Dark & Stormy: Modeling Humor in the Worst Sentences Ever Written

Venkata S Govindarajan
Department of Computer Science,
Ithaca College
vgovindarajan@ithaca.edu

Laura Biester
Department of Computer Science,
Middlebury College
lbiester@middlebury.edu

Abstract

Textual humor is enormously diverse and computational studies need to account for this range, including **intentionally bad humor**. In this paper, we curate and analyze a novel corpus of sentences from the Bulwer-Lytton Fiction Contest to better understand "bad" humor in English. Standard humor detection models perform poorly on our corpus, and an analysis of literary devices finds that these sentences combine features common in existing humor datasets (e.g., puns, irony) with metaphor, metafiction and simile. LLMs prompted to synthesize contest-style sentences imitate the form but exaggerate the effect by over-using certain literary devices, and including far more novel adjective-noun bigrams than human writers.

1 Introduction

Humor is a uniquely human trait, and its interpretation involves language and social understanding, creativity, and common-sense reasoning (Brock, 2017; Martin and Ford, 2018; Attardo, 2020), making it an attractive challenge for Computational Linguistics. Prior work has explored a range of humor-related tasks, including humor detection (Mihalcea and Strapparava, 2006), generation (Hessel et al., 2023), and even editing (Horvitz et al., 2024). However, most available humor datasets focus on familiar forms such as satirical headlines, puns, or short one-liners, leaving other varieties underexplored. In this paper, we introduce a novel dataset for an understudied form of textual humor — intentionally bad humor — through sentences drawn from the **Bulwer-Lytton Fiction Contest**[1] (BLFC).

The BLFC was an annual competition (1982–2024) founded by Professor Scott Rice at San Jose State University, later co-run with his daughter Elizabeth J. Rice. Inspired by the notorious opening line of Edward Bulwer-Lytton's 1830 English novel *Paul Clifford*, it invited participants to craft "an atrocious opening sentence to the worst novel ever written." Each year, the organizers selected 50–60 entries to highlight from thousands of submissions. Figure 1 is an example entry, with dominant literary devices highlighted (see Appendix A for more).

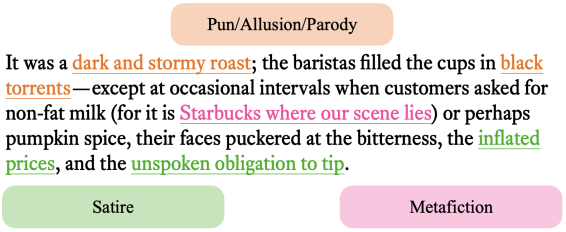

Figure 1: An entry from the Bulwer-Lytton contest, with the key literary devices highlighted.

In this paper, we build and analyze a unique resource focused exclusively on intentionally bad humor with selected entries from the BLFC. Our goal is twofold: (1) to show that the form of humor in these sentences differs significantly from standard humor datasets, and (2) to examine how Large Language Model (LLM) generated Bulwer-Lytton sentences compare with human-written ones.

We find that Bulwer-Lytton sentences differ sharply from standard humor datasets. Humor detection models perform poorly on them; they make greater use of literary devices like metaphor, metafiction and simile; and they contain many more *semantically deviant* (novel) adjective-noun expressions. Moreover, while LLM-generated sentences can imitate the form of Bulwer-Lytton humor, they consistently exaggerate the effect across all our analyses. We share our data and code online at github.com/venkatasg/bulwer-lytton.

