyCentral roasts were scraped to design a survey in which participants blindly evaluated snippets of human and AI roasts, and had to predict the author (AI/human) in a second round of reviewing. The results show that there is no significant difference in how the barbs in human- and AI-generated roasts are rated. Further, a qualitative analysis showed that although the model utilizes specific recurrent phrases to imitate the style of human comedians, both generative LLM detectors and humans performed suboptimally in predicting the true author of the roasts.

1 Introduction and related work

While the most recent language models have shown complex logical reasoning capabilities, mixed results have been reported on the problem of humour generation, which requires world knowledge, semantic reasoning, and linguistic insight. On the one hand, these mixed results may reflect the variety of humour genres that are explored and the prompting strategies utilized in earlier work. Specifically, previous studies have chiefly focused on stand-alone jokes, e.g., Toplyn and Amir (2025); Horvitz et al. (2024); He et al. (2024); Bogireddy et al. (2023); Zeng et al. (2024); Zhang et al. (2024); Mittal et al. (2022). Such out-of-context jokes, however, account just for an estimated 17% of our daily experience of humour (Martin and Ford, 2018). In contrast, performance humour is relatively understudied in Natural Language Generation (NLG) research and has mainly focused on stand-up comedy or sitcom humour, e.g., Li et al. (2023); Mirowski et al. (2024). In this paper, we thus aim to provide new insights into the humour generation problem by presenting initial experiments with GPT4o (OpenAI, 2024) on the unexplored task of comedy roast generation. A *roast* is a form of insult comedy in which guests hurl affectionate barbs at a guest of honour. Specifically, the unit of analysis in this paper is a specific fragment (or "burn") that is uttered by a certain speaker at such a roast event.

Apart from a strong focus on specific genres of humour, the inconclusive findings regarding the humor capabilities of LLMs could also be attributed to a lack of a standardized evaluation paradigm (Lemmens and De Marez, 2026). Previous work has used human evaluators, automatic methods, or a combination of both. In the case of human evaluators, both A/B testing, e.g., Chen et al. (2024), and scoring on Likert scales has been used, e.g., Gorenz and Schwarz (2024), although these scales have not been used consistently in terms of values (5-point vs. 7-point scale) and in what they measure (overall score, humorousness, originality, style, etc.). An advantage of human evaluators is that they can consider a variety of fine-grained aspects of humour when evaluating content, and can offer qualitative insights into their evaluation, in contrast to automatic methods. Disadvantages, on the other hand, include greater expense and slower work-rates, as well as the subjectivity of personal preferences (Romanowski et al., 2025). In this paper, we alleviate this problem by including a large number of human evaluators and by analyzing these personal preferences using a mixed effects statistical analysis.

We begin by scraping Comedy Central roast transcripts from YouTube, and use AI to generate new roast materials. A survey is then used to allow participants to blindly rate the human- and AI-written roasts to determine the comedic abilities of GPT4o. In addition, we investigate the extent to which AIs can predict if a roast was written by another AI, which may produce insights into how to automatically evaluate the quality of AI-generated humour.

We explore the following research questions and hypotheses: (1) Can GPT4o generate roasts that are as funny as human roasts? (2) How accurately can the author (AI or human) of a roast be recognized?

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 65–71

We hypothesize that (1) due to model alignment, and the creative nature of the task, GPT4o is unable to generate roasts of the same funniness as humans. Thus, ratings for human roasts will be significantly higher than those for AI-generated ones. In addition, (2) humans and LLM detectors can both predict if a roast was created by AI or not with relatively high accuracy due to the quality difference between the human and AI generated roasts.

Contributions[1]: Firstly, we provide the materials to scrape all ComedyCentral roasts transcripts on Youtube. Secondly, the results of this study serve as a proof-of-concept that GPT4o can generate roast content that is as funny as that of human comedians, in opposition to our initial hypothesis, while adhering to a specific writing style, as explained in a qualitative analysis. Thirdly, we show that LLM detectors exhibit sub-optimal results on comedy data, although LLM detectors still perform better at this task than humans, who do not exceed random chance.

2 Methodology

A survey was designed to let 64 participants (see Appendix A) blindly rate fragments of human- and AI-generated roasts. To collect data for the survey, we used the Google API to scrape all videos of the @ComedyCentral channel on YouTube containing 'roast' in the title. Then, twenty fragments from this collection of transcriptions were manually selected, following 2 criteria: First, for purposes of comparison, the fragment must be self-contained, and understandable without additional video/audio data. Secondly, 10 distinct roast targets were represented in the subset, i.e. 2 fragments per target.

A random selection of 10 of these 20 fragments were included in the survey, while the remaining 10 were used to generate new roast fragments with GPT4o (OpenAI, 2024). The prompt (Appendix C) first explains what a roast is, using a short version of the Wikipedia definition, and names the specific celebrity to be targeted by the roast. A short Wikipedia biography of the target is then given in the prompt as background information, and a 1-sentence paraphrase of an original roast is provided as a topic. An example roast with a different target is also provided so that the model acquires a sense of the general style and structure of a roast.

Since it was the intention to ask participants to predict whether a roast was written by AI or not, post-processing was required to avoid unintentionally identifying the author with highly revealing text features. Therefore, profane words were removed from the human roasts (if grammatically possible), or replaced with a non-profane synonym, since LLMs rarely generate profanity, even when role-playing, due to model alignment (Huang et al., 2025). In addition, em-dashes were either removed or replaced by commas in the AI-generated roast, as these are a common feature of LLM outputs (Sukhareva, 2025). At this point, our mini-dataset pairs 10 human roasts with 10 AI roasts.

A survey was then created to let participants evaluate these roast fragments[2] in a within-subjects design. This survey consists of two parts: First, participants are asked to indicate whether they are familiar with the target of the roast (variable name: 'familiarity') and to then rate the roasts on a Likert scale (variable name: 'rating') from 1 (not funny) to 5 (extremely funny). To avoid biasing them in their ratings, participants are told that the purpose of the survey is to compare the humour preferences of humans to those of LLMs. This masks the true goal of the study, so as to not reveal that half of the fragments are AI-generated.

In part 2 of the survey, participants were told that half of the roasts were generated by AI, but not told which ones. They were then presented with the same content in the same order as in part 1, and were asked to indicate whether they had seen the roast before on Comedy Central (variable: 'prior encounter') and to predict whether an AI or human wrote it (variable: 'author prediction').

To evaluate the author predictions of participants (AI or human), those predictions were compared with two state-of-the-art generative AI classifiers: Binoculars (Hans, Abhimanyu and Schwarzschild, Avi and Cherepanova, Valeriia and Kazemi, Hamid and Saha, Aniruddha and Goldblum, Micah and Geiping, Jonas and Goldstein, Tom, 2024) and GPTZero (OpenAI, 2023). GPTZero is a multi-step commercial LLM detector that specializes in detecting content generated by ChatGPT, GPT4, Gemini, Claude, and LLaMa models. Binoculars, in contrast, is an open-source zero-shot LLM detection method for text data that leverages normalized cross-perplexity scores between two LLMs to predict whether a text was written by AI or not. In this experiment, we used Falcon-7B and Falcon-

[1]The code and survey data are available on Github: https://github.com/LemmensJens/roast_of_gpt4o.

[2]See Appendix B for an example of the data.

7B-instruct (Almazrouei et al., 2023), and a perplexity threshold of 0.90, as proposed in Hans, Abhimanyu and Schwarzschild, Avi and Cherepanova, Valeriia and Kazemi, Hamid and Saha, Aniruddha and Goldblum, Micah and Geiping, Jonas and Goldstein, Tom (2024).

3 Results

3.1 Statistical analysis

The analysis of the survey data is summarized in Figure 1, where it can be observed that participants predicted that the funnier roasts were human-generated. Additionally, roasts were rated as funnier when participants were more confident that they had seen the roast before. These observations are confirmed by the results of the following statistical analysis (conducted using R (R Development Core Team, 2022) and the *ordinal* package (Christensen and Christensen, 2015)).

Cumulative Link Mixed-Effects Models were used to examine whether 'author', 'familiarity', 'author prediction', and 'prior encounter' have an effect on 'funniness'. Models of increasing complexity were built step-by-step by incrementally including fixed and random effects, with the funniness rating used as the dependent variable. The inclusion of a predictor was assessed by a likelihood ratio test (Baayen, 2008).

The final model included 'author prediction' and 'prior encounter' as fixed effects, and a random intercept for each participant and each roast. The inclusion of 'author' and 'familiarity' as predictors did not significantly improve model fit and were therefore excluded. This suggests that these variables did not have a significant effect on the funniness ratings. That is, there was no statistically significant difference between the funniness ratings of the human-generated roasts and the AI-generated roasts, indicating that, in general, the participants found the AI roasts and the human roasts equally funny. Similarly, whether participants were familiar with the subject of the roast had no influence on how funny they found the roast.

However, 'author prediction' and 'prior encounter' had a significant effect on funniness ratings, as shown in Table 1, which presents the results of the fixed effects from the modelling procedure. From these results, it can be observed that there was a significant positive main effect of 'author prediction' ($\beta = 0.648$, $SE = 0.118$, $z = 5.488$, $p < 0.001$). This indicates that participants tended to find a roast more funny if they thought the roast was human-generated rather than AI-generated.

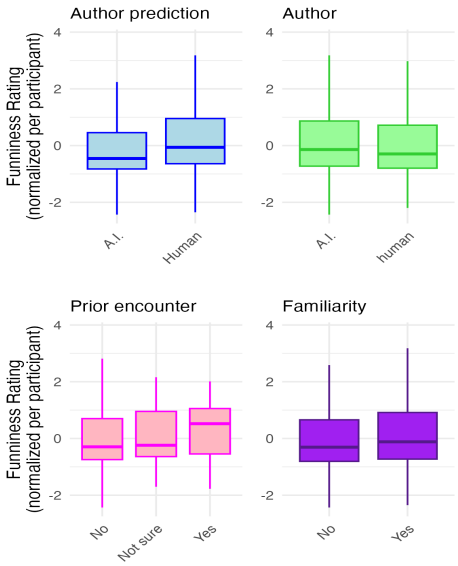

Figure 1: Overview of the funniness ratings: Likert scores normalized per participant (mean and standard deviation). Non-normalized results are included in Appendix E.

The participants were first asked to provide ratings without knowing that some roasts were AI-generated. After it was revealed that not all roasts were human-generated, participants were then asked to predict the author of each roast. As such, this suggests that a participant's perception of a roast's funniness influenced their prediction of its author. Given that there was no significant effect of the actual author on the funniness rating, however, this indicates that there was a strong mismatch between the actual authors and the predicted authors of the roasts. The low performance of the participants on the author prediction task, as further described in Table 2, supports this observation.

In addition, 'prior encounter' (whether they think they have seen the roast before) had a significant linear effect on the funniness rating ($\beta = 0.601$, $SE = 0.192$, $z = 3.136$, $p = 0.002$). This indicates that the more confident a participant's belief is in having encountered a roast before, the funnier they tend to rate it. So prior exposure not only plays a significant role in humour evaluation but also complicates automatic humour assessment, as it increases the subjectivity of such evaluations.

Variable	Estimate	Std. Error	z value	Pr(>|z|)
Author prediction	0.648	0.118	5.488	< 0.001***
Prior encounter [Linear]	0.601	0.192	3.136	0.002**
Prior encounter [Quadratic]	0.032	0.194	0.163	0.870

Signif. codes: 0'***' 0.001'**' 0.01'*'

Table 1: Fixed effects of the final model.

Class	Participants			LLM Detection Systems	
	Precision	Recall	F1-score	GPTZero	Binoculars
Human	48.5 (12.4)	57.3 (19.0)	51.9 (14.2)	58.8 - 100.0 - 74.1	66.7 - 60.0 - 63.2
AI	48.2 (17.3)	40.6 (17.6)	43.0 (16.2)	100.0 - 30.0 - 46.2	63.6 - 70.0 - 66.7
Macro	48.4 (13.8)	49.0 (11.8)	47.5 (12.6)	79.4 - 65.0 - 60.1	65.2 - 65.0 - 64.9

Table 2: Author prediction task. Left: human participants (mean (std.)). Right: LLM detectors (pre, rec, F1).

The results in Table 2 show that the averaged performance of the participants on the author prediction task was not higher than random chance, reflecting both the difficulty of the task and the success of GPT4o. As shown in Table 2, the LLM detectors scored substantially higher than random chance, but by no means showed high accuracy.

3.2 Qualitative analysis

Now, we shall discuss the best and worst rated AI roasts in detail, and analyze why participants predicted either the right or wrong author for these roasts (recall that participants were allowed to put their reasons into words in the questionnaire). In addition, we attempt to identify general trends in AI roasts, in terms of both style and content.

In the best rated AI roast, 3 jokes were made about the height of Kevin Hart. Although this roast was rated the funniest, there are a number of considerations to be made: first, the joke "your big break was playing a Lego in Toy Story" is not factual, since Kevin Hart did not feature in that movie. Further, "You're so tiny, you could hang-glide with a Dorito", seems original, but after a Google search, it became clear that this is an existing "yo mama" joke. The final joke, in contrast, appears to be original and based on relevant facts: "I heard they tried to cast you in Jumanji as a rock, but even Dwayne couldn't find you." This joke builds upon the knowledge that Kevin Hart starred in Jumanji, while incorporating a size-joke (a rock is small, Kevin Hart is small; Kevin Hart is a rock). Simultaneously, there is also a reference to Dwayne "The Rock" Johnson, who also stars in Jumanji.

Conversely, participants who believed this roast was written by a human highlighted that the sentences were rather short, as in spoken language, that the first person view was prominent, and that the imagery was too striking to be AI-generated.

The lowest-rated AI roast was that of Pete Davidson: "Alright, folks, let's talk about Pete Davidson. This guy's career has skyrocketed, but he still lives in his dad's shadow. I mean, Pete, you lost your father on 9/11, but look on the bright side, you're still the second biggest disaster to come out of that day." This roast riffs on a single joke, but does not develop it logically: There is a reference to 9/11, an event in which Davidson's father died, but GPT4o falsely implies that Davidson was also born on the same day. This illogical joke development, however, was not the main reasons why participants incorrectly predicted the author of this roast. Instead, they reported that the main reasons for their decision are that the roast contains "spoken" language ("alright folks") and remarkable brutality.

Considering all roasts used for the survey, we found that GPT4o relies on a number of frequent tics and phrases to imitate the style of comedians: ("..., (am I) right?", "Let's talk about [target]", "But hey, (at least) ..."; "I mean, ..."; "But let's be honest, ..."). Regarding the content, on the other hand, two general trends were observed. First, GPT4o often mentions specific statistics (e.g., "You [Charlie Sheen] were the highest-paid actor on TV at $1.8 million an episode."; "$2 million in settlements? That's one expensive 'extracurricular activity"' [about James Franco]"). Secondly, the model often ends a roast on a positive (if sometimes sarcastic) note. For example: "[David Hasselhoff] You're proof that you don't need to act well when you look good in red trunks, bravo."

4 Conclusion

A small-scale survey indicated that, contrary to our initial hypothesis, GPT4o can generate roast content that is perceived to be as funny as human content, since no statistically significant difference was found in the funniness ratings by the participants of the survey. Interestingly, we found that there were significant effects of "prior encounter" and "author prediction", highlighting important evaluation considerations when using human raters.

Further, a qualitative analysis also showed that GPT4o used specific stylistic tactics to construct roasts, including mentioning specific statistics or facts retrieved via the prompt, ending the roast on a positive note, and using phrases commonly used by comedians to link sentences or jokes. This indicates that there are certain features that can be detected in the writing style of GPT4o, although the

participants did not pick up on this, as evidenced by their no-better-than-random performance on this task. Generative AI predictors, on the other hand, showed higher accuracy than the participants, although performance was still sub-optimal, contradicting our initial hypothesis. However, this result is presumably the effect of the specificity of the genre and the relatively short text lengths.

5 Limitations

The number of roasts included in the survey was limited to 20 fragments to avoid annotator fatigue: participants reported that completing the survey took between 1 and 2 hours. In future work, however, a similar study may be conducted with a larger sample size, using a between-subjects design, and/or using complete roast contributions of specific comedians.

In addition to the limited size of the survey in terms of roasts, the background of participants was relatively homogeneous, given that they were all students from the same Master's program.

Finally, note that the human generated roasts were written for a live performance, i.e. for a specific audience, format, and for verbal use. The GPT generated roasts on the other hand, has no information about the audience, or does not write with the purpose of live, verbal performance. As mentioned, however, the human data was selected to be interpretable without additional context.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models. *Preprint*, arXiv:2311.16867.

R. H. Baayen. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Neha Reddy Bogireddy, Smriti Suresh, and Sunny Rai. 2023. I'm out of breath from laughing! I think? A dataset of COVID-19 Humor and its toxic variants. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1004–1013, New York, NY, USA. Association for Computing Machinery.

Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 2024. Are U a Joke Master? Pun Generation via Multi-Stage Curriculum Learning towards a Humor LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 878–890, Bangkok, Thailand. Association for Computational Linguistics.

