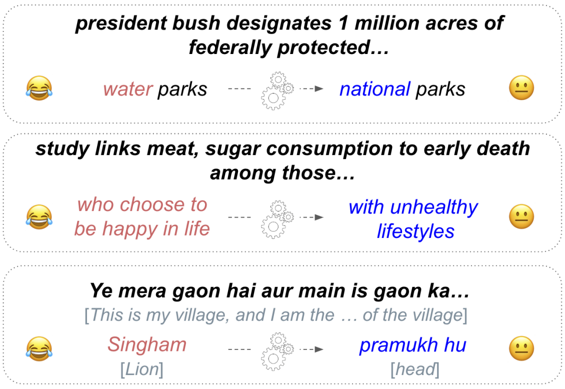
Getting Serious about Humor: Crafting Humor Datasets with Unfunny Large Language Models

Zachary Horvitz[1,*], **Jingru Chen**[1,*], **Rahul Aditya**[1], **Harshvardhan Srivastava**[1],
Robert West[2], **Zhou Yu**[1], **Kathleen McKeown**[1]
[1]Columbia University, [2]EPFL

{zfh2000, jc5898, ra3261, hs3447, zy2461}@columbia.edu
robert.west@epfl.ch, kathy@cs.columbia.edu

Abstract

Humor is a fundamental facet of human cognition and interaction. Yet, despite recent advances in natural language processing, humor detection remains a challenging task that is complicated by the scarcity of datasets that pair humorous texts with similar non-humorous counterparts. We investigate whether large language models (LLMs) can generate synthetic data for humor detection via editing texts. We benchmark LLMs on an existing human dataset and show that current LLMs display an impressive ability to "unfun" jokes, as judged by humans and as measured on the downstream task of humor detection. We extend our approach to a code-mixed English-Hindi humor dataset where we find that GPT-4's synthetic data is highly rated by bilingual annotators and provides challenging adversarial examples for humor classifiers.



president bush designates 1 million acres of federally protected...
water parks → national parks

study links meat, sugar consumption to early death among those...
who choose to be happy in life → with unhealthy lifestyles

Ye mera gaon hai aur main is gaon ka...
[This is my village, and I am the ... of the village]
Singham [Lion] → pramukh hu [head]

Figure 1: Outputs from GPT-4. We leverage language models to *edit away* (or "unfun") humor in existing human-written jokes, resulting in aligned datasets that pair humorous texts with non-humorous counterparts.

1 Introduction

Despite their success on natural language tasks, large language models (LLMs) struggle to reliably detect and explain humor (Baranov et al., 2023; Góes et al.; Hessel et al., 2023), and generate novel jokes (Jentzsch and Kersting, 2023). Notably, humans also struggle to write jokes; even at satirical newspapers like *The Onion*, less than 3% of proposed headlines are printed (West and Horvitz, 2019; Glass, 2008). In contrast, humans are able to consistently edit jokes to *unfun* them, an insight which motivated West and Horvitz (2019) to host a game where internet users competed to edit satirical headlines to make them serious. The resulting dataset, the *Unfun Corpus* (West and Horvitz, 2019), has been a valuable tool for advancing computational humor research. The dataset has been used to study properties of both humor and transformer architectures (West and Horvitz, 2019; Peyrard et al., 2021) and even to generate novel satire (Horvitz et al., 2020). Additionally, recent work has found that despite the relatively small size of the original dataset, humor detection models trained on Unfun data generalize remarkably well to other datasets, while models trained on other humor datasets perform poorly at classifying Unfun-edited data (Baranov et al., 2023).

While useful contributions, Unfun and other aligned humor datasets (Hossain et al., 2019, 2020) are limited in both size and scope, due to their reliance on human annotation. We investigate the alternative of using LLMs to create datasets of aligned humorous and non-humorous texts.[1] Previous work (Jentzsch and Kersting, 2023; Li et al., 2023; Veselovsky et al., 2023) has found that LLMs are limited in their ability to create synthetic humor. We take a new approach, exploiting the asymmetrical difficulty (Josifoski et al., 2023) of synthetic humor generation. Rather than only testing whether LLMs can *generate* humor, we explore their ability to *edit away* humor in existing jokes. Validating and harnessing this capability could provide large

---

[*] Equal contribution.

[1] Our code and datasets are available at `https://github.com/zacharyhorvitz/Getting-Serious-With-LLMs`.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 855–869
August 11-16, 2024 ©2024 Association for Computational Linguistics

paired datasets and support future work on improving humor detection and even generation.

Our contributions include benchmarking against human-curated data in the Unfun corpus, where we find that LLMs like GPT-4 and GPT-3.5 (OpenAI, 2023, 2022) can (1) outperform humans at removing humor from texts and that (2) this ability can be harnessed to generate high quality synthetic data for training humor classifiers. While these models *can* also be prompted to modify unfunny headlines to craft satire, we find that this ability is more inconsistent and lags behind satirical writers. Finally, we consider a code-mixed English-Hindi humor dataset to evaluate whether GPT-4's "unfunning" ability generalizes to other domains and languages. We find that the resulting synthetic unfunny dataset is rated highly by bilingual annotators and poses challenging adversarial data for models trained on the original corpus.