2 Data

Bulwer-Lytton sentences With permission from the contest organizers, we scrapped all entries on the contest website between 1996 and 2024. The

[1] https://www.bulwer-lytton.com/

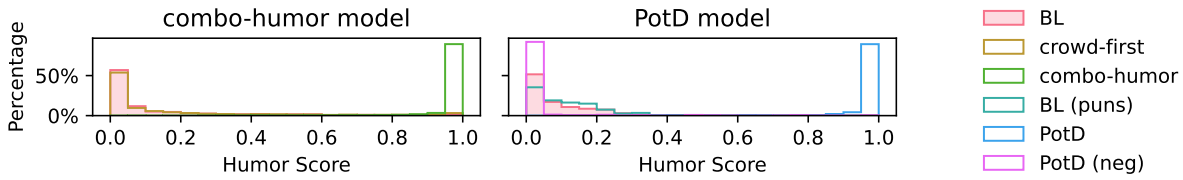

Figure 2: Comparing humor detected in BL sentences to crowd-first/combo-humor with the combined humor model (L). Comparing humor detected in BL sentences with the pun subset of BL/PotD with the pun model (R).

organizers list entries for every year by genre, and also list a grand prize winner, runner-up, and other ad-hoc categories — meta-data that we also collect and associate with each entry. Thus we compile the Bulwer-Lytton (henceforth abbreviated to **BL**) dataset comprising 1778 sentences over 29 years.

Synthetic BL sentences While LLM abilities in humor understanding, generation and manipulation have been uneven (Hessel et al., 2023; Horvitz et al., 2024; Cocchieri et al., 2025), they excel at instruction following (Ouyang et al., 2022) and *imitating the form* of obscure textual forms (Reif et al., 2022). In this work, we conduct a preliminary comparison of LLM-generated and human-written BL sentences to understand how prompting captures—or distorts—the style of intentionally bad humor.

We use a minimal one-shot prompt (see Appendix C) describing the contest and its rules taken from the website, along with the original BL sentence. We generate and analyze 1000 BL sentences from `DeepSeek-V3.1` (DeepSeek-AI, 2024) and OpenAI's `gpt-5-2025-08-07` (OpenAI, 2025c) in this paper. We also present results from `gpt-4.1-2025-04-14` (OpenAI, 2025b) and `gpt-oss-120b` (OpenAI, 2025a) in the appendix.

Baseline datasets We aim to capture how BL sentences differ from two other types of texts: first sentences to novels and texts used in other studies of humor. We study first sentences from a crowd-sourced (henceforth referred to as **crowd-first**) list of first sentences to novels.[2] They range from classics (e.g., Jane Austen, Charles Dickens) to modern bestsellers (e.g., Terry Pratchett, Nora Roberts) to sentences submitted by their amateur creators.

As BL sentences exemplify a genre of humor that has been understudied in NLP, we also want to see how they differ from texts that exemplify other genres of humor. We draw from humor datasets compiled by Baranov et al. (2023), and focus on the combined dataset (henceforth referred to as **combo-humor**) which is a diverse collection of different types of humor. It incorporates puns, satire, ShortJokes from humorous Reddit posts (Chen and Soo, 2018) and jokes from Twitter/ShortJokes from Meaney et al. (2021). When considering pun-specific models and BL sentences labeled "Vile Puns," we incorporate Pun of the Day (henceforth referred to ask **PotD**), jokes which were collected by Yang et al. (2015). We focus on the positive instances from the test split unless otherwise stated.

Table 1 in Appendix B shows that BL sentences (human-written and synthetic) **are substantially longer than typical jokes or novel openings**. We explore how this verbosity contributes to their distinctive form of "bad humor" in following sections.

3 Preliminary Analysis

3.1 Humor Detection

In this section, we examine whether BL sentences are recognized as humorous by existing humor detection models. We use the best performing RoBERTa-based model trained by Baranov et al. (2023) on combined humor datasets, as well as their model to detect puns.

We analyze humor by plotting a histogram of the continuous scores (0–1) from the models for each dataset, using the mean over 5 random seeds. We find that BL sentences are not identified as humorous; Using the combined model, the humor scores are in line with crowd-first (Figure 2, left); results were similar on synthetic BL sentences. Meanwhile, the in-domain sentences from combo-humor (the test split of the dataset the model was trained on) have very high humor scores.