Rune Haubo Bojesen Christensen and Maintainer Rune Haubo Bojesen Christensen. 2015. Package 'ordinal'. *Stand*, 19(2016).

Drew Gorenz and Norbert Schwarz. 2024. How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. *PLOS ONE*.

Hans, Abhimanyu and Schwarzschild, Avi and Cherepanova, Valeriia and Kazemi, Hamid and Saha, Aniruddha and Goldblum, Micah and Geiping, Jonas and Goldstein, Tom. 2024. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, and Naihao Deng. 2024. Chumor 1.0: A Truly Funny and Challenging Chinese Humor Understanding Dataset from Ruo Zhi Ba. *Preprint*, arXiv:2406.12754.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 855–869, Bangkok, Thailand. Association for Computational Linguistics.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. 2025. Safety tax: Safety alignment makes your large reasoning models less reasonable. *Preprint*, arXiv:2503.00555.

Jens Lemmens and Victor De Marez. 2026. Computational Humor Modeling: A Survey on the State of the Art. *ACM Comput. Surv.*, 58(7).

Jianquan Li, XiangBo Wu, Xiaokang Liu, Qianqian Xie, Prayag Tiwari, and Benyou Wang. 2023. Can Language Models Make Fun? A Case Study in Chinese Comical Crosstalk. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7581–7596, Toronto, Canada. Association for Computational Linguistics.

Rod A. Martin and Thomas E. Ford. 2018. *The Psychology of Humor*. Elsevier Inc.

Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. 2024. A Robot Walks into a Bar: Can Language Models Serve as Creativity Support-Tools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians. In *Proceedings of the*

2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, page 1622–1636, New York, NY, USA. Association for Computing Machinery.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating Humorous Puns with Ambiguous Context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2023. GPTZero.

OpenAI. 2024. GPT-4o.

R Development Core Team. 2022. *R: A Language and Environment for Statistical Computing*.

Adrianna Romanowski, Pedro H. V. Valois, and Kazuhiro Fukui. 2025. From Punchlines to Predictions: A Metric to Assess LLM Performance in Identifying Humor in Stand-Up Comedy. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–46, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Maria Sukhareva. 2025. Why em-dashes are common in llm outputs. https://folkertstijnman.com/blog/why-em-dashes-are-common-in-llm-outputs/. Accessed: 2026-04-08.

Joe Toplyn and Ori Amir. 2025. Can AI Make Us Laugh? Comparing Jokes Generated by Witscript and a Human Expert. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 71–78, Online. Association for Computational Linguistics.

JingJie Zeng, Liang Yang, Jiahao Kang, Yufeng Diao, Zhihao Yang, and Hongfei Lin. 2024. "Barking up the Right Tree", a GAN-Based Pun Generation Model through Semantic Pruning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2119–2131, Torino, Italia. ELRA and ICCL.

Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth S. Narayanan, and Salman Avestimehr. 2024. Creating a Lens of Chinese Culture: A Multimodal Dataset for Chinese Pun Rebus Art Understanding. *Preprint*, arXiv:2406.10318.

A Survey participants

In total, 64 responses from unique participants were collected. All participants were Master's students in Professional Communication and Management at the University of Antwerp, of which virtually all students were native speakers of Dutch, but were proficient in English due to their multilingual university education.

B Roast example

Bruce Willis. What a career, right? "The Fifth Element," "The Sixth Sense," "The Whole Nine Yards," "Twelve Monkeys," zero Oscars. And it's not just action movies that made Bruce a star. He's actually a great dramatic actor, too. I loved "The Sixth Sense." It's a great movie and it's a really impressive performance. I don't know how you pretended not to be embarrassed while a 10 year old kid acted circles around you, but you did it. And the ending, I did not see that twist coming. I mean, I shouldn't spoil it, but, it's been like 20 years. It's so good. Okay, so at the end of "The Sixth Sense," Bruce goes back to making bad movies.

C Prompt

You are a stand-up comedian, and it is your task to generate a comedy roast about roastee. A roast is a form of comedy in which a specific individual is subjected to jokes at their expense, intended to amuse the event's wider audience. Roasts are intended to honor a specific individual in a unique way. In addition to jokes and insult comedy, such events may also involve genuine praise and tributes. The assumption is that the roastee can take the jokes in good humour and not as serious criticism or insult. A short example of such a roast can be found below, in which {example_roastee} is the subject:

{example_roast}

To generate a comedy roast, you will draw inspiration from the Wikipedia page of {roastee}, which is included below:

{short_wiki_bio}

Your output must only contain the roast, which must be between 50 and 100 words long. The topic of the roast must be about {summary of the human roast snippet}.

Figure 2: Prompt used for the roast generation.

D Temperature

Initial prompt fine-tuning experiments indicated that a temperature setting of 0.7 was optimal, since the model started to misinterpret the task or generate gibberish at higher temperature settings.

E Non-normalized Likert scores

Author	Mean	Std.
Human	2.31	0.36
AI	2.40	0.25
Combined	2.35	0.30

Table 3: Funniness ratings obtained in the survey (mean and standard deviations).

Phonetic Cues Improve LLM-Based Pun Detection in Short Text

Adith Santosh Thaniserikaran
Purdue University
athanise@purdue.edu

Govind Harikrishnan
Purdue University
gharikr@purdue.edu

Abstract

This paper studies joke detection in short text, focusing only on jokes triggered by lexical ambiguity. Following Attardo and Raskin, we treat these jokes as cases where humor arises from a script opposition activated through a logical mechanism such as homography or homophony. Our framework combines contextual semantic analysis for homographs with phoneme-level similarity for homophones and near-homophones, using CMUdict, weighted Levenshtein distance, and prompt-based reasoning to recover ambiguities that are not visible in spelling alone. Results show that explicit phonetic modeling improves detection of sound-based puns.

1 Introduction

Automatic humor detection has attracted increasing attention in computational linguistics, but recent work suggests that humor understanding remains highly sensitive to the specific linguistic mechanism involved. Studies of caption ranking, pun-focused reasoning, stand-up transcript analysis, and LLM-based humor evaluation show that current systems still struggle when humor depends on structured ambiguity rather than broad surface cues or general fluency (Zhou et al., 2025; Cocchieri et al., 2025; Romanowski et al., 2025; Goes et al., 2023). In particular, pun-centered work highlights that recognizing verbal humor requires models to recover the specific linguistic relation that licenses the joke, rather than merely detect that a text appears playful or surprising (Cocchieri et al., 2025). The core problem addressed in this paper is that existing humor detection systems do not reliably identify puns whose humor depends on different forms of ambiguity, especially when the relevant cue is phonetic rather than purely orthographic or semantic. A model may detect that a sentence is unusual or joke-like without correctly recovering the ambiguity that actually produces the humorous effect. This makes pun detection difficult because successful analysis requires the system to determine not just whether a text is humorous, but what specific linguistic mechanism supports that humor. This paper therefore focuses narrowly on ambiguity-driven pun detection in short text, especially cases where humor arises from either lexical-semantic ambiguity or phonetic similarity. This focus is also motivated by humor theory: following Attardo and Raskin's General Theory of Verbal Humor, verbal jokes can be analyzed in terms of script opposition and logical mechanism, making puns a particularly suitable subtype for computational study because their humor often depends on a clearly identifiable ambiguity trigger (Attardo and Raskin, 1991). Later work on logical mechanisms further reinforces this view by treating ambiguity, analogy, and partial incongruity resolution as central to how verbal jokes are structured and interpreted (Hempelmann and Attardo, 2011; Attardo et al., 2002). To address this problem, we present a combined system for detecting pun-based humor from both written and sound-based cues. The framework integrates lexical preprocessing and part-of-speech filtering, phoneme extraction with the CMU Pronouncing Dictionary (Carnegie Mellon University, 1998), a weighted Levenshtein metric (Levenshtein, 1966) for phonetic similarity, GPT-5.4-based ambiguity detection and phonetic cue augmentation for ambiguity reasoning. Rather than treating humor detection as only a binary joke/non-joke task, the system aims to identify the linguistic mechanism underlying the humor and determine whether the ambiguity is meaningful in context. In this way, the proposed framework is aligned with recent calls for mechanism-aware humor analysis in both computational and theoretical work (Cocchieri et al., 2025; Romanowski et al., 2025; Attardo and Raskin, 1991).

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 72–80

2 Related Work

Related work on computational humor in our paper is most relevant in three areas: humor understanding with large language models, pun-specific evaluation, and humor theory. Recent LLM studies show that humor remains difficult when success depends on recovering the mechanism that makes a text funny, rather than relying on broad fluency or surface plausibility (Zhou et al., 2025; Romanowski et al., 2025; Goes et al., 2023). In particular, work on pun-focused benchmarking shows that current models often recognize apparent wordplay without reliably verifying whether the ambiguity is actually valid, especially in misleading or structurally constrained cases (Cocchieri et al., 2025). This challenge is closely tied to humor theory. In the General Theory of Verbal Humor, verbal jokes are analyzed in terms of script opposition and logical mechanism, and later work argues that ambiguity and partial incongruity resolution are central to how many jokes are structured (Attardo and Raskin, 1991; Attardo et al., 2002; Hempelmann and Attardo, 2011). These theoretical accounts also motivate separating pun-based humor into homographic and homophonic types, since they rely on different forms of ambiguity and therefore may require different computational treatments. Our work builds on these threads by focusing narrowly on short-text puns whose humor depends on lexical-semantic ambiguity, phonetic similarity, or both; we treat weighted Levenshtein distance as a lightweight cueing heuristic informed by prior phonetic-similarity and phonological-alignment work (Kondrak, 2000; Fontan et al., 2016), rather than as a novel phonological similarity metric. Unlike prior LLM-based humor studies that evaluate broader humor understanding or judgment (Zhou et al., 2025; Goes et al., 2023; Romanowski et al., 2025), we target mechanism-level detection of pun-based humor and explicitly model the ambiguity that licenses the joke.

3 System Overview

The proposed system is a hybrid humor-detection pipeline that combines rule-based phonetic analysis with GPT-5.4-based semantic reasoning and classification (OpenAI, 2026). It is designed to detect pun-based humor by modeling two major sources of ambiguity: *phonetic ambiguity*, which underlies homophone-based jokes, and *semantic ambiguity*, which underlies homograph-based jokes. Rather than relying only on an end-to-end language model prediction, the system first extracts linguistically motivated cues and then uses GPT-5.4 to interpret them in context. At a high level, each input sentence passes through a sequence of processing stages. The text is first preprocessed through tokenization, content-word filtering, part-of-speech tagging, and lemmatization. These steps reduce noise and prepare the sentence for both symbolic phonetic analysis and semantic ambiguity detection. The phonetic module then consults the CMU Pronouncing Dictionary to obtain candidate pronunciations for relevant tokens. Using these phoneme representations, the system computes pairwise similarity with a custom weighted Levenshtein distance that accounts for vowel–vowel, consonant–consonant, and vowel–consonant substitutions differently. Word pairs whose normalized similarity exceeds a predefined threshold are treated as candidate homophone or near-homophone cues. When a strong phonetic match is found, the system augments this information directly into the classification prompt. This phonetic cue augmentation step makes hidden sound-based ambiguity explicit for GPT-5.4, enabling the model to reason about puns that may not be obvious from spelling alone. In this way, the phonetic module serves as a symbolic front end that supplies interpretable evidence to the language model. If no strong phonetic ambiguity is detected, the system invokes GPT-5.4 in a separate semantic-analysis stage to identify possible homographs or words with multiple contextual meanings. The model is prompted to return structured output describing any such ambiguity, allowing semantic cues to be incorporated into downstream classification. Finally, GPT-5.4 performs the joke classification step using the sentence together with any detected phonetic or semantic cues. The system outputs a binary prediction (*Joke* or *Non-joke*) along with a short natural-language explanation of the reasoning behind the decision. If the initial explanation is too brief or underspecified, a secondary prompt is used to elicit a clearer justification. This makes the overall pipeline not only predictive but also auditable, since each decision can be traced back to explicit phonetic or semantic evidence.

3.1 Humor-Theoretic Motivation

The system is also grounded in humor theory, particularly incongruity-based accounts of verbal humor (Attardo and Raskin, 1991; Hempelmann and

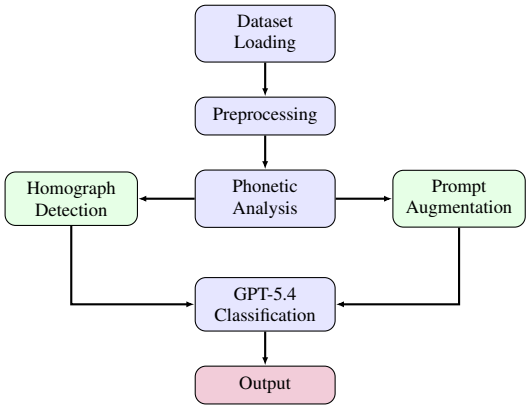

Figure 1: Compact block diagram of the humor detection system.

Attardo, 2011). In both homophone-based and homograph-based jokes, humor arises when a sentence supports more than one plausible interpretation, creating an incongruity between an initial reading and a later, competing one. The joke becomes successful when the listener or reader recognizes this hidden ambiguity and reinterprets the utterance accordingly. In this sense, pun-based humor is not merely a surface phenomenon of unusual words, but a structured interaction between ambiguity, incongruity, and partial resolution in context.

4 Methodology

Our humor-detection pipeline integrates rule-based phonetic modeling, semantic ambiguity resolution, and large language model reasoning. The system is modular, with each component responsible for capturing a distinct linguistic signal associated with humor. Figure 1 provides an overview of the full architecture.

4.1 Text Pre-processing

We begin by normalizing the input text through standard NLP operations, including tokenization, lowercasing, and punctuation removal. Using NLTK (Loper and Bird, 2002), we filter out stopwords and retain only *content-bearing* tokens (nouns, verbs, adjectives, adverbs). Each surviving token is lemmatized and POS-tagged to support downstream phonetic and semantic analysis. This step reduces noise and ensures that ambiguity detection concentrates on linguistically meaningful words.

4.2 Phoneme Mapping and Phonetic Modeling

A central component of the extended humor-detection system involves modeling sound similarity between words. Because written text does not directly encode pronunciation, the system must derive phonetic structure before it can detect homophones or near-homophones. To accomplish this, we rely on the CMU Pronouncing Dictionary (CMUdict), a well-established phonetic lexicon widely used in speech and computational linguistics research. The CMUdict provides pronunciations for more than 134,000 English words and maps each entry to its canonical ARPABET phoneme sequence, a standardized set of phonetic symbols. In our implementation, the system attempts to lookup using both the surface form of a token and its lemmatized form, ensuring coverage for inflected variants (e.g., cats $\rightarrow$ cat). If either the original token or its lemma appears in the CMUdict, its phoneme sequence is stored for further similarity analysis. All pronunciations are retrieved in ARPABET, a compact symbol set representing the phonemes of American English (e.g., CAT $\rightarrow$ K AE T). ARPABET is particularly useful because it disambiguates vowels and consonants clearly. Also, it provides a discrete sequence suitable for edit-distance metrics. This representation enables systematic comparisons between phoneme sequences, which is essential for identifying humor based on subtle sound similarities.

Why a Phoneme-Based Approach? Humor involving homophones and near-homophones depends on similar pronunciation, not spelling. A purely text-based model—such as one trained on embeddings—cannot reliably detect jokes like: "When the band broke up, it wasn't drama — it was just drum!" In orthographic form, drum and drama appear unrelated, yet acoustically they share strong phonetic overlap. A phonetic ambiguity often drives the humor mechanism, and without phoneme-level modeling, these jokes appear semantically incoherent. Therefore, phoneme-based modeling is crucial. Thus, the phonetic layer provides the foundation for identifying sound-driven ambiguity before handing contextual interpretation over to GPT-5.4.

4.3 Weighted Levenshtein Distance for Sound Similarity

After extracting ARPABET phoneme sequences from the CMU Pronouncing Dictionary, the system computes sound similarity using a custom *weighted Levenshtein distance* algorithm. Unlike the standard Levenshtein formulation which assigns a uniform cost to insertions, deletions and substitutions, our method incorporates phonetic knowledge to better model near-homophone similarity relevant to pun-based humor. To account for phonetic closeness, the substitution cost is determined by the phoneme classes of the segments involved:
1) 0 cost for identical phonemes.
2) .5 cost for vowel–vowel or consonant–consonant substitutions.
3) 1 cost for vowel–consonant substitutions.
Insertions and deletions follow the standard unit cost. This scheme captures coarse phonetic similarity by treating substitutions within the same broad phoneme class as less costly than substitutions across classes. For example, vowel–vowel substitutions such as AE $\rightarrow$ EH receive lower cost than vowel–consonant substitutions.