2 Getting Serious with Language Models

We first revisit the Unfun task and resulting dataset, but with language models as players.

2.1 Unfun Dataset

In the original Unfun game (West and Horvitz, 2019), players were tasked with editing existing satirical headlines from *The Onion*,[2] to transform the original satire into corresponding serious headlines. For example (removing "Delicious"):

"Scientists Discover ~~Delicious~~ New Species"

Players were rewarded for preserving token-level similarity with the original satire and for crafting convincingly serious headlines that other players rated as real. The resulting dataset includes approximately 11K unfunned headlines, with a subset rated by players. We leverage Unfun pairs, of satirical headlines and their unfunned counterparts, to benchmark the performance of LLMs at editing humorous texts against humans. We include additional details on data preparation in Appendix A.1.1.

2.2 Unfun Generation

We consider a few-shot setting (Brown et al., 2020), and provide LLMs with a short task description, along with a set of input-output exemplar pairs: *(humorous text, serious text)*. Following Veselovsky et al. (2023), we encourage diversity in our synthetic data by sampling these exemplars from a subset of the existing pairs rated as high-quality by the original human players. For the unfunning task, we consider four popular LLMs: GPT-4 (OpenAI, 2023) and GPT-3.5-TURBO, along with MISTRAL-7B-INSTRUCT and MISTRAL-7B (Jiang et al., 2023).

We also consider a lightweight alternative approach, ROBERTA-SWAP, that replaces low probability tokens using predictions from a ROBERTA masked language model (Liu et al., 2019). This approach is motivated by the Incongruity Theory of Humor (Hutcheson, 1750; Morreall, 2023), which associates humor with surprise, and previous work that has found humorous headlines to have higher perplexities (Peyrard et al., 2021). ROBERTA-SWAP edits satirical headlines by iteratively performing token swaps at k positions. At each selected position, the original token is replaced with the highest probability token predicted by the model at that masked time-step. The k swap positions are selected using the ratio between the probability of the original token and the probability assigned to the language model's prediction. Additional details on unfun generation are included in Appendix A.2.1.

3 Unfun Evaluation

3.1 Experimental Setup

The existing Unfun data enables comparison of human and LLM players, via both **automatic** and **human** evaluations. We first evaluate the quality of synthetically generated data through automated evaluation on the downstream task of Unfun detection, and then follow this with a human evaluation.

3.1.1 Automatic Evaluations

First, following recent work on synthetic data (Li et al., 2023; Veselovsky et al., 2023) we evaluate the data quality of outputs from LLMs by testing whether binary humor classifiers trained on the synthetic outputs can differentiate between actual humorous and unfunned headlines from the original Unfun dataset. We compare training on data from human players and actual satirical headlines to two configurations of synthetic data:

[*Synthetic* unfun; Original satire]
[Human unfun; *Synthetic* satire]

These two configurations enable comparing the "unfunning" and joke writing capabilities of LLMs. Additionally, we consider the alternative of using actual unrelated news headlines as non-humorous examples. Using data from each approach, we

---

[2] `https://www.theonion.com/`

		Data Characteristics		Holdout Accuracy	
Direction	Source	Diversity (TTR)	Edit Dist	Mistral	RoBERTa
Unfun	roberta-swap	0.262	2.7	69.9 (0.9)	62.7 (0.7)
	Mistral	0.257	**2.1**	70.7 (0.7)	61.7 (0.3)
	Mistral Instruct	0.255	2.4	70.9 (0.7)	64.7 (0.5)
	GPT-3.5	0.259	4.5	72.9 (0.2)	65.9 (0.4)
	GPT-4	0.252	3.8	76.5 (0.2)	69.9 (0.5)
	News Headlines	**0.306**	-	66.3 (0.2)	64.1 (0.2)
	Unfun Players	0.271	2.9	**80.3** (0.5)	**72.7** (0.4)
Humor	Mistral	0.244	2.8	66.3 (0.7)	56.3 (0.4)
	Mistral Instruct	0.221	4.5	65.2 (0.8)	58.8 (0.4)
	GPT-3.5	0.24	4.6	69.9 (0.5)	58.7 (0.4)
	GPT-4	0.246	5.5	69.5 (0.7)	59.7 (0.6)
	The Onion	0.262	-	-	-

Table 1: Automatic evaluations of synthetic Unfun data. We consider the two directions of editing away (**Unfun**) and editing in humor (**Humor**). We report median accuracies (and standard error) on a balanced holdout set ($n = 750$) over 5 seeds when fine-tuning Mistral (Jiang et al., 2023) and RoBERTa (Liu et al., 2019) humor classifiers.