We also consider whether models can recognize specific types of humor (puns) in out-of-domain text, by using a humor detection model trained on PotD and analyzing a subset of the BL sentences that were labeled with the genre "Vile Puns." We find that BL sentences have slightly lower scores than the subset of the sentences that are identified as puns (Figure 2, right). However, while both have

[2] https://github.com/janelleshane/novel-first-lines-dataset

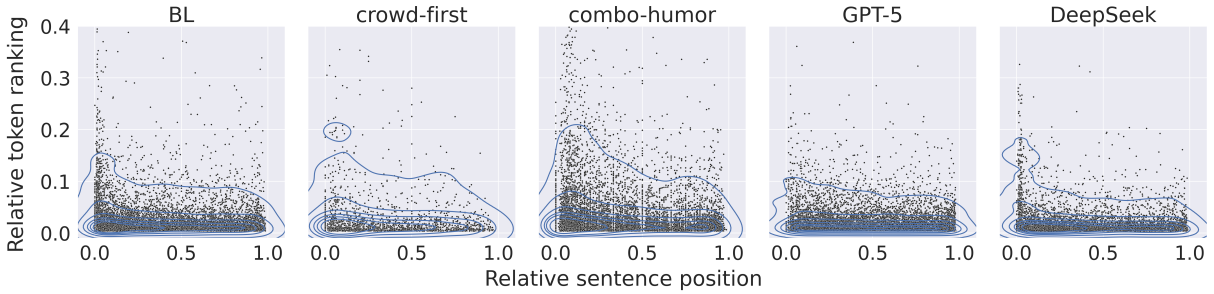

Figure 3: Relative rank of high-surprisal tokens plotted against their relative sentence position for our 5 datasets.

scores that on average exceed the scores on the negative examples from the PotD test set, they are nowhere near the scores on the positive examples from the in-domain PotD data. Closer inspection reveals that puns are often *one of multiple features* that are employed to make a BL sentence humorous, whereas in the PoTD dataset, the examples are typically short sentences which are only funny due to a pun. Our dataset **demonstrates a failure of pun detection models trained on short jokes to recognize the use of puns in a longer context**.

3.2 Literary Devices

To further analyze how BL sentences differ from standard humor datasets, we automatically extract sets of literary devices that are used in sentences using a prompt-based framework adapted from TopicGPT (Pham et al., 2024). Compared to traditional topic modeling algorithms such as LDA (Blei et al., 2003), we find that this method is able to extract higher-level features capturing the *style* of writing rather than content. We added humor-related seed topics and rework the prompts to refer generically to "features" rather than "topics" to better fit the type of label we hope to extract. We execute the generation and alignment processes with GPT-4.1 using 300 sentences with 100 sampled from each of the BL dataset, the combo-humor dataset (to capture features related to humor), and the crowd-first dataset (to capture features common in first sentences). This process results in a set of eight features, and we validate that features aligned with the text using an intruder task (see Appendix D).

After extracting features, we run the assignment framework on all BL sentences (including synthetic) and up to 1000 randomly sampled instances from each baseline dataset. Figure 4 demonstrates how the literary devices in the BL dataset differ from the other datasets. We find that BL sentences clearly exceed the baselines in Irony, Metafiction,

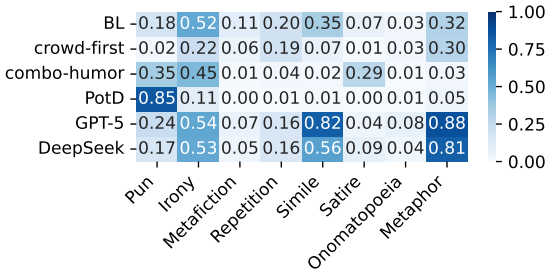

Figure 4: Literary device presence across datasets.

and Simile. They fall between the crowdsourced first sentences and existing humor datasets in their use of Puns and Satire. Furthermore, we observe a tendency of the synthetic BL sentences to overuse certain literary devices including Simile, Metaphor, and to a lesser extent Onomatopoeia.