However, this weighting is intentionally coarse: it does not distinguish finer phonological features such as voicing, place of articulation, or manner of articulation. Therefore, the resulting similarity score should be interpreted as a lightweight heuristic for identifying candidate near-homophones, rather than as a full phonological similarity model. From a humor-theoretic perspective, this step is important because homophone-based puns rely on phonetic ambiguity between expressions that sound alike but differ in meaning. Such ambiguity creates the incongruity that underlies many forms of verbal humor: the listener initially interprets one form, then recognizes a competing sound-linked interpretation that shifts the meaning of the utterance. By identifying word pairs with strong phonetic overlap, the weighted Levenshtein module serves not only as a similarity measure but also as a computational mechanism for detecting one of the central triggers of pun-based humor.

Score Normalization Let P_1 and P_2 denote the phoneme sequences of two words. If $D_{weighted}(P_1, P_2)$ is the minimal weighted edit distance between them, the similarity score S is computed as:

$$S = 1 - \frac{D_{weighted}(P_1, P_2)}{\max(|P_1|, |P_2|)}. \quad (1)$$

This normalization bounds similarity in the interval $[0, 1]$, allowing consistent comparison across words of different lengths.

Homophone Threshold

Following empirical evaluation and linguistic validation from the project, a threshold of:

$$S > 0.8 \quad (2)$$

is used to identify homophones or near-homophones. This value was chosen because it reflects a high degree of phonetic similarity while still allowing for small pronunciation differences that commonly occur in near-homophonic word pairs. In practice, true homophones and plausible near-homophones tend to achieve scores close to 1, whereas unrelated pairs with only partial sound overlap usually fall below this range. As a result, the 0.8 cutoff provides a practical balance between retaining meaningful phonetic ambiguities and reducing false positives. Pairs exceeding this threshold are passed to the language model as candidate phonetic ambiguities, improving precision in the detection of sound-based puns. The weighted Levenshtein framework thus provides a linguistically motivated and computationally efficient mechanism for detecting phonetic similarity, serving as the core of the system's homophone-based humor recognition.

4.4 Homograph Detection via GPT-5.4

If no strong phonetic ambiguity is found, the system switches to detecting semantic ambiguity. Many jokes rely on homographs—words with multiple possible senses—whose interpretation depends on context. While lexical resources such as WordNet (Miller, 1995) contain sense inventories, they lack contextual reasoning, coverage for creative language, and the ability to decide which sense-shift is relevant to humor. To address this, we use GPT-5.4 as a context-aware semantic ambiguity detector. This stage also reflects humor theory, since homograph-based jokes depend on lexical ambiguity that activates multiple semantic scripts at once. Humor emerges when the surface reading of a word is suddenly replaced or enriched by another valid sense, producing an incongruous reinterpretation. GPT-5.4 is prompted to identify words in the sentence that could function as homographs and to list their alternative interpretations in a structured JSON format. This design ensures interpretability and easy downstream parsing. The exact prompt used in our system is shown below:

```
 You are a linguistic analyzer.
Find any word in the following sentence
that could have multiple meanings
(homograph pun).
List its possible interpretations
clearly.

Sentence: "SENTENCE_HERE"

Answer in JSON format like:
{"word": "flies", "meanings": ["move
quickly", "insects"]}
If no such word, return {"word": null}.
```

GPT-5.4 returns either a homograph candidate with its sense set or a null indicator. This structured output allows us to attach explicit semantic rationale to the joke classification stage. Because GPT-5.4 can leverage context, pragmatic cues, and world knowledge, we use it as a flexible context-aware alternative to dictionary-based sense lookup for humor-related ambiguity.

4.5 Phonetic Cue Augmentation

Phonetic ambiguity is not directly observable from written text, and large language models often fail to recognize homophone-based wordplay unless the relevant sound information is explicitly supplied. To address this, we use a phonetic cue augmentation mechanism: whenever the phonetic module detects a high-scoring homophone pair (similarity ≥ 0.80), the pair is programmatically inserted into the GPT-5.4 classification prompt.

In theoretical terms, this step helps the model recover an incongruity that may remain hidden in written form. Human listeners can often notice such ambiguity through pronunciation or contextual expectation, but text-only systems may miss it unless the latent phonetic relation is made explicit.

This augmentation provides GPT-5.4 with side information about the possible sound-based ambiguity, allowing the model to reason about humor mechanisms that would otherwise remain hidden. The augmented prompt explicitly highlights the homophone pair and requests a joke judgment using ambiguity-aware reasoning. The core prompt template is shown below:

```
 You are a humor detector.
This sentence likely relies on a
*homophone pun* (sound-based).
Analyze whether the text expresses
deliberate wordplay or simple absurdity.

Detected homophone pair: "W1" "W2"
Phonetic similarity = SCORE

Examples:
- "The chef couldn't bear it when the
bare bear stole his pie." → Joke
- "The knight trained every night before
the tournament." → Non-joke

Now classify the following:
"SENTENCE_HERE"
Answer:
```

The augmented pair narrows the model's search space by making the candidate sound-based relation explicit. This does not guarantee a correct humor judgment, but it provides the classifier with an interpretable cue that may otherwise be implicit in written text. Prompt augmentation therefore acts as a bridging layer between the rule-based phonetic module and the LLM classifier, allowing the system to combine deterministic similarity scores with contextual language-model reasoning.

4.6 Classification and Rationale Generation

The final stage integrates the outputs of the phonetic module, homograph analysis, and conditionally augmented prompts to perform humor classification. GPT-5.4 receives a structured prompt that includes (1) the input sentence, (2) any detected homophone pair or homograph information, and (3) examples of how ambiguity influences joke interpretation. GPT-5.4 then produces a label ("Joke" or "Non-joke") followed by a short natural-language explanation.

To maintain deterministic behavior, the system post-processes the model's response by extracting the text after the "Answer:" tag and reading the first line as the predicted label. Heuristic matching (e.g., checking for "joke" vs. "non-joke") ensures robustness to minor variations in formatting. The explanation is taken from subsequent lines containing keywords such as "reason" or "because". If the model provides no explanation or only a very short one (fewer than three words), the system automatically issues a secondary clarification prompt:

```
  You classified the text below as
"LABEL".
Now, explain your reasoning in 2–3
sentences with clear semantic logic.
Focus on ambiguity, wordplay, or
literalness.

Text: "SENTENCE_HERE"
```

This re-prompting mechanism ensures that each prediction is accompanied by a human-readable post-hoc rationale grounded in identifiable linguistic cues. We treat these explanations as useful decision summaries rather than faithful accounts of the model's internal reasoning. The final output includes the predicted label, the refined rationale, and any detected homophone or homograph cues,

creating an auditable record of the cues used by the pipeline.

5 Experimental Setup

5.1 Dataset

The evaluation was conducted on a subset of 500 short text instances drawn from the SemEval 2017 Task 7 pun detection dataset (Miller et al., 2017), consisting of 250 jokes and 250 non-jokes. The joke examples primarily contain pun-based humor, including both phonetic wordplay and semantic ambiguity, while the non-joke examples consist of proverbs, aphorisms, literal statements, and witty but non-humorous expressions. This balanced setup allows the system to be evaluated not only on its ability to detect humor, but also on its ability to distinguish genuine jokes from linguistically ambiguous yet non-humorous text.

Category	Count	Percentage
Joke	250	50.0%
Non-joke	250	50.0%
Total	500	100.0%

Table 1: Label distribution of the 500-example evaluation subset drawn from the SemEval–2017 Task 7 pun detection dataset.

5.2 LLM Configuration

All GPT-based components in the pipeline, including homograph detection, ambiguity reasoning, and final joke classification, were implemented using **gpt-5.4** (OpenAI, 2026). Unless otherwise specified, all model calls were executed with **temperature = 0.0** and a maximum of **160 completion tokens**. The zero-temperature setting was chosen to reduce output variability and encourage consistent classification behavior across runs, while the token limit was sufficient to allow both label prediction and brief natural-language justification without unnecessarily increasing generation length.

5.3 Task

The task is formulated as a binary classification problem: given an input sentence, the system predicts whether it should be labeled as *Joke* or *Non-joke*. In addition to the final label, the system also generates a short explanation describing the reasoning behind the prediction, allowing us to analyze whether the model relies on phonetic similarity, semantic ambiguity, or other contextual cues.

5.4 System Variants

We evaluate a hybrid humor-detection system together with several baselines and ablations in order to measure the contribution of each component.

Full system. The full system combines three sources of information: (1) direct LLM-based classification, (2) phonetic similarity cues derived from pronunciation-based matching, and (3) homograph or semantic ambiguity cues identified through language-model reasoning. These signals are integrated through prompt-guided reasoning to produce the final prediction.

Baselines. Two main baselines are considered:

- **plain_gpt**: a direct LLM classifier with no additional ambiguity-focused guidance.
- **ambiguity_prompt**: an LLM classifier explicitly prompted to focus on ambiguity and wordplay when making its prediction.

Ablation settings. To understand the contribution of each component, we evaluate the following ablations:

- **phonetic_only**: prediction based only on phonetic similarity cues.
- **no_prompt**: the hybrid system without the phonetic cue augmentation component.
- **no_homograph**: the hybrid system without homograph or semantic ambiguity reasoning.
- **no_phonetic**: the hybrid system without phonetic similarity cues.

These comparisons allow us to determine whether the full model truly benefits from the integration of symbolic and LLM-based signals, and which components are most responsible for performance gains.

5.5 Evaluation Metrics

We report *accuracy*, *macro F1*, and per-class F1 scores for the *Joke* and *Non-joke* classes. Accuracy provides an overall measure of correctness, while macro F1 gives equal weight to both classes and is therefore more informative in assessing balanced classification quality. Per-class F1 scores are included to show whether the system performs equally well on humorous and non-humorous instances, or whether it is biased toward one class.

5.6 Evaluation Goal

The primary goal of the evaluation is not only to measure overall classification performance, but also to determine whether phonetic and semantic ambiguity cues provide meaningful improvements over a plain LLM baseline. The ablation study is therefore central to the analysis, as it reveals which components contribute useful complementary information and which have limited practical impact in the current system design.

6 Results

Mode	Acc.	MF1	J-F1	NJ-F1
plain_gpt	0.710	0.692	0.767	0.617
ambiguity	0.684	0.655	0.755	0.556
phonetic_only	0.504	0.368	0.075	0.661
no_prompt	0.724	0.709	0.775	0.644
no_homograph	0.732	0.718	0.780	0.656
no_phonetic	0.610	0.551	0.714	0.389
full_system	**0.732**	**0.720**	**0.780**	**0.660**

Table 2: Performance of baseline and ablation variants on the 500-example evaluation set. MF1 = Macro F1, J-F1 = Joke F1, NJ-F1 = Non-joke F1.

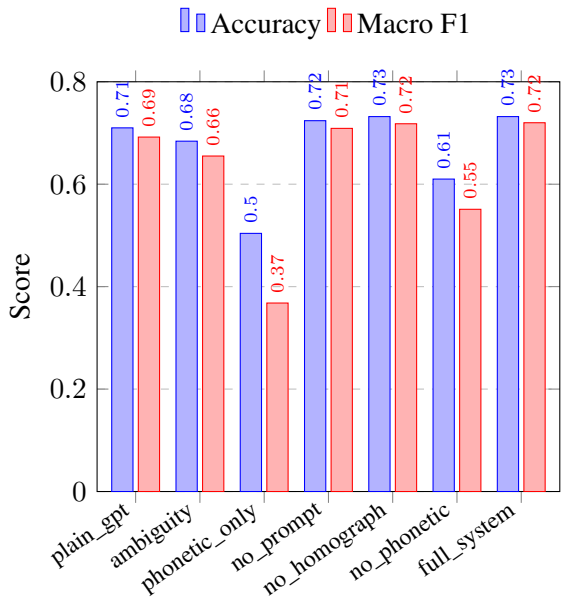

Figure 2: Accuracy and Macro F1 across baseline and ablation variants. The plot visually summarizes the same results reported in Table 2, with the largest drop occurring when phonetic cues are removed.

7 Discussion

The results show that the proposed hybrid system provides a *modest* but consistent improvement over the plain GPT baseline on the 500-example evaluation set. The full system achieves an accuracy of 0.732 and a macro F1 score of 0.720, compared with 0.710 accuracy and 0.692 macro F1 for the plain GPT setting. Although this gain is not large, it suggests that explicit linguistic cues can provide useful complementary information beyond direct LLM classification alone.

The ablation study provides a clearer picture of which components are responsible for this improvement. The strongest finding is the effect of removing phonetic information: the `no_phonetic` variant drops substantially to 0.610 accuracy and 0.551 macro F1. This indicates that phonetic similarity is the most important auxiliary signal in the current architecture. At the same time, the `phonetic_only` condition performs poorly overall, with an accuracy of 0.504 and a macro F1 score of 0.368. Taken together, these two results suggest that phonetic cues are not sufficient on their own, but they become highly valuable when combined with contextual reasoning from GPT-5.4. In other words, the phonetic module functions best as a complementary signal rather than as a standalone humor detector.

By contrast, removing the homograph module has almost no effect on overall performance. The `no_homograph` setting achieves 0.732 accuracy and 0.718 macro F1, which is nearly identical to the full system. This suggests that, in the current implementation, semantic ambiguity reasoning contributes far less than phonetic modeling. One possible explanation is that GPT-5.4 already captures some homograph-based ambiguity implicitly during classification, reducing the added value of a separate homograph-detection stage. Another possibility is that the semantic ambiguity module is currently too weak or too loosely integrated to provide a measurable benefit.

The role of prompt design is also informative. The `ambiguity_prompt` baseline performs worse than the plain GPT baseline, indicating that simply directing the model to attend to ambiguity does not reliably improve humor classification. In fact, making ambiguity too salient may encourage the model to over-predict humor in sentences that are merely proverb-like, ironic, or stylistically unusual. The `no_prompt` ablation performs slightly worse than the full system, suggesting that phonetic cue augmentation contributes a small but positive effect. This indicates that prompt steering is more useful when tied to explicit phonetic evidence than when applied as a general instruction.

An additional pattern emerges from the class-wise results. The full system achieves a much

higher recall for *Joke* instances than for *Non-joke* instances, indicating that the model remains biased toward predicting humor. This means that many non-joke examples are still misclassified as jokes, especially when they contain lexical ambiguity, stylistic oddity, or witty phrasing. This is an important limitation of the current system and reflects a deeper challenge in computational humor: not all ambiguous language is humorous, and not all surprising language is intended as a joke. Distinguishing deliberate pun-based humor from aphorisms, ironic statements, and creative but non-humorous expressions remains difficult even for LLM-based systems.

Overall, these findings support a nuanced conclusion. The hybrid system does not dramatically outperform the plain LLM baseline, but it does show that phonetic modeling provides meaningful complementary value when combined with contextual reasoning. The experiments also reveal that the current gains are driven primarily by phonetic cues rather than by homograph-based semantic ambiguity modeling. Future work should therefore focus on improving non-joke discrimination, strengthening the semantic ambiguity component, and evaluating whether more targeted integration of homograph reasoning can yield larger and more balanced performance gains.

8 Limitations

A notable limitation of the current system is its inability to recognize *multi-word or cross-token homophones*. While the model performs reliably for single-word homophones and homographs, the phonetic analysis pipeline assumes that each candidate homophone corresponds to a single lexical token. During preprocessing, the system queries the CMU Pronouncing Dictionary only for individual words, and phonetic similarity is computed exclusively between isolated tokens.

As a result, the system does not evaluate whether a *sequence of words* may form a phonological unit whose pronunciation matches another lexical item. This limitation becomes apparent in jokes such as:

> *My therapist told me to express my feelings, so now I scream for ice cream.*

Here, the humor relies not on the single word "scream," but on the multi-word sequence "I scream," which phonetically overlaps with "ice cream." Because the system analyzes "I" and "scream" independently, it never constructs the combined phoneme sequence necessary to compare against the pronunciation of "icecream." Consequently, the weighted Levenshtein algorithm cannot identify the intended sound-based ambiguity.

Although GPT may still classify the sentence as a joke based on contextual reasoning, the phonetic ambiguity is not captured at the algorithmic level.

Addressing this limitation would require extending the phoneme extraction module to operate over character n-grams or token n-grams, enabling the system to capture a broader range of sound-based humor that arises beyond single word boundaries. Unlike APUN-Bench, which evaluates pun understanding from audio, our system only approximates sound-based ambiguity from written text using CMUdict, leaving prosody and pronunciation variation for future work (Su et al., 2026).

9 Conclusion and Future Work

This work presented a unified humor-detection framework for identifying both homograph-based and homophone-based wordplay in written text. The system integrates phoneme extraction using the CMU Pronouncing Dictionary, a weighted Levenshtein phonetic similarity metric, and contextual semantic analysis through GPT-5.4. By combining symbolic linguistic methods with large language models, it provides both reliable classification and interpretable explanations of the linguistic mechanisms underlying humor.

The evaluation results show that the pipeline can recognize semantic ambiguity, phonetic similarity, and context-dependent humor across jokes and non-jokes. Overall, the findings suggest that computational humor detection benefits from hybrid architectures that combine structured linguistic processing with LLM-based inference.