Direction	Source	Rated Real	*Slightly* Funny / Funny	Grammatical	Coherence
Unfun	roberta-swap	30%	15% / 5%	93%	86%
	Mistral Instruct	21%	50% / 14%	**100%**	96%
	GPT-3.5	51%	23% / 3%	**100%**	98%
	GPT-4	49%	21% / 3%	**100%**	**99%**
	News Headlines	**81%**	**2% / 0%**	99%	93%
	Human Players	33%	21% / 7%	94%	92%
Humor	Mistral Instruct	21%	34% / 9%	99%	93%
	GPT-3.5	11%	54% / 8%	**100%**	94%
	GPT-4	10%	45% / 10%	**100%**	**98%**
	The Onion	**4%**	**68% / 24%**	99%	97%

Table 2: Human evaluations of synthetic Unfun data. We consider $n = 100$ samples per approach. We collect three annotations per example and assign labels by majority agreement.

fine-tune RoBERTa and Mistral-7B for humor classification. Our test set comprises a subset of headline pairs from the Unfun corpus that were highly rated in the original game. Additional evaluation details are provided in Appendix A.4.

3.1.2 Human evaluations

To perform our human evaluations, we recruited 10 university students as annotators, all of whom were American and native English speakers. Annotators were tasked with rating headlines as *real/satire/neither*. In the case of the "satire" label, we also task the annotators with rating *funniness* ([$0 =$ *not funny*, $1 =$ *slightly humorous*, $2 =$ *funny*]). If the annotator selects "neither", we ask them to rate the headline's *grammaticality* ($\{0, 1\}$) and *coherence* ($\{0, 1\}$). We gather three annotations for each sample and assign labels based on majority vote. We include additional information on our human evaluations and annotation scheme in Appendix A.3 and C.1

3.2 Results

Automatic Evaluations Table 1 contains the automatic evaluations on the Unfun corpus. Notably, when validated on human data, humor classifiers trained on GPT-4's synthetic unfun data are very performant, incurring the smallest accuracy drop relative to human-edited training data ($\Delta_{Mistral} = -3.8\%$ and $\Delta_{RoBERTa} = -2.8\%$). In contrast, classifiers trained with real news head-

Source	Edit Dist	Humor	Coherence
Non-Humor	-	16.8%	92.8%
GPT-4 Unfuns	6.6	16.0%	93.6%
+ GPT-4 Filter	6.9	3.6%	89.3%
Humor	-	48.0%	93.6%

Table 3: Human evaluations and edit distance of original and synthetic English-Hindi Tweet data (Khandelwal et al., 2018). $n = 125$ per approach.

lines as unfunny data perform poorly, highlighting the importance of aligned data for this task. However, we find that not all aligned data is created equal, and that classifiers perform significantly worse when trained on synthetic *humor* data relative to human-edited data ($\Delta < -10\%$). Even data from our ROBERTA-SWAP unfun baseline dramatically outperforms, or is on par with, all synthetic humor approaches. The edit distances demonstrate that each approach retains a large portion of the original humorous text. However, GPT-4 and GPT-3.5 tend to modify headlines more than human players (3.8 and 4.5 vs 2.9).

Human Evaluations Table 2 displays the results from our human evaluations. All approaches for generating synthetic humor significantly underperform *Onion* headlines on funniness and realness ratings ($p < 0.05$). Notably, we do not observe a significant improvement between GPT-3.5 and GPT-4. In contrast, synthetic unfuns from both GPT-3.5 and GPT-4 were significantly more likely than human unfuns to be rated as real news headlines. They were also rated as similarly unfunny and more grammatical and coherent. Surprisingly, our simple ROBERTA-SWAP approach also performed comparably with Unfun players on funniness and real headline metrics, but underperformed on coherence. Together, these results indicate that current LM-based methods underperform satirical writers on *humor generation*, but can outperform human crowd-workers at *editing away* humor in satire to craft aligned datasets.

4 Extending Unfun to Other Languages

Recent work has found that GPT-4 exhibits strong multilingual capabilities (Møller et al., 2023; Jiao et al., 2023; Ahuja et al., 2023). Motivated by these findings, we investigate whether its ability to edit away humor generalizes to other languages and forms of joke.

4.1 Experimental Setup

We consider an existing corpus of code-mixed English-Hindi tweets, previously annotated as humorous or non-humorous (Khandelwal et al., 2018). Here, we prompt GPT-4 to unfun humorous tweets. To remove low quality results, we secondarily filter outputs that GPT-4 still classifies as humorous. We provide additional details on dataset preparation in Appendix A.1.2 and English-Hindi unfun generation in A.2.

We perform a **human evaluation** with bilingual annotators who rated these unfunned outputs from GPT-4 alongside samples from the original dataset. We also run an **automatic evaluation**, testing the performance of humor classifiers trained with different proportions of synthetic non-humorous data. We evaluate on holdout synthetic data rated by the annotators as coherent and successfully non-humorous. For the humor classifier, we fine-tune an XLM-ROBERTA model (Conneau et al., 2020) previously fine-tuned on English-Hindi Twitter data (Nayak and Joshi, 2022).

4.2 Results

Tables 3 and 4 contain the human evaluations and automatic results for English-Hindi data. GPT-4 edited texts were rated comparably to non-humorous human tweets despite being derived from humorous tweets, which were rated as humorous by our annotators (48%) of the time. Filtering with GPT-4 yielded a smaller sample (56/125) that was rated as much less humorous (3.6%). These results demonstrate that GPT-4 is able to reliably unfun English-Hindi tweets, but with more edits than American satirical headlines (6.6 vs 3.8). Additionally, unfunned data can provide a challenging adversarial dataset. In Table 4 we evaluate the performance of humor classifiers on human-vetted unfunned data. When trained on the original dataset, the classifier fails to generalize to the unfunned samples and performs poorly (23% accuracy). Incorporating synthetic training data improves this metric at a cost to accuracy on humorous examples in the original dataset. Together, these results provide evidence that the humor classifier relies on superficial features to identify humorous text, and that, even with fine-tuning, the model struggles to recognize synthetic unfunny data.