4 Incongruity in BL sentences

4.1 Surprisal

The incongruity and semantic script theories of humor (Raskin, 1979; Forabosco, 1992) describe humor as arising from violated expectations followed by reinterpretation. Building on this view, computational studies have modeled incongruity through information-theoretic measures like *surprisal*. Humorous headlines tend to have higher perplexities (Peyrard et al., 2021), and swapping low-probability (high surprisal) tokens with more predictable ones can remove humor (Horvitz et al., 2024). West and Horvitz (2019) further found that humor tends to reside in the later part of satirical headlines, aligning with the idea that humor emerges from late-stage expectation violation. In this section, we analyze the distribution of surprisal across sentence positions to test whether **Bulwer-Lytton sentences exhibit different patterns of incongruity** from standard humor datasets.

High surprisal tokens We extract the log-probabilities for each token in our datasets us-

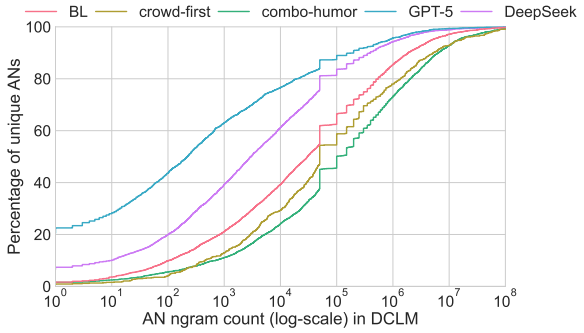

Figure 5: Cumulative distribution of percentage of ANs within a dataset against their count in the DCLM corpus.

ing the `minicons` package (Misra, 2022) and the `gemma-3-1b-pt` LM (Team Gemma et al., 2025). We define high surprisal tokens as those outside the top 1000 tokens ranked by probability as predicted by the LM, and plot their relative rank (predicted rank normalized by the vocabulary size) over relative sentence position for that token in its sentence in Figure 3. We sample 1000 sentences from crowd-first and use the entirety of all other datasets.

High-surprisal tokens from BL sentences differ in 2 ways from sentences in combo-humor. Firstly, high-surprisal tokens from combo-humor occur more often at the beginning half of the sentence, with a small peak at the end; BL sentences have a much flatter distribution of surprising tokens. Secondly, BL sentences have **more surprising tokens per sentence** (5.1 ± 3.1) than combo-humor (2.0 ± 1.4). Both GPT-5 and DeepSeek inflate the number of high-surprisal tokens, but curiously DeepSeek (6.2 ± 2.6) does this less than GPT-5 (9.2 ± 2.8). Figure 3 shows that DeepSeek's distribution aligns much more closely with that of human-written BL sentences than with GPT-5, which is relatively flat. The difference in behavior of these two LLMs adds to contemporaneous work demonstrating how different LLMs exhibit significant differences in behavior on the same task/prompt due to their distinct training pipelines (Shao et al., 2025).

4.2 Novel Adjective-Noun expressions

Jarring and novel adjective–noun combinations disrupt semantic expectations and draw attention to the prose itself (Westbury and Hollis, 2019), heightening the bad humor exemplified by BL sentences. Inspired by Vecchi et al.'s (2017) work showing that the 'deviance' of novel adjective-noun pairs can be predicted by their distributional occurence a corpus, we analyze the distribution of Adjective-Noun (AN) bigrams across all our datasets and their occurrence in a large pre-training corpus of natural language.

AN counts As in Vecchi et al. (2017), 'deviant' or novel ANs are identified as ones that have a low count in a natural language corpus. We expect BL sentences to contain more novel AN bigrams than our baseline datasets. We extract all ANs from all our datasets (we sampled 1000 sentences for crowd-first) using Stanza (Qi et al., 2020) and query their count in the DCLM pre-training corpus (Li et al., 2024) using Infini-gram (Liu et al., 2024). BL sentences have more ANs per sentence on average (2.7) than crowd-first (0.77) or combo-humor (0.55), highlighting how multiple ANs are a unique feature for BL sentences. GPT-5 BL sentences had 2.56 ANs per sentence and DeepSeek sentences had 5.3, once again showing a curious difference in behavior from different LLMs to the same prompt that we also observed in §4.1.