Future work could incorporate n-gram phoneme modeling to capture multi-word homophonic sequences, requiring phrase-level phoneme synthesis and similarity computation over variable-length units. The framework could also be extended to detect humor arising from other mechanisms, such as morphological ambiguity, metaphorical reinterpretation, and pragmatic incongruity.

References

Salvatore Attardo, Christian F. Hempelmann, and Sara Di Maio. 2002. Script oppositions and logical mech-

anisms: Modeling incongruities and their resolutions. *Humor*, 15(1):3–46.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: joke similarity and joke representation model. *Humor*, 4(3–4):293–347.

Carnegie Mellon University. 1998. The CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Accessed 2026-05-12.

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, July 27–August 1, 2025. Association for Computational Linguistics.

Lionel Fontan, Isabelle Ferrane, Jérôme Farinas, Julien Pinquier, and Xavier Aumont. 2016. Using phonologically weighted levenshtein distances for the prediction of microscopic intelligibility. In *Proceedings of Interspeech 2016*, pages 650–654, San Francisco, USA.

Fabricio Goes, Piotr Sawicki, Marek Grzes, Dan Brown, and Marco Volpe. 2023. Is gpt-4 good enough to evaluate jokes? Unpublished manuscript. Manuscript.

Christian F. Hempelmann and Salvatore Attardo. 2011. Resolutions and their incongruities: Further thoughts on logical mechanisms. *Humor*, 24(2):125–149.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 288–295. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *arXiv preprint cs/0205028*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

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.

OpenAI. 2026. Gpt-5.4. https://developers.openai.com/api/docs/models/gpt-5.4. Accessed 2026-05-12.

Adrianna Romanowski, Pedro H. V. Valois, and Kazuhiro Fukui. 2025. From punchlines to predictions: A metric to assess llm performance in identifying humor in stand-up comedy. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–46. Association for Computational Linguistics.

Yuchen Su, Shaoxin Zhong, Yonghua Zhu, Ruofan Wang, Zijian Huang, Qiqi Wang, Na Zhao, Diana Benavides-Prado, and Michael Witbrock. 2026. Words at play: Benchmarking audio pun understanding in large audio-language models. *arXiv preprint arXiv:2603.18678*.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T. Rogers, Lalit K. Jain, Robert D. Nowak, Bob Mankoff, and Jifan Zhang. 2025. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in llms. *arXiv preprint arXiv:2502.20356*.

Does Bigger Mean Funnier? Evaluating Humor Generation Across the Qwen3 Model Family

Jatin Agrawal
LTRC, IIITH
jatin.agrawal@research.iiit.ac.in

Radhika Mamidi
LTRC, IIITH
radhika.mamidi@iiit.ac.in

Abstract

We investigate whether scaling model parameters improves humor generation through a controlled ablation study. Using five Qwen3 variants (8B–235B, dense and MoE), we generate jokes across 50 themes. Beyond evaluating humor scaling, this work serves as an empirical study into the nature of LLM versus human evaluations on highly subjective creative tasks. While an automated judge yields a perfect monotonic ranking between parameter count and win rate, human annotators find no significant aggregate difference in humor quality. Restricting to themes where annotators agree reveals a *significant* preference for the largest model ($p = 0.039$), suggesting scaling effects exist but are masked by a "quality floor"—a baseline level of structural competence below which models rarely fall, making their outputs perceptually indistinguishable. Crucially, our analysis of bias characteristics shows that the automated judge exhibits substantial positional and length biases compared to human evaluators, further suggesting that LLMs may systematically distort quality differences on subjective tasks.

1 Introduction

Scaling laws suggest that increasing model size yields predictable improvements in language modeling (Kaplan et al., 2020). However, whether these trends extend to creative generation tasks remains unclear. Humor requires world knowledge, pragmatic inference, and timing—capabilities that may not automatically emerge with additional parameters. As LLMs are widely deployed in creative applications, understanding their ability to generate genuinely funny content is practically important, especially since state-of-the-art models often struggle with repetition and nuance (Jentzsch and Kersting, 2023; Hessel et al., 2023).

We present a dual-purpose investigation. Primarily, we conduct the first controlled, within-family ablation study of parameter count on humor generation quality. By restricting testing to the Qwen3 family, we isolate parameter count from confounds like training data and architecture, evaluating both dense and Mixture-of-Experts (MoE) models to distinguish between a model's total parameter capacity and the subset of "active" parameters it uses per token. Secondarily, we empirically examine the fundamental differences between automated LLM-as-a-judge evaluation and human annotation on highly subjective tasks.

Our contributions are:

1. A controlled within-family ablation study of humor generation across five Qwen3 models (8B–235B), combining automated pairwise judging with blinded human annotation.
2. A high-agreement subset analysis revealing that scaling effects are significant when annotators can distinguish joke quality ($p = 0.039$), but are masked in the aggregate by a "quality floor."
3. Extensive analysis showing that automated LLM judges exhibit substantial procedural artifacts—notably a 73.5% positional bias and significant length bias ($p < 0.0001$)—potentially overstating model differences that humans do not perceive.
4. An open-source pipeline, dataset, and annotation data for reproducibility, available at https://github.com/J10Official/Humor-Parameter-Ablation.

2 Related Work

Humor Generation and Scaling. ChatGPT repetitively produces the same jokes (Jentzsch and Kersting, 2023), and broad model comparisons often confound architectural benefits with training data advantages (Evstafev, 2025). While recent work focuses on generation via prompting strate-

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 81–94

gies (Tikhonov and Shtykovskiy, 2024; Kim and Chilton, 2025), the underlying capability scaling of humor remains unexplored. Power-law scaling holds for language modeling loss (Kaplan et al., 2020), but whether creative humor scales smoothly, emerges suddenly (Wei et al., 2022), or plateaus remains open. Our controlled *within-family* ablation isolates the parameter count variable.

Evaluation Disconnects and Subjectivity. LLM judges exhibit vulnerabilities like positional and verbosity biases, struggling on open-ended subjective tasks (Wang et al., 2023). Our work extends these findings to creative generation, where we observe substantial procedural biases. Subjectivity in humor evaluation causes low agreement (Castro et al., 2018; Mittal et al., 2021), but as Sandri et al. (2023) argue, this disagreement is informative rather than merely noisy. We show that aggregate human disagreement conceals structured patterns, dropping sharply when models hit a "quality floor."

3 Methodology

3.1 Models

We evaluate five instruction-tuned models from the Qwen3 family (Qwen Team, 2025): three dense models (8B, 14B, 32B) and two Mixture-of-Experts (MoE) models that route to a subset of active parameters per token: 30A3 (30B total, 3B active) and 235A22 (235B total, 22B active). All models share the same training pipeline and tokenizer, differing in parameter count and architectural approach. The inclusion of both dense and MoE architectures allows us to disentangle the effects of total parameter count from active (per-token) parameter count.

3.2 Data

We curate 50 everyday humor themes (e.g., *"Autocorrect text message fails," "Airport security lines"*). Themes span a broad range of everyday situations—including technology, social interactions, workplace life, and domestic scenarios—while avoiding domain-specific knowledge requirements, culturally sensitive topics, and themes likely to produce offensive content. The full theme list is available in the repository. Each model generates one joke per theme using a standardized prompt (Appendix A.1), with `max_tokens=128` and thinking disabled.[1]

[1]The Qwen3 `/no_think` control token suppresses chain-of-thought reasoning, ensuring purely generative output.

3.3 Automated Evaluation

We perform pairwise comparisons using DeepSeek-V3.2 as a judge via OpenRouter. We deliberately select a non-Qwen judge to avoid potential family bias in evaluation.[2] For each theme, all $\binom{5}{2} = 10$ model pairs are compared in both orderings (A-first and B-first), yielding 20 comparisons per theme. This symmetric design ensures that any positional bias affects all models equally in aggregate. We aggregate results using Bradley-Terry (BT) scoring (Bradley and Terry, 1952), which estimates latent "ability" parameters from pairwise outcomes, and report win rates computed as the fraction of comparisons each model wins.

3.4 Human Evaluation

Two independent groups of six annotators participate in a blinded evaluation of three models (32B, 30A3, 235A22) across 40 themes (themes 11–50). The first 10 themes were reserved for pilot testing, and the two smaller models (8B, 14B) are excluded to keep annotation tractable. Evaluation is conducted simultaneously using a split projected screen, with full details of the annotation protocol provided in Appendix A.2.

4 Results & Discussion

We report automated judge outcomes and artifacts, followed by human ranking results and a subset analysis focusing on themes where annotators can reliably identify a winner.

4.1 Automated Judge Findings

Table 1 presents the LLM judge results across all 50 themes. Win rates increase monotonically with total parameter count, yielding a perfect Spearman correlation ($\rho = 1.0$, $p < 0.001$). The log-linear relationship (Figure 1) shows a clear positive trend, though the three mid-sized models cluster tightly between 49–55%. Furthermore, calculating a theme-level cluster bootstrap standard error (SE) of the Bradley-Terry scores—which accounts for intra-theme performance correlation—reveals that the 95% confidence intervals for these models frequently overlap. This internal variation indicates that while the LLM's aggregate point estimates scale perfectly, the underlying margin of victory

[2]One theme received 19 of 20 planned comparisons due to an API timeout, yielding 999 rather than 1000 total comparisons.

Model	Win Rate	BT Score (± SE)	Avg Rank
235A22	61.0%	0.276 ± 0.030	2.63
32B	54.6%	0.224 ± 0.024	2.84
30A3	51.6%	0.202 ± 0.020	3.18
14B	49.2%	0.187 ± 0.024	2.85
8B	33.5%	0.110 ± 0.014	3.50

Table 1: LLM judge results (50 themes, 999 comparisons). BT = Bradley-Terry score (normalized to sum 1) with cluster-level bootstrap standard error. Win rates and BT scores correlate perfectly with total parameter count ($\rho = 1.0$), though overlapping SE bounds suggest narrow actual margins of victory.

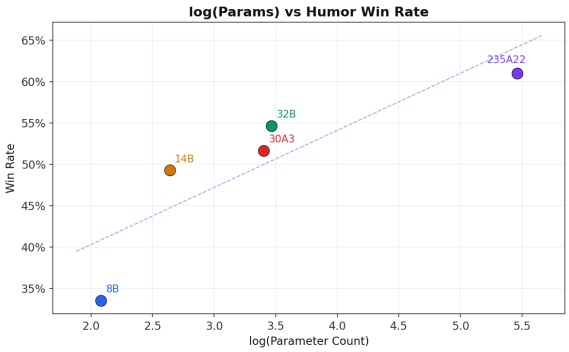

Figure 1: Log parameter count vs. LLM judge win rate. The x-axis represents the log-scaled total parameter count, and the y-axis indicates the fraction of pairwise comparisons won. Error bars denote 95% cluster-bootstrap confidence intervals. The perfect positive correlation ($\rho = 1.0$) indicates that the LLM judge systematically prefers larger models, despite overlapping confidence intervals for mid-sized models.

among consecutive models is narrow, foreshadowing the rating ambiguity observed in human judgments.

The automated judge exhibits severe procedural artifacts, notably positional and length biases. Position A (the joke presented first) wins 73.5% of comparisons ($\chi^2 = 215.9$, $p < 0.001$). While positional bias in LLM judges has been documented in factual evaluation settings (Zheng et al., 2023; Wang et al., 2023), our observed rate of 73.5% is substantially higher than the 60–65% typically reported, suggesting that subjective creative tasks may amplify this bias.

Furthermore, the judge exhibits a strong length bias: when the two jokes differ in character count, it selects the longer joke 62.2% of the time (binomial $p < 0.0001$). This contrasts sharply with our human evaluators, who show no significant preference for longer jokes (Spearman $\rho = -0.065$, $p = 0.082$). The LLM's reliance on length as a proxy for humor highlights a fundamental disconnect between automated metrics and human subjective appreciation.

Model	Mean Rank	Borda	Ranked 1st
235A22	1.900	1.100	40.8%
32B	2.033	0.967	30.4%
30A3	2.067	0.933	28.7%

Table 2: Human evaluation results (12 annotators, 40 themes, 240 total rankings). Mean rank: 1 = best, 3 = worst. Borda score: 0–2 scale. Unlike the automated judge, human evaluators show no statistically significant overall preference for larger models.

To quantify reliability, we find that only 52.7% of transposed model pairs produce consistent winners—not significantly above chance (one-sided binomial test, $p = 0.12$).

4.2 Human Evaluation Findings

Human annotators show a consistent but weaker trend than the LLM judge (Table 2). The largest model (235A22) achieves the best mean rank (1.90) and is ranked first 40.8% of the time, compared to 30.4% for 32B and 28.7% for 30A3. However, a Friedman test reveals no significant difference between models ($\chi^2 = 3.73$, $p = 0.155$). Pairwise Wilcoxon signed-rank tests confirm that no model pair differs significantly (all $p > 0.10$). The rank distribution (Appendix Figure 4) shows that 235A22 receives a disproportionate share of 1st-place rankings (41% vs. 33% expected by chance), but bootstrap 95% confidence intervals for mean rank overlap substantially (Appendix B.2, Figure 5).

Agreement is low (Krippendorff's $\alpha = 0.098$ (Krippendorff, 2011)), consistent with prior humor studies (Castro et al., 2018; Mittal et al., 2021) and reflecting subjectivity rather than annotation error (Sandri et al., 2023). Notably, both groups independently rank 235A22 first (Group 1: 1.88; Group 2: 1.92), suggesting the top-level trend is robust despite low per-item agreement.

Model performance varies substantially across themes (Appendix B.2, Figure 6). This observation motivates our high-agreement subset analysis (§4.3): themes where annotators *can* distinguish between models may reveal scaling effects masked in the aggregate.

4.3 High-Agreement Subset Analysis

We therefore quantify per-theme distinguishability by computing the fraction of annotator pairs that agree on which model produced the funniest joke. With 6 annotators per theme there are $\binom{6}{2} = 15$ pairs; if all 6 agree on the winner, agreement is 1.0, while under random ranking the expected agreement is 0.33. We define high-agreement themes as those with pairwise winner agreement ≥ 0.4 (above chance).[3]

Of the 40 themes, 22 (55%) meet the high-agreement threshold, with agreement scores ranging from 0.40 to 1.00.

Table 3 presents the subset analysis. On the 22 high-agreement themes, the Friedman test is **significant** ($\chi^2 = 6.47$, $p = 0.039$; Kendall's $W = 0.025$, indicating a small effect size), and a pairwise Wilcoxon test shows that 235A22 significantly outperforms 30A3 ($W = 3404$, $p = 0.020$). On the 18 low-agreement themes, the Friedman test shows no effect whatsoever ($\chi^2 = 0.35$, $p = 0.839$), and all three models are statistically indistinguishable (all mean ranks within 0.07 of 2.0).

The subset-restricted figures confirm this pattern visually. Appendix Figure 7 shows that on the high-agreement subset, 235A22's confidence interval separates clearly from 30A3, and Appendix Figure 8 shows a pronounced shift in first-place rankings toward 235A22. The per-theme heatmap restricted to high-agreement themes (Appendix Figure 9) reveals sharper color contrasts between models compared to the all-theme version (Appendix B.2), confirming that model differences are concentrated in this subset.

4.4 Architectural Efficiency: Dense vs. MoE

Comparing the 32B (dense) and 30A3 (MoE) models reveals a sharp divergence in how judges evaluate architectural efficiency. To the LLM judge, the 30A3 (51.6% win rate) performs remarkably close to the 32B dense model (54.6%) despite utilizing just 3B active parameters per token, implying total parameter capacity suffices for structural competence. Conversely, human evaluators show a directional preference for dense compute: in the high-agreement subset, the 32B dense model (mean rank 2.023) outperforms the 30A3 MoE (2.144), though this difference is not statistically significant (Wilcoxon $W = 3916.5$, $p = 0.259$). This suggests that while sparse routing maps adequate world knowledge to formulate a coherent joke, the advantage of dense per-token reasoning for humor remains suggestive rather than conclusive.

Subset	Mean Rank			Friedman p
	235A22	32B	30A3	
High-agree (22)	**1.833**	2.023	2.144	**0.039***
Low-agree (18)	1.981	2.046	1.972	0.839
All (40)	1.900	2.033	2.067	0.155

Table 3: Subset analysis by inter-annotator agreement. On high-agreement themes, the Friedman test reaches significance ($p < 0.05$), and 235A22 significantly outperforms 30A3 (Wilcoxon $p = 0.020$). Low-agreement themes show no model differences, confirming that scaling benefits are only perceptible when evaluating distinguishable topics.

4.5 Qualitative Analysis: Why Scaling Helps

To connect our quantitative findings to humor mechanisms, we examine representative outputs through the lens of the Incongruity-Resolution framework (Suls, 1972), which posits that humor arises when an expectation is established and then subverted by an incongruous but interpretable punchline.