		Original Dataset		
Source	Unfuns	Balanced Accuracy	Humor	Non-Humor
Original	22.6 (3.7)	67.9 (0.9)	80.3 (3.5)	56.9 (5.1)
(25%) Synth Unfuns	34.0 (8.4)	67.7 (1.7)	78.4 (3.3)	55.4 (5.9)
(50%) Synth Unfuns	57.7 (6.0)	62.1 (0.6)	68.4 (5.7)	55.9 (4.7)

Table 4: Automatic evaluations with English-Hindi synthetic data. We report median accuracies (and standard error) on a holdout set from the original dataset ($n = 591$) and the human-vetted unfuns ($n = 97$). We also report median class-level accuracies for the original dataset.

5 Discussion

Our results indicate that current LLMs struggle to generate humor, but can outperform crowd-workers at editing away (or *unfunning*) humor. We hypothesize that maximum likelihood training, combined with autoregressive sampling techniques, does not endow models with the creative spark required for joke writing, and instead lends itself to making high probability, reasonable substitutions to replace incongruous twists. Our evaluations on code-mixed English Hindi Twitter data indicate that, for GPT-4, this ability can impressively generalize to other languages and settings to create novel Unfun-like datasets. We are excited for future work that harnesses this capability and resulting data to improve humor detection and generation systems, and also to demystify fundamental properties of humor.

6 Limitations

We consider two settings, English satirical headlines and code-mixed English-Hindi tweets. Humor practices and references vary by culture (Alden et al., 1993; Jiang et al., 2019), and we leave investigating cultural impacts on LLMs and humor to future work. In both of our evaluations, the subjectivity of humor presents a challenge for our evaluations (Warren et al., 2021). We see evidence of this in Table 3, where only 48% of tweets previously annotated as humorous were also rated as humorous by our annotators, and where 16% of non-humorous tweets were rated as humorous. This likely reflects differences in background knowledge and context between annotators. Additionally, we note that human Unfun players were incentivized to perform minimal edits, which may have affected their human evaluation metrics and lowered edit distances. On average, however, GPT-4 performs less than one additional word edit, and several approaches, including ROBERTA-SWAP, were performant with lower edit distances than human players. Another concern is data contamination (Sainz et al., 2023), and that a portion of the text from the Unfun corpus could have been trained on and memorized by the LLMs we evaluated. We investigate this concern in Appendix A.6. We note that our results on English-Hindi data show that GPT-4's abilities generalize to a dataset where these pairs do not already exist on the internet.

7 Ethical Statement

Humor brings joy to people and plays a critical role in building and maintaining social relationships (Basso, 1979). However, its importance presents a double-edged sword; offensive and hurtful humor can cause real harms, and reinforce prejudice (Benatar, 1999). As a result, with their widespread adoption, it will be paramount for AI systems to be more capable of identifying and appropriately navigating jokes. We believe that our work on benchmarking LLM humor abilities and building challenging detection datasets is an important step in this direction. However, one possible concern is that malicious actors could leverage our *unfunning* approach to circumvent existing safeguards. In our experimentation, we found numerous settings where GPT-4 refused to generate jokes for offensive topics, but had no trouble editing texts to remove humor and offensiveness. This could enable building large parallel datasets of (offensive-text, non-offensive counterparts) that could then be used to train models for offensive joke generation.

Acknowledgements

We would like to thank Eric Horvitz for guidance that helped shape the direction of this work. We are also grateful to Nicholas Deas, Debasmita Bhattacharya, and Maximillian Chen for their feedback. Additionally, we would like to extend our gratitude to Amith Ananthram, Samir Gadre, Fei-Tzin Lee, Matthew Toles, Elsbeth Turcan, Melanie Subbiah, Emily Allaway, Tymon Nieduzak, Rattandeep Singh, Prabhpreet Singh Sodhi, and Apoorva Joshi for support on human evaluations.

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Krithika Ramesh, Samuel C. Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai. *ArXiv*, abs/2303.12528.

Dana L. Alden, Wayne D. Hoyer, and Chol Lee. 1993. Identifying global and culture-specific dimensions of humor in advertising: A multinational analysis. *Journal of Marketing*, 57:64 – 75.

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.

K.H. Basso. 1979. *Portraits of 'the Whiteman': Linguistic Play and Cultural Symbols among the Western Apache*. Cambridge University Press.

David Benatar. 1999. Prejudice in jest: When racial and gender humor harms. *Public Affairs Quarterly*, 13(2):191–203.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Ira Glass. 2008. Tough room.

Fabrício Góes, Piotr Sawicki, Marek Grze´s, Daniel Brown, and Marco Volpe. Is gpt-4 good enough to evaluate jokes?

Jack Hessel, Ana Marasovic, 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, Toronto, Canada. Association for Computational Linguistics.