Figure 5 plots the cumulative percentage of all unique ANs in each dataset as a function of their (log-scaled) frequency in the DCLM corpus, allowing us to see, for any frequency threshold, what share of ANs occur at or below that count. As expected, BL sentences have many more novel ANs per sentence than either of our baseline datasets — 10% of BL ANs occur less than 100 times in DCLM compared to 4% and 6% for crowd-first and combo-humor.

Furthermore, Figure 5 (and Figure 8 in Appendix E) show that synthetic BL sentences have even more novel ANs than the human-written BL sentences. Together with our analysis in previous sections, this paints a consistent picture of the synthetic BL sentences **imitating the form of human-written BL sentences while consistently exaggerating its notable features**. Table 2 in Appendix E contains examples of novel ANs from our datasets.

5 Conclusion

We introduce a new dataset of intentionally bad humor with texts from the Bulwer-Lytton Fiction Contest. Humor detection models fail to identify these sentences as humorous, and literary theme analysis further reveals how devices like metaphor, repetition and metafiction differentiate the humor in this dataset. Analysis of synthetic BL sentences reveals that they imitate the form of BL sentences, but exaggerate several of their key features.

Limitations

Data and focus Our work is a preliminary investigation into 'bad' humor as exemplified by entries in the Bulwer Lytton Fiction Contest, and we wish to motivate how our novel corpus differs from existing humor datasets. Intentionally bad humor can occur in many forms — sentences in the BLFC are simply one form of that. Whether or not a sentence constitutes bad humor or a good entry in the BLFC is an inherently subjective question. We leave the quantification and further analysis of the boundaries of when humor is 'bad' or 'good' to future work, and rely on the subjective choices by the organizers of the BLFC for what constitutes a BL sentence.

Humor Detection Our experiments use a limited number of humor detection models; it is possible that other models would perform better on out-of-domain data. However, we believe that the complexity of BL sentences and their tendency to encapsulate multiple forms of humor means that other models likely would also struggle to quantify their humor.

Literary Device Analysis We focus on only eight key literary devices identified using TopicGPT with GPT-4.1. In preliminary experiments, we found that other LLMs identified additional literary devices such as personification, but also identified false positives (e.g., the inclusion of "Craigslist" in the sentence). Following a pilot of the intruder task discussed in Appendix D, we decided that the literary devices generated and assigned using GPT-4.1 gave us the best overview of the dataset while excluding noisy features. Despite this, a small number of sentences were labeled by GPT-4.1 in a way that could not be correctly parsed during the assignment phase. In some cases, a feature was assigned with the qualifier ("not assigned") in the description of why it was assigned, and in other cases features were assigned that did not actually exist in our set of eight literary devices. This introduces a small amount of noise, but our feature validation process shows that the labels are meaningful despite this.

Synthetic BL sentences Our work focuses on understanding how LLMs generate and mimic the style of BLFC entries with minimal prompting — hence the one-shot example with simple instructions from the rules on the website. We focus on how these sentences differ stylistically from the human-written BL sentences in this paper, and leave evaluation of whether the generated sentences are good entries for the BLFC to future work.

Surprisal While we analyze the incongruity in BL sentences through token probabilities, this doesn't imply *directly* that the humor resides in regions of high surprisal. This would require human annotation and further analysis (as in West and Horvitz (2019)) which goes beyond the scope of this short paper. Our goal with the the analysis in this paper is to motivate that the distribution (and number) of low-probability tokens is clearly different between BL sentences and standard sentences of sentential humor, as well as between human-written and synthetic BL sentences.

Acknowledgments

This research is partially supported by start-up funds and computational resources provided by Ithaca College and Middlebury College. We would also like to thank Kasia Bartoszynska and Hugh Egan in the English department at Ithaca College for helpful feedback and discussions in the early stages of this work.