On the highest-agreement theme (*Astrological signs*, agreement = 1.0, 235A22 wins unanimously), the models produce strikingly different outputs:

- **8B:** *"I asked my friend if she believed in horoscopes, and she said, 'I don't believe in them, but I do believe in the power of a good reading to make you feel seen."'* — No incongruity or punchline; the response reads as a sincere comment.
- **32B:** *"I checked my horoscope this morning and it said I'd have a lucky day—turns out the 'lucky' part was the only part they got right."* — Weak incongruity; the subversion ("lucky" being the only accurate part) is vague and lacks a concrete comic image.
- **235A22:** *"I checked my horoscope, and it said today is a great day to reconnect with someone from your past—so I texted my ex. Turned out, the universe meant your* dentist." — Classic incongruity-resolution: the setup establishes a

[3]The result is robust to threshold choice: at 0.35, the same 22 themes qualify ($p = 0.039$); at 0.50, only 6 themes qualify but the rank ordering (235A22 > 32B > 30A3) is preserved ($p = 0.08$, non-significant due to reduced power). The gap between 0.40 and 0.45 is sharp: no themes have agreement in $(0.40, 0.467)$, so 0.40 is a natural boundary.

romantic expectation (texting an ex), then the punchline subverts it with a mundane, dreaded alternative (the dentist). The delayed reveal and italicized final word demonstrate precise comedic timing.

This pattern recurs across high-agreement themes. On *Getting older/Turning 30* (agreement = 0.667, 32B wins), the 32B model produces a dark, unexpected punchline (*"half my friends are dead and the other half are on LinkedIn"*), while 30A3 delivers an extended metaphor that dilutes the comic punch. On *Coffee shop baristas* (agreement = 0.667, 235A22 wins), 235A22's punchline (*"she looked at me like I'd asked for a time machine"*) delivers a sharp, concrete visual incongruity, whereas 30A3's version (*"a side of judgment?"*) relies on a generic quip.

Conversely, on low-agreement themes like *Dealing with landlords* (agreement = 0.2), all three models produce structurally competent but interchangeable jokes—each relies on a familiar complaint-as-metaphor structure without genuine incongruity, explaining why annotators cannot distinguish them.

These examples suggest that scaling benefits manifest specifically as improved *incongruity construction*: larger models more reliably establish an expectation and subvert it with a concrete, surprising punchline, whereas smaller models tend toward generic observations or over-extended metaphors that lack a clear comic turn.

4.6 The Human-LLM Disconnect

The divergence between automated and human evaluation is not merely a matter of degree but of kind. While human evaluators correctly judge structurally competent but equally unfunny jokes as ties, the LLM relies on surface-level proxies (fluency, exact length) that correlate with model size to force a monotonic ranking despite overlapping Bradley-Terry confidence intervals. This is starkly illustrated by the low-agreement themes (§4.3): on themes like *Dealing with landlords*, where all models produce indistinguishable outputs, human annotators appropriately assign near-uniform ranks (all within 0.07 of 2.0), whereas the LLM judge still reports clear preferences—preferences that, as our bias analysis shows, track joke length and presentation order rather than humor.

5 Conclusion

Through our controlled ablation study across the Qwen3 family, we find that scaling parameter count yields a consistent but *non-significant* trend in human-judged humor quality overall. However, restricting analysis to high-agreement themes reveals a *significant* preference for the largest model ($p = 0.039$), indicating that scaling effects are real but frequently masked by a theme-dependent "quality floor."

Crucially, our study highlights the potential danger of using LLM evaluators for creative generation tasks. In stark contrast to human annotators, the automated judge reports a perfect scaling correlation ($\rho = 1.0$) by systematically relying on substantial procedural artifacts, exhibiting 73.5% positional bias and 62.2% length bias. Automated judges may distort subjective quality differences and can be unreliable as standalone tools for assessing creativity.

Limitations

Our study has several limitations that highlight directions for future work. First, regarding evaluation robustness, each model generated only a single joke per theme, precluding the assessment of within-model variance. Furthermore, our dataset is relatively small, and our automated evaluation relies on a single LLM judge (DeepSeek-V3.2), leaving inter-judge reliability untested. Second, the human evaluation was constrained by a small sample size and overall low annotator agreement, and manual ratings were only obtained for three of the five evaluated models. Third, by using the `/no_think` token, we explicitly disabled model reasoning to isolate generative capability. However, recent findings suggest that deliberative inference significantly improves an LLM's humor competence, a factor our design excludes (Narad et al., 2025). Fourth, it remains unclear whether larger models generate novel jokes or retrieve memorized examples (Amir, 2025; Jentzsch and Kersting, 2023). Although our judging prompt includes originality as a criterion, the pipeline parses only the final `WINNER: A/B` token rather than per-criterion scores, so we cannot assess how the judge weighs originality against structural proxies. Finally, while we ground our qualitative analysis in the Incongruity-Resolution framework (§4.5), extending this to additional humor theories (e.g., benign violation, superiority) remains future work.

References

O. Amir. 2025. Are AI-generated jokes truly original? Charting the "joke space.". In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. 2018. A crowd-annotated Spanish corpus for humor analysis. In *Proceedings of the 6th International Workshop on Natural Language Processing for Social Media*, pages 7–11.

Iurii Evstafev. 2025. Optimizing humor generation in LLMs: Temperature configurations and architectural trade-offs. *arXiv preprint arXiv:2504.02858*.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Ali Farhadi, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from The New Yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 688–714.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging for large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pages 325–340. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Hyunwoo Kim and Lydia B. Chilton. 2025. AI humor generation: Cognitive, social and creative skills for effective humor. *arXiv preprint arXiv:2502.07981*.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC)*.

Anirudh Mittal and 1 others. 2021. So you think you're funny?: Rating the humour quotient in standup comedy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

R. Narad, S. Suresh, J. Chen, P. S. Dysart-Bricken, B. Mankoff, R. Nowak, and 1 others. 2025. Which LLMs get the joke? probing non-stem reasoning abilities with HumorBench. *arXiv preprint arXiv:2507.21476*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Elisa Sandri, Elisa Leonardelli, and Sara Tonelli. 2023. Why don't you do it right? Analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2428–2440.

Jerry M. Suls. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Jeffrey H. Goldstein and Paul E. McGhee, editors, *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, pages 81–100. Academic Press.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor mechanics: Advancing humor generation with multistep reasoning. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Experimental Setup

A.1 Prompts

Joke generation prompt. The following system prompt is used for all five Qwen3 models, with `max_tokens=128` and thinking disabled (`/no_think`):

> *"You are a stand-up comedian known for sharp, original observational humor. When given a theme, write one short joke about it — either a classic setup/punchline or a punchy one-liner. The joke must be self-contained, land clearly, and end on the punchline. Output only the joke text, nothing else — no title, no explanation, no commentary."*

The user message is simply the theme text (e.g., "Autocorrect text message fails").

Judging prompt. The DeepSeek-V3.2 judge receives the following system prompt:

> *"You are an expert comedy critic with a sharp, discerning sense of humor. Your task is to decide which of two jokes is funnier."*

The user message follows this template:

Theme: "{theme}"

Joke A:
{joke_a}

Joke B:
{joke_b}

Which joke is funnier? Consider:
- Comedic timing and structure
- Originality of the idea
- Clarity and strength of the punchline
- How well it fits the theme

First, write 2–3 sentences explaining your reasoning. Then on the final line write ONLY:
WINNER: A
or
WINNER: B

A.2 Annotation Protocol & Interface

Annotation protocol. Annotators are split into two independent groups of six. Evaluation is conducted simultaneously using a split projected screen that displays different themes to each group (Figure 2). For each theme, three anonymized jokes are presented in a random order for 60 seconds while annotators independently rank them from funniest to least funny.

B Additional Results & Visualizations

B.1 LLM Judge Per-Theme Heatmap

B.2 Human Evaluation: All 40 Themes

Figures 4, 5, and 6 show the full aggregate results across all 40 themes, including the 18 low-agreement themes where models produce indistinguishable outputs. These figures complement the high-agreement subset analysis presented in the main text (§4.3).

B.3 High-Agreement Subset Figures

Figures 7, 8, and 9 provide visualization for the high-agreement subset analysis presented in the main text (§4.3).

C Qualitative Examples & Data

C.1 Example Jokes

Table 4 shows example model outputs for selected themes, illustrating cases where models are clearly distinguishable (high agreement) and cases where outputs converge (low agreement).

C.2 High-Agreement Theme List

The 22 themes meeting the high-agreement threshold (pairwise winner agreement ≥ 0.4) are listed below with their agreement scores and per-model mean ranks.

Figure 2: The split-screen annotation interface projected in the classroom. The left half displays a theme for Group 1 (themes 11–30), while the right half simultaneously displays a different theme for Group 2 (themes 31–50). This layout allows 12 annotators (two groups of six) to evaluate simultaneously without seeing each other's responses. Each set of three randomized, anonymized jokes (A, B, C) is displayed for 60 seconds, a duration determined via pilot testing, while annotators independently rank them from funniest to least funny.

Theme	**32B**	**30A3 (MoE)**	**235A22 (MoE)**
High agreement — Astrological signs (agreement = 1.0, 235A22 wins unanimously)			
Astrological signs	I checked my horoscope this morning and it said I'd have a lucky day—turns out the "lucky" part was the only part they got right.	I told my zodiac sign I was breaking up, and it said, "But we're meant to be!" I said, "Yeah, I know, that's why I'm going to find someone who's not on my level."	I checked my horoscope, and it said today is a great day to reconnect with someone from your past—so I texted my ex. Turned out, the universe meant your *dentist*.
Low agreement — Dealing with landlords (agreement = 0.2, three-way tie at rank 2.0)			
Dealing with landlords	I told my landlord his rent increase was like a horror movie—it had a slow build, a shocking revelation, and left me screaming into the void.	I told my landlord I was going to start a band, he said, "Great, but can you at least pay rent on time?"	I asked my landlord if the apartment included a gym, and he said, "No, but the stairs are free."

Table 4: Example themes illustrating high- and low-agreement cases. Full generated outputs for all 50 themes across all 5 models are available in the repository. The high-agreement examples show clear qualitative differences in punchline strength, whereas low-agreement jokes demonstrate similar structural competency but equal lack of comedic effect.

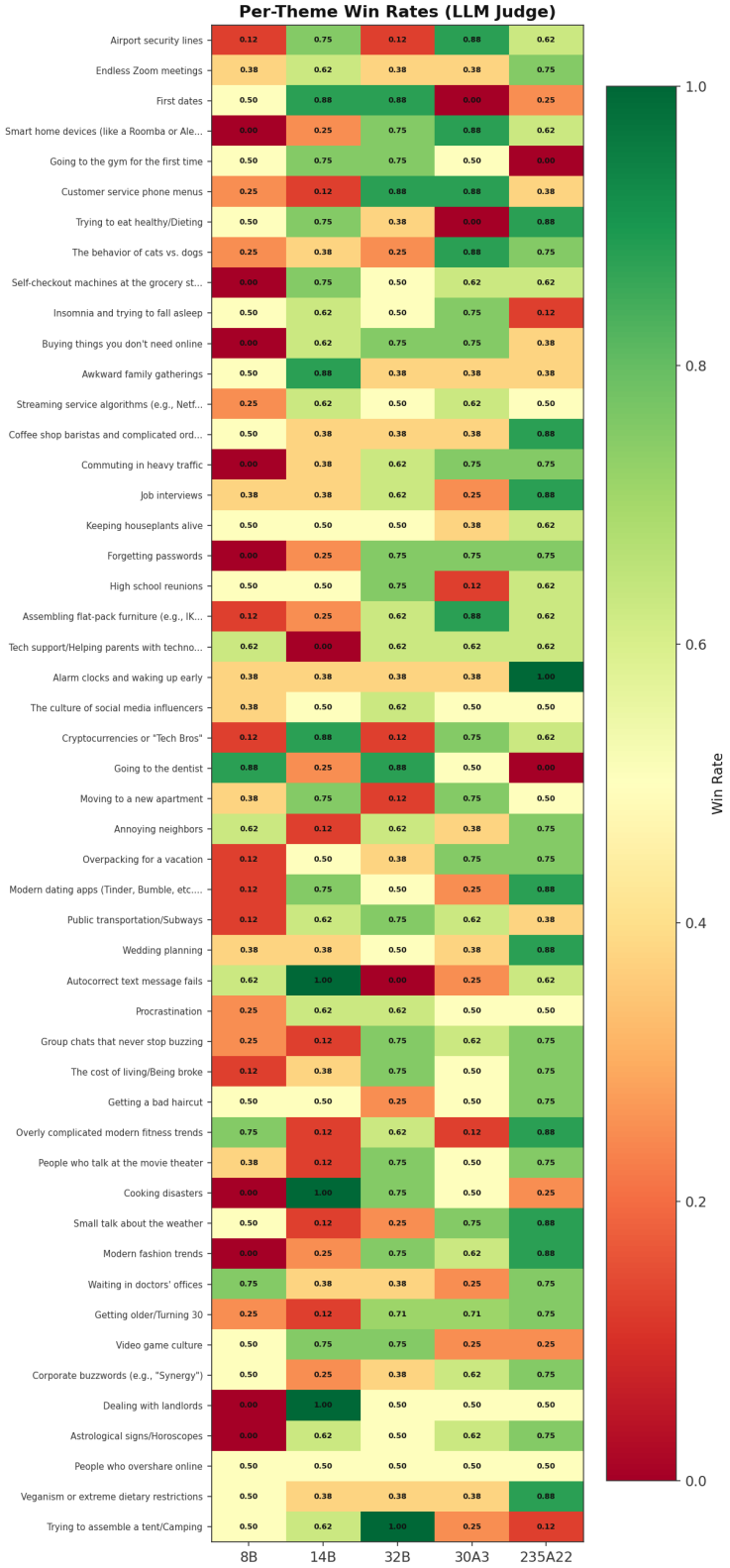

Figure 3: LLM judge per-theme win rates across all 5 models (50 themes). Warmer colors (green) indicate higher win rates. The 235A22 model shows consistently high win rates while 8B shows consistently low win rates, but per-theme variation is substantial. Reading horizontally reveals a given model's performance variance across topics, while reading columns vertically illustrates the judge's size preference on specific topics.

Theme	Winner	Agreement	32B	30A3	235A22
Astrological signs/Horoscopes	235A22	1.000	3.00	2.00	**1.00**
Trying to assemble a tent/Camping	32B	1.000	**1.00**	2.67	2.33
Getting a bad haircut	235A22	0.667	2.00	2.83	**1.17**
People who talk at the movie theater	30A3	0.667	2.00	**1.33**	2.67
Getting older/Turning 30	32B	0.667	**1.17**	2.17	2.67
Coffee shop baristas	235A22	0.667	2.00	2.83	**1.17**
Procrastination	30A3	0.467	2.83	**1.33**	1.83
The cost of living/Being broke	30A3	0.467	1.83	**1.50**	2.67
Veganism/dietary restrictions	235A22	0.467	1.83	2.83	**1.33**
Awkward family gatherings	235A22	0.467	2.50	2.00	**1.50**
Group chats	235A22	0.400	2.17	2.50	**1.33**
Modern fitness trends	235A22	0.400	2.17	2.50	**1.33**
Cooking disasters	30A3	0.400	2.00	**1.67**	2.33
Modern fashion trends	32B	0.400	**1.50**	1.83	2.67
Waiting in doctors' offices	32B	0.400	**1.50**	2.17	2.33
People who overshare online	235A22	0.400	2.17	2.33	**1.50**
Buying things online	235A22	0.400	2.33	2.33	**1.33**
Streaming service algorithms	32B	0.400	**1.67**	2.50	1.83
Job interviews	235A22	0.400	1.83	2.33	**1.83**
Forgetting passwords	30A3/235A22	0.400	2.67	**1.67**	**1.67**
Cryptocurrencies/"Tech Bros"	30A3	0.400	2.67	**1.50**	1.83
Annoying neighbors	32B	0.400	**1.67**	2.33	2.00

Table 5: High-agreement themes (≥ 0.4 pairwise winner agreement). Bold indicates best (lowest) mean rank. 235A22 wins 9 of 22 themes, 32B wins 8, 30A3 wins 4, and 1 is tied.

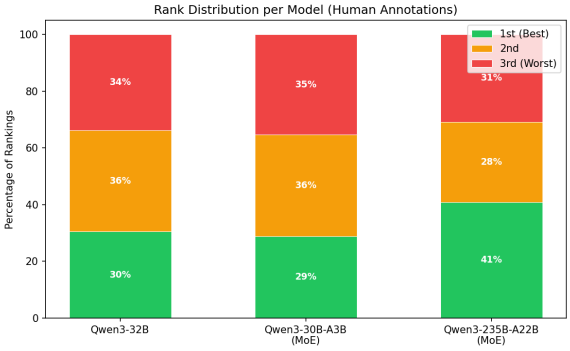

Figure 4: Distribution of 1st, 2nd, and 3rd place rankings across all 40 themes. Bar segments denote the percentage of times each model received a specific rank from human evaluators. While 235A22 receives 41% of first-place votes vs. 33% expected by chance, the overall Friedman test is non-significant ($p = 0.155$).