Zachary Horvitz, Nam Do, and Michael L. Littman. 2020. Context-driven satirical news generation. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 40–50, Online. Association for Computational Linguistics.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. 2020. Stimulating creativity with funlines: A case study of humor generation in headlines.

F. Hutcheson. 1750. *Reflections Upon Laughter: And Remarks Upon the Fable of the Bees*. Garland Publishing.

Sophie Jentzsch and Kristian Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models.

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.

Tonglin Jiang, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore. Association for Computational Linguistics.

Ankush Khandelwal, Sahil Swami, Syed S. Akhtar, and Manish Shrivastava. 2018. Humor detection in english-hindi code-mixed social media content : Corpus and baseline system.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *ArXiv*, abs/2310.07849.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *ArXiv*, abs/2304.13861.

John Morreall. 2023. Philosophy of Humor. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Summer 2023 edition. Metaphysics Research Lab, Stanford University.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report.

Maxime Peyrard, Beatriz Borges, Kristina Gligorić, and Robert West. 2021. Laughing heads: Can transformers detect what makes a sentence funny?

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *ArXiv*, abs/2305.15041.

Caleb Warren, Adam Barsky, and A. Peter McGraw. 2021. What makes things funny? an integrative review of the antecedents of laughter and amusement. *Personality and Social Psychology Review*, 25(1):41–65. PMID: 33342368.

Robert West and Eric Horvitz. 2019. Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

A Appendix

A.1 Data Preparation

A.1.1 Unfun Corpus

We use the February 2, 2023 Unfun (West and Horvitz, 2019) database backup,[3] and consider all valid unfunned headlines (i.e. not *None*). This results in 11831 pairs. A subset of these have ratings from other players. We use these to curate a **high quality** evaluation subset of pairs where:

- There is at least one annotation.
- The satirical headline has a funniness rating ≥ 0.8.
- The unfunned headline has a funniness rating ≤ 0.2.

The resulting 867 pairs were split among prompt examples (10%), dev (30%), and test (60%) shards. For our training set, we consider the remaining headlines, again ensuring that there is no overlap with other shards. The resulting dataset has many instances where there are multiple unfunned counterparts for each satirical headline. As an additional step, we randomly filter our training, dev, and test shards so that there is only one unfunned headline per satirical headline. This results in a training set of 3882 unfuns, a dev set of 186 unfuns, and a test set of 375 unfuns, in each case, these are included alongside their corresponding satirical headlines. For an additional training data baseline, we also retrieve an equal number of real news headlines included in the Unfun database.

A.1.2 Code-Mixed English-Hindi Humor

We use the version of the English-Hindi Humor dataset by Khandelwal et al. (2018) hosted on GitHub.[4] We use the provided labels for the available data. Notably, a portion of annotated samples appear to be unavailable. We divide the available dataset ($n = 2951$) into training, dev, and test shards (60%, 20%, 20%). Additionally, we filter tweets containing links.

A.2 Data Generation Details

We include our full prompts in Appendix B. For decoding hyperparameters, we use *top-p* $= 0.85$ and $\tau = 1.0$ for all LLMs.

---

[3] `https://github.com/epfl-dlab/unfun`

[4] `https://github.com/Ankh2295/humor-detection-corpus`

A.2.1 Unfun Data Generation

To generate synthetic Unfun for each LLM approach, we prompt each model with 8 randomly sampled in-context pairs from examples from our high quality subset that was set aside for prompting. For our ROBERTA-SWAP baseline, we replace tokens in the original satirical headline using a ROBERTA-BASE[5] model. To select each replacement, we iterate over and individually mask each token in the headline, and then predict the masked token:

$$\hat{x}_i = \arg\max_{x} P(x \mid x_{\neq i}, \theta_{\text{RoBERTa}})$$

The position with the largest ratio between the predicted token and the original token probabilities is selected as the swap position:

$$\text{swap position} = \arg\max_{i}[\frac{P(\hat{x}_i \mid x_{\neq i}, \theta_{\text{RoBERTa}})}{P(x_i \mid x_{\neq i}, \theta_{\text{RoBERTa}})}]$$

We then replace x_i with $\hat{x}_i$, and repeat this procedure k times. We set $k = 3$ in our experiments.

A.2.2 Hindi-English Data Generation

Unlike for Unfun, we do not have existing pairs of (un-humorous, humorous) English Hindi tweets. To remedy this, we first generated 50 examples in a zero-shot setting on our training set, and then selected nine high quality results to serve as our prompt. We additionally prompt GPT-4 with humorous and non-humourous texts to classify the resulting unfunned tweets as humorous or non-humorous. We filter unfunned tweets if they are still classified as humorous.

A.3 Human Evaluations

We recruited 10 university students as annotators for the **Unfun task**. All annotators were American and native English speakers. For the **English-Hindi** dataset, we worked with three bilingual (Hindi and English) speakers. For both evaluations, we gathered three unique annotations per example, and assigned labels based on majority votes. Our Unfun evaluation assumes that any headline labeled as satirical or as real headline is grammatical and coherent. In contrast, we do not consider the grammatical label for English-Hindi data, due to the varied syntactic styles of tweets.