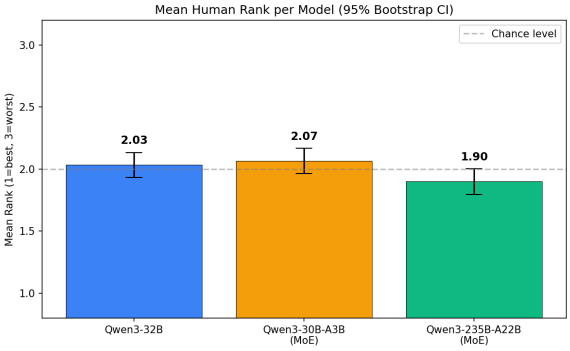

Figure 5: Mean human rank per model with 95% bootstrap CIs across all 40 themes. Points denote the average human rank (lower is better), and vertical lines show 95% confidence intervals. All confidence intervals overlap and include the chance level (2.0). This visualizes the aggregate "quality floor," showing no significant separation between models.

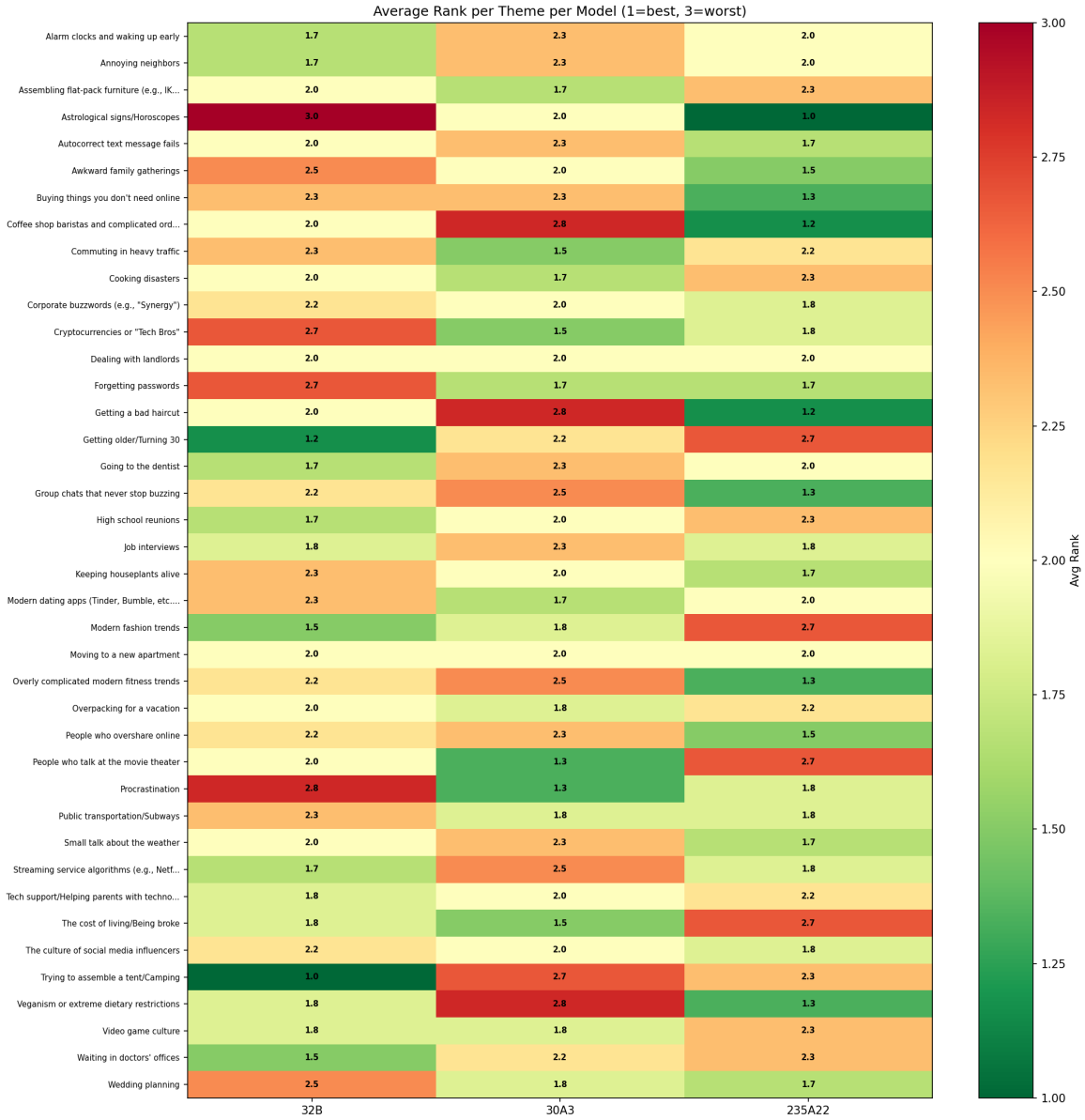

Figure 6: Per-theme average human rank across all 40 themes (1 = best, 3 = worst). Some themes show strong differentiation (dark green for the winner), while many cluster near 2.0 (yellow), indicating that annotators could not distinguish the models. The predominance of yellow cells highlights how frequently models produced indistinguishable outputs.

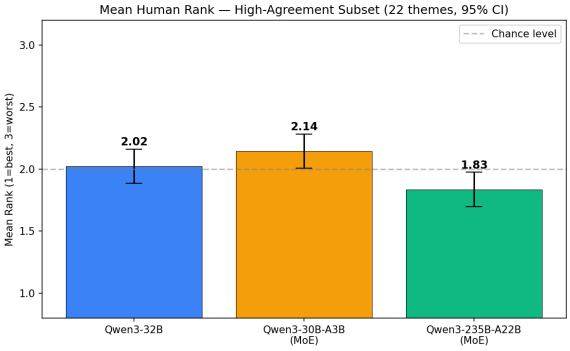

Figure 7: Mean human rank on the 22 high-agreement themes (95% bootstrap CI). Points denote the average human rank on this restricted subset, and vertical lines show 95% confidence intervals. The clear separation of 235A22's interval from 30A3's visually confirms the significant statistical preference ($p = 0.039$) found in this subset.

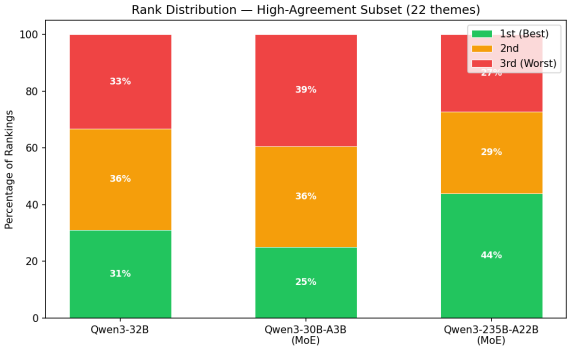

Figure 8: Rank distribution on the 22 high-agreement themes. Bar segments denote the percentage of times each model was ranked 1st, 2nd, or 3rd within this high-agreement subset. Compared to the aggregate distribution, 235A22 secures a noticeably larger majority of first-place votes here.

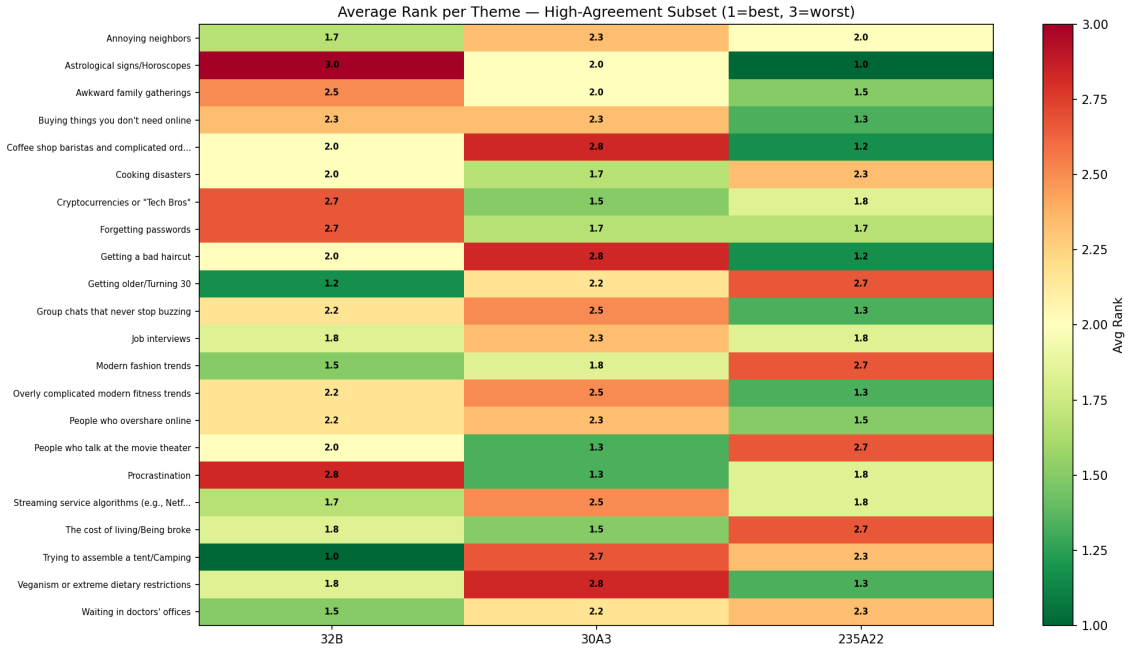

Figure 9: Per-theme average human rank on the 22 high-agreement themes (1 = best, 3 = worst). Dark green indicates better (lower) mean ranks, while light yellow indicates chance-level performance. Sharper color contrasts indicate clearer model differentiation compared to the full 40-theme heatmap. The absence of yellow tie-columns emphasizes that these are themes where distinct, perceptible quality differences emerged.

Navigating the Joke Space: Towards Automated Originality Assessment of AI-Generated Humor

Ori Amir[1,2,3], Huyen Dieu Ngo[1], Joe Toplyn[3,4], Kevin P. Hickerson[5]

[1]Fulbright University Vietnam [2]HaHator, LLC
[3]Semantic AI and Creativity Lab, East Texas A&M University
[4]Twenty Lane Media, LLC [5]California Institute of Technology

Correspondence: ori.amir@fulbright.edu.vn oriacadem@gmail.com

Abstract

This study validates automated, corpus-based methods for quantifying joke originality using "topic handles" — key nouns or noun phrases capturing a joke's script opposition and logical mechanism (per the General Theory of Verbal Humor). Using a reference corpus of one million jokes in English from Reddit, we compute Pointwise Mutual Information (PMI) in three variants (raw co-occurrence, semantic-cluster smoothing, and word-decomposition) and two embedding-based measures (handle-level conceptual distance and full-text corpus novelty via Sentence-BERT). We evaluate these measures on 400 LLM-generated jokes (200 each from GPT-4o and GPT-5.4) and 80 jokes from the Witscript-powered JEST benchmark, rated by three professional comedians for originality and funniness. Corpus novelty and concept distance between the most semantically distant handle pair both correlated significantly with human originality ratings (ρ = .37); PMI-based measures showed weaker but significant associations (ρ = .23–.25) on the most original handle pair. A Lasso-based composite of the three strongest predictors achieved ρ = .40 (cross-validated), capturing 82% of the theoretically predictable variance given inter-rater agreement. These results demonstrate that handle-based PMI and semantic novelty metrics offer practical, quantitative tools for assessing originality in AI-generated humor, advancing objective evaluation of computational creativity.

1 Introduction

Large Language Models (LLMs) can generate jokes on demand, producing humor that rivals the quality of professional comedy writers, as evidenced by audience reactions in comedy club settings (Toplyn and Amir, 2025). However, the originality of LLM-generated jokes remains uncertain (Jentzsch and Kersting, 2023; Veale, 2024). While traditional plagiarism detection can identify verbatim matches, LLMs excel at rephrasing, prompting critics to label them as plagiarism machines (Chomsky et al., 2023) or "stochastic parrots" (Bender et al., 2021). Unlike rule-based joke generation algorithms, which follow transparent processes (Binsted and Ritchie, 1997), LLM-based systems function as opaque "black boxes," obscuring the mechanisms behind their joke creation. Numerous studies investigating the comedic capabilities of LLMs often implicitly assume that the generated jokes are original. However, there is evidence that successful LLM humor is often not original (Jentzsch and Kersting, 2023; Veale, 2024).

To address this issue, Amir (2025) proposed characterizing the "Joke Space" — the set of all possible verbal jokes in English — as a framework for evaluating originality. Even when considering only the core elements of a joke, which make a joke truly unique, the space of all possible jokes is estimated to be at minimum 1 to 32 billion jokes, several orders of magnitude larger than the total number of jokes any LLM or human comedian could plausibly have been exposed to (estimated as 150 million for modern LLMs; Figure 1). This suggests that, in principle, LLMs could be prompted to generate entirely novel jokes. Despite this vast potential, however, both AI systems and human comedians frequently produce non-original jokes. This tendency can be attributed either to prior exposure to similar jokes, or to a focus on highly salient topics (Amir and Biederman, 2016; Amir et al., 2022; Shahaf et al., 2015) or emotionally charged content (Hempelmann and Ruch, 2005).

Determining whether any individual joke is original therefore requires identifying its "essence": the core elements that define its humor independent of surface wording. Drawing on the General Theory of Verbal Humor (GTVH; Attardo and Raskin, 1991), Amir (2025) defines a joke's essence as its Script Opposition (SO) and Logical Mechanism

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 95–105

(LM), with other elements such as language and narrative strategy representing optimization rather than essence. At the implementation level, this essence can be approximated by a joke's topic handles — two key nouns or short noun phrases whose pairing results in the comedic effect (Toplyn, 2014), and the associative link connecting them. Drawing on the definition of originality as statistical infrequency (Dumas et al., 2021; Runco et al., 2024), a joke is more original when its handles co-occur more rarely in natural language generally and in joke corpora in particular. Despite this body of prior work, quantitative corpus-based originality assessment of individual AI-generated jokes remains understudied (Jentzsch and Kersting, 2023; Loakman et al., 2025; Veale, 2024). Most research on AI creativity has instead relied on general frameworks such as the Torrance Test or the Alternative Uses Test, which measure originality as statistical rarity relative to a reference population (Lamb et al., 2018; Runco et al., 2024). These approaches, however, are domain-general and may fail to capture the essential elements of a joke that determine whether two jokes are distinct or, instead, share a common premise based on a particular combination of script opposition and logical mechanism (Ruch et al., 1993; Cain et al., 2024).

Toplyn (2014) proposed a practical operationalization of this framework, originally designed for human comedy writers and later adapted for generative computational humor work (e.g., Toplyn, 2023). The approach involves extracting (usually two) handles from the joke topic and identifying unexpected or pseudo-logical associative links between them (see also Dean, 2000). The handles are typically the key nouns or verbs in the joke topic that are critical for establishing the script opposition and logical mechanism.

This work explores automated methods for assessing joke originality by integrating widely used PMI and semantic embedding techniques with a topic handle extraction framework. We validate these methods against originality judgments by professional comedians, and examine how the relationship between originality and funniness varies across joke sources and generation methods.

2 Methods

2.1 Data and Resources

As a reference corpus, we used the SocialGrep/one-million-reddit-jokes dataset, comprising one million jokes in English drawn from joke subreddits and covering a wide variety of styles (SocialGrep, 2021). For experimentation, 400 AI-generated jokes were produced using the OpenAI API: 200 by GPT-4o and 200 by GPT-5.4. Jokes were generated in batches of 10 using the following system prompt: "You are a professional comedian. You provide raw joke content without any headers or metadata." The user prompt randomly varied the comedic style across five categories (absurdist, satirical, dry wit, wordplay, and observational), with the style selected uniformly at random for each batch, to encourage output diversity. The full user prompt was:

> Generate n original and clever jokes in the style of 'style'. STRICT FORMATTING RULES: (1) Do NOT include any topic names, categories, or labels. (2) 'title' MUST be the setup or question only. (3) 'selftext' MUST be the punchline or answer only. (4) Do NOT use common clichés. Return ONLY a raw JSON array of objects.

All calls used a temperature of 1.0 to maximize output diversity. The model was instructed to return jokes as structured JSON objects with a title field (setup) and selftext field (punchline); the two fields were concatenated to form the full joke text. To prevent repetition, jokes were deduplicated based on normalized full-text content (lowercased, whitespace-collapsed) before inclusion; batches were re-requested as needed until 200 unique jokes per model were collected. An additional 80 items from the JEST benchmark (Toplyn and Amir, 2026) were used for validation. The JEST benchmark was generated using Witscript (Toplyn, 2023), a software system for automated joke writing: 60 items were produced by Witscript in response to short texts and selected to achieve approximately equal representation of funniness ratings of 1 (Somewhat Funny), 2 (Funny), and 3 (Very Funny). The remaining 20 items were generated by the OpenAI model GPT-5.1 to match the benchmark jokes in style and length but to contain no humor; each was assigned a funniness rating of 0 (Not Funny). While the benchmark's funniness categories were by design, the category assignments were confirmed by ratings from experienced professional comedy writers, ensuring that the stratification reflects genuine expert judgment rather than design alone. GPT-4o was selected as it is the most studied LLM in the context of computational humor

and was also the model underlying Witscript, the system used to generate the JEST benchmark. GPT-5.4, the most recent OpenAI model at the time of data collection, was included to assess whether newer models produce more original jokes.