In Table 2, headlines are only rated "Real" if a majority of annotators rated the headline as "Real" (not "Satire" or "Neither"). Headlines are rated "Slightly Funny" if a majority of annotators assigned the headline *funniness* ≥ 1, and "Funny" with *funniness* $= 2$. Our full instructions for both human evaluations are included in Appendix C.1. Tables 5 and 6 display inter-annotator agreement statistics.

Human Label	Krippendorff
Real	0.507
Funny	0.333
Very Funny	0.214
Grammar	0.271
Coherence	0.214

Table 5: Krippendorff's α results on Unfun dataset.

Human Label	Krippendorff
Coherence	0.206
Humorous	0.377

Table 6: Krippendorff's α results on English-Hindi dataset.

A.4 Automatic Evaluations

On the **Unfun dataset**, for each synthetic Unfun approach, we generate data using the corresponding original 3882 training examples as inputs. We then evaluate classifiers trained on each dataset on the filtered high quality holdout data. To generate humor, we provide the unfunned example as input. To edit away humor, we provide the original satirical headline. We also provide in-context pairs drawn from the high quality prompt examples (See A.1.1). For our Real News baseline, we randomly select 3882 real news headlines to serve as non-humorous examples.

On the **English-Hindi dataset**, we compare training on the original dataset to training on data where (25%) and (50%) of non-humorous examples have been replaced by GPT-4 Filtered unfunned data. We evaluate classifiers on a holdout set from original dataset ($n = 591$), and also set of Unfuns ($n = 97$), derived from humorous examples in our holdout set and rated by our annotators as both coherent and non-humorous. All results for both datasets are computed over 5 seeds.

---

[5] `https://huggingface.co/FacebookAI/roberta-base`

A.5 Humor Classifier Training

For the Unfun task, we fine-tune MISTRAL (Jiang et al., 2023)[6] and ROBERTA (Liu et al., 2019)[7] models. For Hindi-English, we consider HING-ROBERTA (Nayak and Joshi, 2022)[8]. All models are trained with the AdamW optimizer (Loshchilov and Hutter, 2019) and a constant learning rate. Due to the class imbalance in the available English-Hindi dataset (39% non-humorous, 61% humorous), we weight the loss by the inverse proportion of class frequency.

We fine-tune our MISTRAL classifier with 4-bit quantized LoRA (Dettmers et al., 2023) and the addition of a classification head. For all classifiers, we first perform hyperparameter tuning on the original human authored datasets.

For the **Unfun dataset** we consider:

- Learning Rates $\in \{5e-5, 2.5e-5, 1.25e-5, 6.25e-6, 3.125e-6, 1.5625e-6\}$
- Batch Size $\in$ $[32]$ (Due to resource constraints)

For the **English-Hindi** Dataset dataset we consider:

- Learning Rates $\in \{5e-5, 2.5e-5, 1.25e-5, 6.25e-6, 3.125e-6, 1.5625e-6\}$
- Batch Size $\in \{256, 128, 64, 32, 16, 8\}$

After selecting the highest performing configuration, we run each experiment with 5 seeds ($[1234, 2345, 3456, 4567, 5678]$). We include the most performant hyperparameters in Table 7. All model trains use a single NVIDIA A100 GPU. We estimate the total compute budge to be 200 hours.

A.6 Considering Memorization

We investigate whether data contamination and memorization is affecting our results by testing how often synthetic unfuns or humor appear in the original Unfun corpus. We find that only a small fraction of outputs appear to match human-unfunned text or satire headlines. We include results in Table 8. Of these, the majority represent simple edits, indicating that the models may have rediscovered trivial unfuns. For example:

"Egypt plunges into state of ~~Middle East~~ *crisis*"

B Prompts

B.1 Unfun Task Prompts

B.1.1 Humor Generation

Chat Models

"You are a helpful assistant that edits realistic headlines to make them humorous."
{"role": "user", "content": <Unfunned Headline>},
{"role": "assistant", "content": <Satire Headline>}

Completion Models

"The following realistic headlines can be edited to be humorous:"
"<Unfunned Headline> -> <Satire Headline>"

B.1.2 Unfun Generation

Chat Models

"You are a helpful assistant that edits humorous headlines to make them realistic."
{"role": "user", "content": <Satire Headline>},
{"role": "assistant", "content": <Unfunned Headline>},
...

Completion Models

"The following humorous headlines can be edited to be realistic:"
"<Satire Headline> -> <Unfunned Headline>"

B.2 English-Hindi Task Prompts

B.2.1 Unfun Generation

Chat Models

"Kya ye diye hue tweet ka humor wala part hata kar use normal bana sakti ho? Aur jitna ho sake utna punctuation use same rakhne ki koshish karna" [Can you remove the humorous part of the given tweets and make them normal? And try to keep the punctuation as much the same as possible.].

---

[6] `https://huggingface.co/mistralai/Mistral-7B-v0.1`

[7] `https://huggingface.co/FacebookAI/roberta-base`

[8] `https://huggingface.co/l3cube-pune/hing-roberta`

Model	Learning Rate	Batch Size
Mistral (QLoRA)	6.25e-06	32
RoBERTa	1.25e-05	32
Hing-RoBERTa	1.5625e-06	8

Table 7: The training configurations for our automatic evaluations, after hyperparameter tuning.