2.2 Data Processing and Semantic Indexing

The title and body fields of each reference corpus entry were concatenated into a single text sequence. All jokes were lowercased and tokenized using a regular expression that retains word tokens of three or more characters; standard stop words (e.g., 'the,' 'and,' 'with') were removed. An inverted index was constructed mapping each token to the set of joke IDs in which it appears. To ensure statistical stability and reduce sparsity, only tokens with a corpus frequency of five or more were retained in the working vocabulary. All vocabulary items were encoded using all-mpnet-base-v2 for subsequent semantic clustering and comparison (Reimers and Gurevych, 2019).

All jokes in the dataset were encoded as dense vector representations using all-mpnet-base-v2. These embeddings were L2-normalized so that all vectors share a uniform magnitude, which allows cosine similarity to be computed efficiently via dot product. To enable fast retrieval of the most semantically similar jokes for any given input, we indexed the normalized embeddings using the FAISS library with an IndexFlatIP configuration (Johnson et al., 2019).

2.3 Topic Handle Extraction

Topic handles were defined as key nouns or short noun phrases representing the core reusable subject matter of a joke, grounded in Toplyn (2014) finding that jokes typically link two discrete concepts pseudo-logically. Handle extraction was performed by prompting GPT-5.4 (temperature 0.1) using the following system prompt:

> You are an expert comedy analyst. Extract the topic handles from a given joke. A topic handle is a key noun or short noun phrase representing the joke's core, reusable subject matter. Handles must be concrete and specific (e.g., 'dog', 'depressed penguin'). Do not include generic nouns (e.g., 'man', 'thing', 'person') or verbs/adjectives that only enable punchline logic (e.g., 'smell', 'think'). Return ONLY a JSON object: 'handles': ['handle1', 'handle2', ...] . All handles must be lowercase strings.

The output was stripped of whitespace and stop words. Multi-word handles were additionally decomposed into their component words (for example, 'depressed penguin' into 'depressed' and 'penguin'), yielding both compound and atomic handles for each joke.

Two examples illustrate the method. For the GPT-5.4 joke "I tried becoming more spontaneous. I've scheduled another attempt for Thursday," the extractor returned: 'spontaneity,' 'schedule,' 'Thursday' — capturing the core script opposition cleanly, with 'Thursday' as minor noise that strengthens the joke through specificity but is not necessary for its logic. For the JEST joke "Giraffes only sleep 30 minutes a day... takes 25 minutes to fluff the neck pillow," it returned: 'giraffes,' 'neck pillow,' 'neck,' 'pillow' — where 'neck pillow' and 'giraffe' are the appropriate key handles, with 'neck' and 'pillow' as the decomposed components.

It is worth noting that this extraction method differs from Toplyn's (2021; 2023) Witscript approach in two respects. First, whereas Toplyn extracts handles from the joke setup during the writing process, here extraction is performed on the full joke text after generation. Second, the prompt used here typically returned more than two handle candidates per joke. For the handle-based measures (PMI variants and handle-level conceptual distance), subsequent analysis therefore enumerated all pairs of handles and computed each measure for every pair, then selected the most original pair per joke. This is conceptually consistent with Witscript's algorithm, in which one criterion for selecting topic handles is maximizing the semantic distance between them (Toplyn, 2021). The corpus novelty measure, which operates on the full joke text rather than handle pairs, was not subject to this pairwise procedure. Our goal was to demonstrate that even this relatively basic extraction approach can capture meaningful originality signals, with methods leveraging deeper joke understanding, a skill that modern LLMs exhibit to a useful extent (Narad et al., 2025), as a natural direction for future work.

2.4 PMI-Based Familiarity Measures

Based on the hypothesis that jokes utilizing less common concept combinations are more original, we adopt Pointwise Mutual Information (PMI) to quantify the frequency of handle co-occurrence

in the reference corpus. For each pair of handles, a raw PMI score is computed from exact co-occurrence counts. Let D(h) denote the set of jokes containing handle h, and N = 1,000,000 be the total number of jokes. The raw PMI is defined as:

$$\mathrm{PMI}(h_1, h_2) = \log_2 \left(\frac{|D(h_1) \cap D(h_2)| \cdot N}{|D(h_1)| \cdot |D(h_2)|} \right) \quad (1)$$

Higher PMI values indicate more frequent co-occurrence, implying a more familiar conceptual pairing and lower originality. Because exact handle co-occurrence is often sparse (especially for specific or multi-word concepts), we introduce a cluster-based PMI variant. Vocabulary items are first grouped into 15 semantic clusters using K-means over pre-trained sentence embeddings. Each handle is mapped to its corresponding cluster, and PMI is computed over the union of joke sets associated with each cluster.

To further address sparsity in multi-word handles, a decomposed PMI score is computed as the mean of raw PMI across all pairs of component words. This preserves partial lexical associations when the full phrase is absent or too rare in the corpus. For each joke, all unordered handle pairs are enumerated and raw, cluster-based, and decomposed PMI scores are computed for each pair.

2.5 Embedding-Based Semantic Measures

Because PMI is insensitive to semantic meaning, an embedding-based method is used to quantify originality through conceptual distance and holistic similarity. Two distinct embedding-based measures were used. At the handle level, conceptual distance between two handles is computed as one minus their cosine similarity:

$$d(h_1, h_2) = 1 - \cos(\mathrm{emb}(h_1), \mathrm{emb}(h_2)) \quad (2)$$

A larger distance indicates a more semantically dissimilar and thus more novel conceptual pairing.

At the joke level, the full-text similarity score s_full is determined as the maximum cosine similarity between the joke and any joke in the reference corpus (of over one million jokes), retrieved via the pre-built FAISS index. A high s_full indicates that the joke is semantically redundant with existing material; a low score suggests greater originality. For analysis, this score is transformed to a corpus novelty measure as 1 - s_full, so that higher values consistently indicate greater originality.

2.6 Human Ratings

Three raters evaluated the jokes for originality and funniness. All three raters are professional comedians, each with at least a decade of experience performing and/or writing comedy in the United States. Each joke was rated on two 4-point scales (0–3). Originality was defined as how original the joke felt relative to the rater's own prior exposure to similar humor: 0 indicated an exact repeat (the rater had heard this specific joke before), 1 indicated low originality (not the exact joke, but something very similar, with the same setup and punchline), 2 indicated a familiar premise (the joke was new to the rater but used a recognizable formula or trope), and 3 indicated full originality (both the premise and execution felt fresh). Funniness was rated on a scale from 0 (not funny; likely to get no reaction) to 3 (very funny; likely to get a laugh), with intermediate values of 1 (almost funny; might get a smile) and 2 (funny; likely to get a chuckle). Raters were instructed that some texts might not be intended to be funny and to read each text carefully before rating. Each rater evaluated 80 jokes from each of the three sources: GPT-4o, GPT-5.4, and the JEST benchmark (i.e., 240 jokes per rater, 480 ratings per rater across both dimensions).

2.7 Statistical Analysis

Inter-rater agreement was assessed on the 80 human-rated jokes per source, as well as pooled across all 240 jokes. Fleiss' κ with quadratic weighting served as the primary measure of absolute rating agreement. Quadratic-weighted Fleiss' κ is the multi-rater generalization of Cohen's weighted κ, penalizing larger disagreements proportionally to the square of the distance between categories, which is appropriate for ordinal scales with more than two raters. Mean pairwise Spearman ρ (the average of all three pairwise rank correlations) served as a complementary measure of relative agreement, capturing the degree to which raters agreed on the rank ordering of jokes independently of differences in scale use. Bootstrapped 95% confidence intervals and p-values were derived from a single resampling pass (2,000 item-level resamples) using the recentering method.

Differences in mean originality and funniness ratings across the three joke sources were tested using Kruskal-Wallis H tests on item-level means, with post-hoc pairwise Mann-Whitney U tests and Holm-Bonferroni correction. Within each source,

the relationship between originality and funniness at the item level was quantified using Spearman ρ, and pairwise differences between the three resulting correlations were tested using both Fisher's z-transformation and a bootstrap test of the difference in ρ, with Holm correction across the three pairs.

To assess the validity of the proposed automated originality measures, Spearman rank-order correlations with human originality and funniness ratings were examined. Spearman's ρ was used throughout as all measures departed significantly from normality (Shapiro-Wilk W ranging from .48 for PMI Cluster Max to .91 for Corpus Novelty, all $p < .001$). All measures were oriented prior to analysis so that higher values consistently indicate greater originality: the full-text BERT similarity score was transformed as 1 - BERT_full to produce a corpus novelty score; the three PMI-based scores were negated (× -1) so that rarer concept pairings yield higher values; the handle-level BERT distance measure required no transformation as it already reflects conceptual distance. For the handle-pair measures (PMI variants and handle-level conceptual distance), scores were summarized across pairs by selecting the most original pair per joke — defined as the pair with the lowest raw PMI (greatest rarity) for PMI measures, and the pair with the greatest cosine distance for the conceptual distance measure. Where all handle pairs for a given joke were absent from the corpus (rendering PMI technically undefined), the PMI score was imputed to the theoretical maximum originality value, corresponding to the minimum possible co-occurrence frequency in a corpus of this size ($-\log_2(N) \approx 15.3$ after negation). Jokes for which no handle pairs could be formed (single-handle jokes) were excluded from PMI-based analyses. Correlation analyses were restricted to the subset of jokes for which all three human raters provided ratings (n = 240: 80 JEST benchmark, 80 GPT-4o, 80 GPT-5.4). Ninety-five percent confidence intervals for all correlations were computed via the Fisher z-transformation.

To assess whether combining measures improves prediction, a composite originality score was constructed using Lasso regression, a penalized regression method that automatically shrinks redundant predictor coefficients to zero. The regularization strength (alpha) was selected via 10-fold cross-validation. Predictive validity of the composite was evaluated using both 10-fold and leave-one-out cross-validation, with out-of-fold Spearman correlations serving as the performance metric. Jokes missing any measure (primarily single-handle jokes lacking pair-based scores) were excluded from the composite analysis (n = 235).

2.8 Implementation

Both the computation of the originality measures and the subsequent analyses were implemented in Python. All semantic operations were performed using the all-mpnet-base-v2 Sentence-Transformer model. All 400 LLM-generated jokes were produced with the OpenAI API.

3 Results

3.1 Inter-Rater Agreement

Pooled across all 240 jokes (80 per source), inter-rater agreement on originality was low. Quadratic-weighted Fleiss' κ was .187 (95% CI [.10, .26], $p < .001$) and mean pairwise Spearman ρ was .240 (95% CI [.15, .32], $p < .001$). Pairwise Spearman correlations ranged from $\rho = .083$ (95% CI [-.05, .21], $p = .209$) to $\rho = .372$ (95% CI [.25, .48], $p < .001$).

Agreement on funniness was substantially higher. Quadratic-weighted Fleiss' κ was .45 (95% CI [.36, .53], $p < .001$) and mean pairwise Spearman ρ was .477 (95% CI [.40, .56], $p < .001$). Pairwise Spearman correlations ranged from $\rho = .353$ (95% CI [.24, .47], $p < .001$) to $\rho = .648$ (95% CI [.55, .73], $p < .001$).

3.2 Originality and Funniness Ratings by Joke Source

Mean originality ratings differed significantly across the three sources (Kruskal-Wallis $H(2) = 45.80$, $p < .001$). The JEST benchmark jokes received the highest mean originality rating M = 2.48 (SD = 0.45), followed by GPT-4o jokes M = 2.09 (SD = 0.48) and GPT-5.4 jokes M = 1.92 (SD = 0.55). Post-hoc pairwise comparisons showed that all three sources differed significantly from one another after Holm correction (Benchmark vs. GPT-4o: $U = 4653.5$, $p < .001$; Benchmark vs. GPT-5.4: $U = 4980.0$, $p < .001$; GPT-4o vs. GPT-5.4: $U = 3857.5$, $p = .022$).

Mean funniness ratings did not differ significantly across sources (Kruskal-Wallis $H(2) = 4.74$, $p = .094$), with means of M = 1.50 (SD = 1.06) for Benchmark jokes, M = 1.24 (SD = 0.60) for GPT-4o jokes, and M = 1.21 (SD = 0.62) for GPT-5.4 jokes. Levene's test indicated significant het-

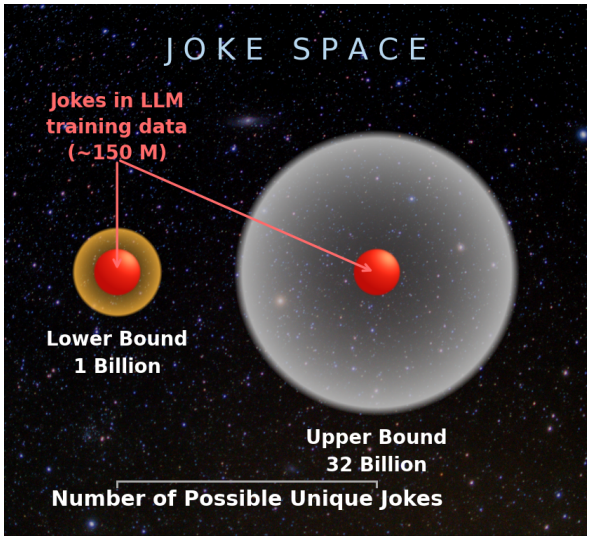

Figure 1: The "Joke Space" versus LLM training coverage. Sphere ("planet") volumes are proportional to estimated unique English joke counts, derived from vocabulary combinatorics under the General Theory of Verbal Humor: noun pairs linked by 5 association bridges each yield the Lower Bound (1B jokes, 14k-noun adult vocabulary) and Upper Bound (32B jokes, 80k WordNet synsets). Both dwarf the estimated jokes in LLM training data (150M; red core, sized identically in both spheres), suggesting LLMs have encountered only a small fraction of all possible jokes and thus have meaningful potential for generating genuinely novel humor. Full derivation is in Amir (2025). Note: all spheres are to scale (the red core represents 15% of the Lower Bound planet's volume and less than 0.5% of the Upper Bound's) though volume differences are difficult to appreciate in a 2D projection.

erogeneity of variance in funniness ratings across sources ($F(2, 237) = 30.69$, $p < .001$), with Benchmark jokes showing considerably greater spread than the two AI-generated sets (as expected per the benchmark's design – see Introduction).

3.3 Relationship Between Originality and Funniness Within Each Source

The relationship between item-level originality and funniness differed markedly across sources (see Figure 2). For Benchmark jokes the correlation was positive (ρ = .554, $p < .001$), for GPT-4o jokes it was negative (ρ = -.247, $p < .001$), and for GPT-5.4 jokes it was weakly positive (ρ = .239, $p < .001$). All three pairwise differences between these correlations were statistically significant after Holm correction by both Fisher's z and bootstrap tests: Benchmark vs. GPT-4o ($\Delta\rho$ = .80, $p < .001$), Benchmark vs. GPT-5.4 ($\Delta\rho$ = .32, p = .018), and GPT-4o vs. GPT-5.4 ($\Delta\rho$ = -.49, p = .004).

3.4 Correlations Between Automated Measures and Human Originality Ratings

Correlation analyses were restricted to the 240 jokes rated by all three raters (80 per source). Within this subset, the mean human originality rating was M = 2.16 (SD = 0.55) and the mean human funniness rating was M = 1.32 (SD = 0.79), both on a 0–3 scale.

All five automated measures correlated positively and significantly with human originality ratings (Figure 3). The two strongest predictors were Corpus Novelty (ρ = .372, 95% CI [.26, .48], $p < .001$) and Concept Distance Max (ρ = .369, 95% CI [.25, .47], $p < .001$), which were nearly identical in magnitude despite capturing distinct aspects of originality: Corpus Novelty reflects how semantically distant an entire joke is from the nearest joke in the reference corpus, while Concept Distance Max reflects the semantic distance between the two most distant handles within the joke itself. Their moderate inter-correlation (ρ = .392, 95% CI [.28, .50], $p < .001$) suggests they capture partially overlapping but meaningfully distinct signals.

The three PMI-based measures showed weaker but consistently significant associations, all in the range ρ = .23–.25: PMI Raw Max (ρ = .231, 95% CI [.11, .35], $p < .001$), PMI Decomposed Max (ρ = .232, 95% CI [.11, .35], $p < .001$), and PMI Cluster Max (ρ = .248, 95% CI [.12, .36], $p < .001$). All three index the co-occurrence rarity of the joke's most original handle pair in the reference corpus. As detailed in Section 3.6, PMI Raw Max and PMI Decomposed Max are near-identical by construction, while PMI Cluster Max captures related but partially independent information through semantic cluster smoothing rather than exact token co-occurrence.