Model	Unfun	Satire
GPT-3.5	3/200	0/200
GPT-4	7/200	0/200
Mistral	2/200	1/200
Mistral Instruct	2/200	0/200
roberta-swap	0/200	-

Table 8: The number of overlapping samples between human-curated headlines and synthetic headlines in our test examples ($n = 200$).

{"role": "user", "content": <Context Funny Tweet>},
{"role": "assistant", "content": <Context Unfunned Tweet>}

B.2.2 Unfun Filtering

Chat Models

"You are a pattern-following assistant used to rigorously determine whether a Hindi tweet is intended to be humorous. Given a Hindi tweet, respond only with either of Yes or No. Yes if it is humoruous and No if it is not humorous"
{"role": "user", "content": <Context Tweet>},
{"role": "assistant", "content": <Context Yes/No Label>}

C Human Evaluation Instructions

C.1 Unfun Task Instructions

Each annotator has been assigned a series of text samples to review. First, you are asked to evaluate whether the text sounds like a

- *r) real news headline (like from a non-humorous news website)*
- *OR s) satirical news headline (like from a humorous newspaper like TheOnion.)*
- *OR n) neither (text that would not appear in either setting, because it is ungrammatical, or incoherent.*

If you rate a headline as n (neither), you will be further prompted to rate it as a grammatical [no=0,yes=1 (for a news headline) and coherent [no=0,yes=1].
If you rate a headline as s (satire), you will be prompted to subjectively rate the quality of humor:

- *0 - not funny*
- *1 - slightly humorous / there is some identifiable joke*
- *2 - funny*

Content Warning: Several headlines may contain references to upsetting content.
EXAMPLES: **Satirical Headlines**

- nhl not quite sure why it has a preseason
- america's sweetheart dumps u.s. for some douchebag
- apple: new iphone good
- cat general says war on string may be unwinnable
- fire chief grants fireman 3-day extension on difficult fire

News Headlines

- the word 'doofuses' may cost ex-yahoo ceo bartz $10 million
- 2 meteorites hit connecticut

- world outraged by north korea's latest nuke test
- poverty rate hits 17-year high
- philippines: 5 foreign terror suspects in south

C.2 English-Hindi Task Instructions

The following task instructions specify additional information based on the original instructions provided to annotators in (Khandelwal et al., 2018).

Each annotator has been assigned a series of text samples to review. First, you are asked to evaluate whether the text is h) humorous n) non-humorous
Secondarily, you will be asked to rate whether a text is coherent [no=0,yes=1] A tweet should be marked as coherent, even if you don't have all the required background knowledge, as long as you can reasonably understand its meaning.
Additional info:

- *Any tweets stating any facts, news or reality should be classified as non-humorous.*
- *Tweets which consisted of any humorous anecdotes, fantasy, irony, jokes, insults should be annotated as humorous*
- *Tweets stating any facts, dialogues or speech which did not contain amusement should be put in non-humorous class.*
- *Tweets containing normal jokes and funny quotes should be placed in the humorous category.*
- *Some tweets consist of poems or lines of a song but modified. If such tweets contain satire or any humoristic features, then they could be categorized as humorous otherwise not.*

Content Warning: Several tweets may contain references to upsetting/offensive content.
EXAMPLES (We give the English Translations of each in brackets but they were not presented to the annotators):

Humorous Tweets

- Jhonka hawa ka aaj bhi chhup ke hilaata hoga na #Samir #HawaKaJhonka #BeingSalmanKhan [*Does the breeze still sway secretly today? #Samir #HawaKaJhonka #BeingSalmanKhan*)
- Working on a Sunday, chand rupye kamaane ke liye insaan apni khushiyon ka bhi sauda kar leta hai. [*Working on a Sunday, to earn a few rupees, a person sometimes even sacrifices their happiness.*]
- DJ wale babu bhosdike ab to gaana baja de iska.. bol bol ke kaan se khoon nikaal diya hai isne [*DJ wale babu, play the song now.. he has made our ears bleed by talking so much.*]
- Is Arvind Kejriwal new Che Guavara ? RT @ashutosh83B Is Rahul Gandhi new Arvind Kejariwal ? [*Is Arvind Kejriwal the new Che Guevara? RT @ashutosh83B Is Rahul Gandhi the new Arvind Kejriwal?*]
- Sukh bhare din beete re bhaiya, Babadook aayo re [*Brother, may the days filled with joy pass by. The Babadook has arrived.*]

Non-Humorous Tweets

- Apne support wale MLAs ko farmhouse main band kar lenge. Parade karayenge. Takhta palat karenge. Akhand chutiyap. [*We will lock up our supporting MLAs in the farmhouse. Parade them. Flip the throne. Absolute nonsense.*]
- Hrithik Roshan is using Vodafone. [*Hrithik Roshan is using Vodafone.*]
- PLEASE STOP MAKING JOKES ON SALMAN KHAN. BHAI BOLA NAHI CHALA RAHA THA GAADI TO NAHI CHALA RAHA THA. #BHAIROXX [*Please stop making jokes on Salman Khan. Bhai was not driving the car if he said he was not driving the car. #BHAIROXX*]

- Bhaari sankhya mein vote karein, aapke TL par wph hi nazar aayega [*Vote in large numbers, wph will only appear in your TL.*]

D Reference Examples

Tables 9, 10, and 11 include reference samples for English synthetic unfun outputs, English satire outputs, and English-Hindi unfun outputs respectively.

Original Satire	**tom petty to play some new stuff he's been working on at super bowl**	**jaguars offensive line not sure they can open big enough hole for maurice jones drew**	**obama takes surprise caller during weekly radio address**
Human	tom petty to play some new stuff he's been working on at coachella	jaguars offensive line not sure they can open stable positioning hole for maurice jones drew	obama takes caller during weekly radio address
GPT-3.5	tom petty to perform classic hits at super bowl	jaguars offensive line not sure they can create sufficient gap for maurice jones drew	obama takes surprise caller during live radio interview
GPT-4	tom petty to perform new material at super bowl	jaguars offensive line unsure if they can open big enough hole for maurice jones drew	obama takes unexpected caller during weekly radio address
MISTRAL	tom petty to play some new stuff he's been working on at superbowl	jaguars offensive line not sure they can open big enough hole for joe flacco	obama takes surprise caller during weekly radio address
MISTRAL INSTRUCT	tom petty to play some new songs he's been working on at super bowl halftime show	jaguars offensive line not sure they can clear big enough opening for maurice jones drew	obama takes surprise caller during weekly interview
ROBERTA-SWAP	he wants to play some new stuff he's been working on at the bowl	jaguars offensive line not sure they can dig big enough hole for maurice jakes	panama takes in migrants during weekly radio address

Table 9: Unfun model output reference samples.

Original Unfun	**obama, congress must strike a deal on budget by march 1**	**espn sports segment reveals science behind tony romo's precision pass**	**jimmie johnson disqualified after pre-race inspection reveals car to be violating rules**
Human	obama, congress must reach deal on budget by march 1, and then april 1, and then april 20, and then april 28, and then may 1	espn sports segment reveals science behind tony romo's shittiness	jimmie johnson disqualified after pre-race inspection reveals car to be airplane
GPT-3.5	obama, congress must strike a deal on budget by march 1 or face detention	espn sports segment reveals magic behind tony romo's precision pass	jimmie johnson disqualified after pre-race inspection reveals car to be a Transformer in disguise
GPT-4	obama, congress must strike a deal on budget by march 1 or risk being grounded	espn sports segment reveals magic spell behind tony romo's precision pass	jimmie johnson disqualified after pre-race inspection reveals car to be a spaceship
MISTRAL	obama, congress must strike a deal on budget by march 1	espn sports segment reveals science behind tony romo's precision penis	jimmie johnson disqualified after pre-race inspection reveals car to be black
MISTRAL INSTRUCT	obama, congress must strike a deal on budget by march 1, or else there will be no more kittens	espn sports segment reveals science behind kate middleton's royal wave	jimmie johnson disqualified after pre-race inspection reveals car to be violating laws and joking about it

Table 10: Satire model output reference samples.

Original Humorous Tweet	**GPT-4 English-Hindi Unfuns**
Ab ki baar.. #MaaBetaFarar..	Ab ki baar.. yeh log farar hain..
Husbands should be like Vim bar, gale kam aur chale zyada.	Patidev ko samarpit aur lambe samay tak saath dena chahiye.
O naadan parindey ghar aaja. Parinda: naadan tera baap.	O naadan parindey ghar aaja. Parinda: Mujhe ghar aane do.
Neend aaja nahi to kal se tujhe KRK bulaunga	Neend aaja nahi to kal se tujhe alag naam se bulaunga
Bhai ab itne velle bhi nahi hai ki #IndVsBan test match dekhenge	Bhai ab itne samay nahi hai ki #IndVsBan test match dekhenge
Asli toofan andar hai, jail ke andar. #SalmanVerdict	Asli samasya jail ke andar hai. #SalmanVerdict
Vodafone use karne se acha to ek kabootar pal lo.	Vodafone use karne se acha to kisi aur network provider ka use karo.

Table 11: GPT-4 English-Hindi unfunned reference samples. See Table 12 for English translations.

Original Humorous Tweet	**GPT-4 English-Hindi Unfuns**
This time.. #MotherSonGone..	This time.. these people are gone..
Husbands should be like Vim bar, less talk and more work.	Husbands should be dedicated and support for a long time.
Oh naive bird, come home. Bird: Your dad is naive.	Oh naive bird, come home. Bird: Let me come home.
If sleep doesn't come, from tomorrow I will call you KRK.	If sleep doesn't come, from tomorrow I will call you by a different name.
Bro, we're not that free to watch the #IndVsBan test match.	Bro, we don't have that much time to watch the #IndVsBan test match.
The real storm is inside, inside the jail. #SalmanVerdict	The real problem is inside the jail. #SalmanVerdict
It's better to raise a pigeon than to use Vodafone.	It's better to use another network provider than Vodafone.

Table 12: Translation of GPT-4 English-Hindi unfunned reference samples.