3.5 Correlations Between Automated Measures and Human Funniness Ratings

None of the five automated measures of originality showed a significant correlation with human funniness ratings (all $p > .08$; Figure 3). Corpus Novelty showed a small negative trend (ρ = -.059) and PMI Cluster Max a small positive one (ρ = .073), but neither approached significance. This pattern of null results suggests the automated framework captures something specific to originality rather than to joke quality more broadly.

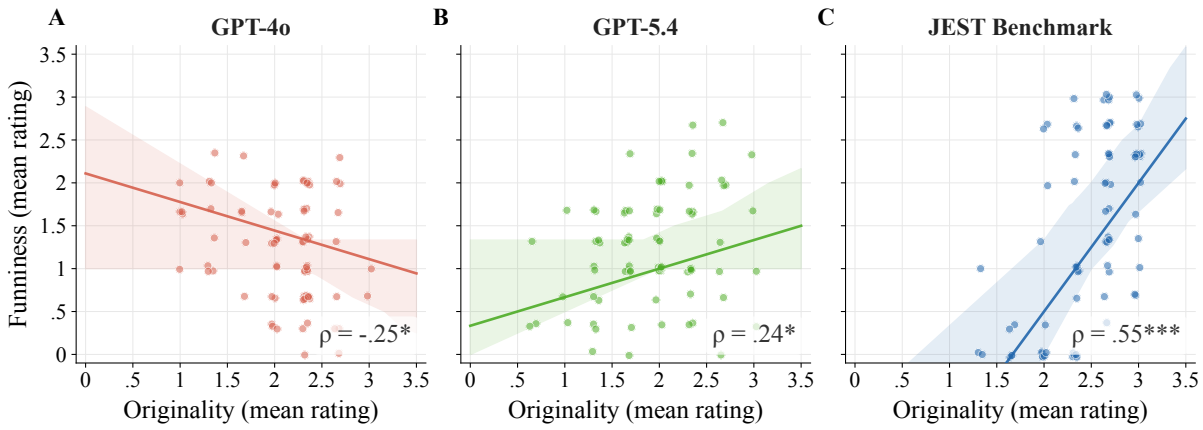

Figure 2: Scatterplots showing the relationship between mean originality and mean funniness ratings by three expert raters for jokes from each source. (A) GPT-4o, (B) GPT-5.4, (C) JEST benchmark. Each data point represents a single joke rated by all three raters, with coordinates reflecting the mean rating across raters. The line shows the Theil-Sen fit with 95% bootstrap confidence band (1,000 resamples). ρ = Spearman rank correlation. * p < .05, *** p < .001.

3.6 Inter-correlations Among Automated Measures

The inter-correlations among automated measures are shown in Figure 3. PMI Raw Max and PMI Decomposed Max were near-perfectly correlated (ρ = .903, 95% CI [.88, .92], p < .001). The two measures return identical values when all handles are single words. The small residual difference arises from the minority of jokes containing at least one two-word handle, which tend to be rarer or absent in the reference corpus. PMI Cluster Max showed moderate correlations with both PMI Raw Max (ρ = .243, 95% CI [.12, .36], p < .001) and PMI Decomposed Max (ρ = .276, 95% CI [.15, .39], p < .001), confirming that the cluster-based smoothing approach captures related but not interchangeable information. Corpus Novelty correlated moderately with Concept Distance Max (ρ = .392, 95% CI [.28, .50], p < .001) and PMI Cluster Max (ρ = .304, 95% CI [.18, .42], p < .001), but showed little relationship with PMI Raw Max (ρ = .098, p = .128) or PMI Decomposed Max (ρ = .088, p = .170), indicating that full-text semantic novelty and handle-pair co-occurrence rarity capture largely independent aspects of originality. Concept Distance Max correlated moderately with all PMI measures (ρ = .222–.440, all p < .001), suggesting partial convergence between semantic distance and co-occurrence rarity, while retaining sufficient independence to contribute distinct information.

3.7 Composite Originality Score

To assess whether combining automated measures improves prediction of human originality ratings, we constructed a composite score using Lasso regression with the alpha parameter selected via 10-fold cross-validation. Applied to the five automated originality measures, Lasso assigned non-zero weights to only three predictors: Corpus Novelty (51%), Concept Distance Max (43%), and PMI Decomposed Max (6%), zeroing out PMI Raw Max and PMI Cluster Max entirely — consistent with their redundancy with PMI Decomposed Max and weaker independent contribution respectively. The in-sample composite correlated with human originality at ρ = .417 (95% CI [.31, .52], p < .001, n = 235 — excluding single-handle jokes). Cross-validation within the 235-joke subset, for which all human ratings and all originality measures were available, yielded ρ = .401 (10-fold) and ρ = .388 (LOO), confirming that the composite generalizes beyond the training sample, though the gain over the best individual measure (Concept Distance Max, ρ = .369) is modest.

To examine whether joke length confounds the composite, word count was added as an additional predictor. Length correlated substantially with human originality ratings (ρ = .358, p < .001), and the Lasso assigned it 27% of the composite weight, partially displacing Corpus Novelty (51% → 41%) and Concept Distance Max (43% → 31%). The cross-validated composite improved marginally (LOO ρ = .39 → .41). However, from a GTVH perspective, joke length should not affect the core script opposi-

tion or logical mechanism that defines originality, and conceptual novelty is independent of how elaborately it is expressed. While we acknowledge that the craft of comedy writing involves careful word selection, empirical evidence suggests professional comedians would consider a reworded joke with the same premise (operationalized as these two structural components) as joke theft (Cain et al., 2024). Length may either create an illusion of originality in human raters or correlate with aspects of originality not directly captured by our measures, but neither justifies its inclusion in a theoretically grounded assessment tool. We therefore retain the length-free composite as our primary measure.

Applying the Spearman-Brown formula to the mean pairwise inter-rater agreement (ρ = .24, k = 3) yields an estimated reliability of the mean originality rating of ρ = .49, the theoretical ceiling for predicting it. The cross-validated composite (ρ = .40) reaches approximately 82% of this ceiling, suggesting the automated measures capture most of the reliably predictable variance in human originality judgments.

4 Discussion

This work explores automated measures of humor originality assessment. Several of the measures correlated significantly with originality judgments of experienced human comedy writers. The hope is that such measures will prove useful for judging the likelihood that human- or AI-generated jokes are novel.

4.1 Relationship between Perceived Originality and Funniness

Three joke sources were used: naïve prompting of GPT-4o and GPT-5.4 to generate original jokes, and the JEST benchmark, where humor was generated in response to short texts using the commercially available Witscript application (accessed November 2025) and curated to achieve a wide distribution of funniness ratings. Figure 2 shows a significant pattern where, in GPT-4o, a negative correlation emerged between perceived originality and funniness by human raters. That pattern reversed for GPT-5.4 and peaked with the strongest positive correlation for the Witscript-generated JEST benchmark. That more original jokes would be perceived as funnier is consistent with the long-held view among some humor researchers as well as comedy writers that surprise is either an essential element of humor or, if not necessary, at least enhances amusement (Hurley et al., 2011; Suls, 1972). However, these results are also consistent with the view that surprise or novelty is not a sufficient condition for humor. The negative correlation between originality and funniness observed in the earliest model (GPT-4o) is consistent with Veale's (2024) findings that this model, when prompted plainly to generate jokes, produces either funny but unoriginal jokes clearly retrieved from its training set or unfunny attempts at original humor. GPT-5.4's moderate positive correlation suggests it may generate somewhat better original jokes. However, even a superficial review of its outputs reveals heavy reliance on variations of classical joke themes. For instance, five of the 200 jokes were variations of puns involving a mirror reflecting: "Why did the mirror become a therapist? It had a gift for helping people reflect" and "Why did the mirror get promoted? It really knew how to reflect well on the company."

Since the JEST benchmark used an LLM-powered application that is algorithmically designed for joke generation, as well as specific setups (short texts) in the prompts, it does not rely as much on preexisting jokes. Consequently, its jokes were rated reliably more original than the jokes produced by the general-purpose models in response to generic prompts (i.e., "generate 10 original jokes").

4.2 Humor Detection vs. Originality

A common criticism of LLM creativity, especially in domains like humor where novelty and surprise are highly valued, is that the statistical pattern completion the model is based on renders it a "stochastic parrot" rather than a source of true creativity (Bender et al., 2021). Classical machine learning approaches for automated humor detection concluded that the best way to recognize humor is to find words that appear more often in jokes compared to non-jokes (sometimes referred to as the "bag of words" approach; Mihalcea and Strapparava, 2006). However, as we have seen, words that are infrequently used in jokes also signal their novelty. Modern approaches for humor detection make use of semantic embedding models with broader context representations (Tasnia et al., 2023). However, as our results show, semantic embedding distance also signals originality. There is thus an inherent tension: the statistical features that make a joke recognizable as humor (e.g., frequent concept pairings and semantic proximity to known jokes) are precisely the features that make it unoriginal.

	1	2	3	4	5	6	7	M	SD
1. Human Originality	—							2.16	0.55
2. Human Funniness	0.289***	—						1.32	0.79
3. Corpus Novelty	0.372***	-0.059	—					0.36	0.10
4. PMI Raw Max	0.231***	0.104	0.098	—				1.49	2.12
5. PMI Cluster Max	0.248***	0.073	0.304***	0.243***	—			0.16	0.20
6. PMI Decomposed Max	0.232***	0.085	0.088	0.903***	0.276***	—		1.69	2.62
7. Concept Distance Max	0.369***	0.092	0.392***	0.222***	0.440***	0.230***	—	0.86	0.10

Figure 3: Spearman rank-order correlation matrix for human ratings and automated originality measures (n = 240, jokes rated by all three raters). Cell shading reflects the magnitude and direction of each correlation (blue = positive, white = 0, red = negative). The lower triangle displays the Spearman ρ coefficient with significance stars and 95% confidence intervals computed via Fisher z-transformation. All measures entering the correlations are oriented so that higher values indicate greater originality; M and SD for PMI measures are shown in the raw (un-negated) scale, where lower values indicate greater originality. Corpus Novelty is the complement of maximum cosine similarity between the joke and the reference corpus (1 - BERT_full); PMI Raw, PMI Cluster, and PMI Decomposed quantify the co-occurrence rarity of the most original handle pair per joke (i.e., the pair with the lowest co-occurrence in the reference corpus); Concept Distance is the embedding-based semantic distance of the most semantically distant handle pair (1 - cosine similarity). Undefined PMI values for jokes where all handle pairs were absent from the corpus are imputed to the theoretical maximum originality score. Human Originality and Human Funniness are mean rater scores on a 0–3 scale. * p < .05, ** p < .01, *** p < .001.

An LLM trained to produce text that looks like humor is therefore also trained, implicitly, to produce humor that looks like existing humor.

4.3 Towards Automated Originality Assessment of AI-Generated Humor

We explored several variant methods for probabilistic automated joke originality assessment. The methods can be reduced to identifying how rarely a joke's most distinctive handle pair co-occurs in a large corpus of jokes, or how semantically distant that pair is. The measures vary in specificity (cluster vs. raw PMI) and sensitivity to context (PMI vs. embedding). A Lasso-based composite of the three strongest predictors: Corpus Novelty, Concept Distance Max, and PMI Decomposed Max — modestly outperformed individual measures on cross-validated prediction of human originality ratings (ρ = .40), suggesting that the measures capture partially complementary signals. Remarkably, when considering the modest human experts' inter-rater reliability, this composite automated measure captures 82% of the reliably predictable variance in their originality judgments. These methods are ideal for cases in which it is clear from the context that a text segment is an instance of humor. The conflict described in Section 4.2 between humor detection and novelty judgment remains to be resolved.

Limitations

All but one of the proposed methods rely on the assumption that jokes can be reduced to a few concepts extractable as "handles" and linked together. While the positive correlation between these measures and human originality ratings is promising, it is possible that this approach only works for a subset of humor.

Since we do not have access to all the humor ever produced nor to the full set of jokes included in the training set of most modern LLMs, a ground truth originality score is impossible to obtain. We must therefore rely on the judgments of human comedy writers. However, even the most experienced comedians have likely been exposed to a smaller set of jokes than modern LLMs (Amir, 2025; Brawer and Amir, 2021), so they are likely to rate some jokes as original that are nevertheless available in an LLM training set.

This work relies exclusively on the SocialGrep (2021) dataset. While this dataset includes over a million highly diverse short jokes sourced from

Reddit, it still likely fails to capture the full richness of human humor. This is fine as a proof of concept, but for real-world applications a broader and more diverse set of reference humor should be used.

References

Ori Amir. 2025. Are AI-generated jokes truly original? Charting the "Joke Space". In *Proceedings of the 16th International Conference on Computational Creativity*, pages 302–308, Campinas, Brazil. Association for Computational Creativity.

Ori Amir and Irving Biederman. 2016. The neural correlates of humor creativity. *Frontiers in Human Neuroscience*, 10:597.

Ori Amir, Konrad J. Utterback, Justin Lee, Kevin S. Lee, Suehyun Kwon, Dave M. Carroll, and Alexandra Papoutsaki. 2022. The elephant in the room: Attention to salient scene features increases with comedic expertise. *Cognitive Processing*, 23(2):203–215.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 4(3-4):293–347.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Kim Binsted and Graeme Ritchie. 1997. Computational rules for generating punning riddles. *Humor: International Journal of Humor Research*, 10(1):25–76.

Jacob Brawer and Ori Amir. 2021. Mapping the "funny bone": Neuroanatomical correlates of humor creativity in professional comedians. *Social Cognitive and Affective Neuroscience*, 16(9):915–925.

Kathleen Cain, Steven Gimbel, Lindsay Howard, Brittany Maronna, and Sean Beirne. 2024. Joke synonymy sensitivity among working comedians and the General Theory of Verbal Humor. *Humor: International Journal of Humor Research*, 37(4):513–528.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. The false promise of ChatGPT. *The New York Times*.

Greg Dean. 2000. *Step by step to stand-up comedy*. Heinemann, Portsmouth, NH.

Denis Dumas, Peter Organisciak, and Michael Doherty. 2021. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, 15(4):645.

Christian F. Hempelmann and Willibald Ruch. 2005. 3WD meets GTVH: Breaking the ground for interdisciplinary humor research. *Humor: International Journal of Humor Research*, 18(4):353–387.

Matthew M. Hurley, Daniel C. Dennett, and Reginald B. Adams, Jr. 2011. *Inside jokes: Using humor to reverse-engineer the mind*. MIT Press.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Carolyn Lamb, Daniel G. Brown, and Charles L. Clarke. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys*, 51(2):1–34.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who's laughing now? An overview of computational humour generation and explanation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 780–794, Hanoi, Vietnam. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

Reuben Narad, Siddharth Suresh, Jiayi Chen, Pine S. L. Dysart-Bricken, Bob Mankoff, Robert Nowak, Jifan Zhang, and Lalit Jain. 2025. Which LLMs get the joke? Probing non-STEM reasoning abilities with HumorBench. *Preprint*, arXiv:2507.21476.

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

Willibald Ruch, Salvatore Attardo, and Victor Raskin. 1993. Toward an empirical verification of the General Theory of Verbal Humor. *Humor: International Journal of Humor Research*, 6(2):123–136.

Mark A. Runco, Burak Turkman, Selcuk Acar, and Ahmed M. Abdulla Alabbasi. 2024. Examining the idea density and semantic distance of responses given by AI to tests of divergent thinking. *The Journal of Creative Behavior*, 59(3):e1528.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1065–1074.

SocialGrep. 2021. One million Reddit jokes [Dataset]. Hugging Face. https://huggingface.co/datasets/SocialGrep/one-million-reddit-jokes (accessed April 2025).

Jerry M. Suls. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Jeffrey H. Goldstein and Paul E. McGhee, editors, *The psychology of humor: Theoretical perspectives and empirical issues*, volume 1, pages 81–100. Academic Press, New York.

Radiathun Tasnia, Nabila Ayman, Afrin Sultana, Abu N. Chy, and Masaki Aono. 2023. Exploiting stacked embeddings with LSTM for multilingual humor and irony detection. *Social Network Analysis and Mining*, 13(1):43.

Joe Toplyn. 2014. *Comedy Writing for Late-Night TV*. Twenty Lane Media.

Joe Toplyn. 2021. Witscript: A system for generating improvised jokes in a conversation. In *Proceedings of the 12th International Conference on Computational Creativity*, pages 22–31, Mexico City, Mexico (Virtual). Association for Computational Creativity.

Joe Toplyn. 2023. Witscript 3: A hybrid AI system for improvising jokes in a conversation. *Preprint*, arXiv:2301.02695.

Joe Toplyn and Ori Amir. 2025. Can AI make us laugh? Comparing jokes generated by Witscript and a human expert. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 71–78, Online. Association for Computational Linguistics.

Joe Toplyn and Ori Amir. 2026. JEST: A benchmark for rating the funniness of short texts. In *Proceedings of the 17th International Conference on Computational Creativity*, Coimbra, Portugal. Association for Computational Creativity. In press.

Tony Veale. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLMs. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*, Jönköping, Sweden. Association for Computational Creativity.

Author Index