References

Salvatore Attardo. 2020. *The Linguistics of Humor: An Introduction*. Oxford University Press.

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.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Alexander Brock. 2017. Modelling the complexity of humour – Insights from linguistics. *Lingua*, 197:5–15.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in neural information processing systems*, 22.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor Recognition Using Deep Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

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.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Giovannantonio Forabosco. 1992. *Cognitive aspects of the humor process: The concept of incongruity*. Walter de Gruyter, Berlin/New York Berlin, New York.

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, 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.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 14200–14282. Curran Associates, Inc.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. In *First Conference on Language Modeling*.

Rod A Martin and Thomas Ford. 2018. *The Psychology of Humor: An Integrative Approach*. Academic press.

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.

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

Kanishka Misra. 2022. `minicons`: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

OpenAI. 2025a. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI. 2025b. Introducing GPT-4.1 in the API.

OpenAI. 2025c. Introducing GPT-5.

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.

Maxime Peyrard, Beatriz Borges, Kristina Gligorić, and Robert West. 2021. Laughing Heads: Can Transformers Detect What Makes a Sentence Funny? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3899–3905. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A Prompt-based Topic Modeling Framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A Recipe for Arbitrary Text Style Transfer with Large Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. 2025. Spurious Rewards: Rethinking Training Signals in RLVR. *Preprint*, arXiv:2506.10947.

Team Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2017. Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces. *Cognitive science*, 41(1):102–136.

Robert West and Eric Horvitz. 2019. Reverse-Engineering Satire, or "Paper on Computational Humor Accepted despite Making Serious Advances". In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 7265–7272.

Chris Westbury and Geoff Hollis. 2019. Wriggly, Squiffy, Lummox, and Boobs: What Makes Some Words Funny? *Journal of Experimental Psychology: General*, 148(1):97.

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.

A Example BL sentences

The opening sentence from the novel *Paul Clifford* by Edward Bulwer-Lytton that inspired the contest:

(1) It was a dark and stormy night; the rain fell in torrents—except at occasional intervals, when it was checked by a violent gust of wind which swept up the streets (for it is in London that our scene lies), rattling along the housetops, and fiercely agitating the scanty flame of the lamps that struggled against the darkness.

Below, we list additional examples from the Bulwer-Lytton fiction contest:

(2) a. It was a dark and stormy roast; the baristas filled the cups in black torrents—except at occasional intervals when customers asked for non-fat milk (for it is Starbucks where our scene lies) or perhaps pumpkin spice, their faces puckered at the bitterness, the inflated prices, and the unspoken obligation to tip.

b. She had a body that reached out and slapped my face like a five-pound ham-hock tossed from a speeding truck.

c. Space Fleet Commander Brad Brad sat in silence, surrounded by a slowly dissipating cloud of smoke, maintaining the same forlorn frown that had been fixed upon his face since he'd accidentally destroyed the phenomenon known as time, thirteen inches ago.

d. Having committed to memory the route maps leading from Alaska to L.A. and determined to avenge the more than 300 years of ridicule his people had suffered for the popularly assumed whimsy of Eskimo naming conventions, Robson Pollowow ("He who moves mountains") glanced one last time into his hastily rigged rear view mirror before releasing the hand brake on the Paw Whal Dee Glacier.

B Dataset statistics

Dataset	Avg. token len (S.D.)	# Instances
BL	70.5 (23.1)	1778
crowd-first	24.5 (30.8)	7901
combo-humor	19.0 (8.9)	5416
PotD	14.3 (5.1)	493
GPT-5	73.7 (6.7)	1000
DeepSeek	94.1 (26.6)	1000
GPT-4.1	81.1 (15.5)	1000
`gpt-oss-120b`	95.4 (27.7)	1000

Table 1: Average token length (with standard deviation in parentheses) per sentence in each of our datasets. Sentences tokenized with `gemma-3-1b-pt`.

C One-shot prompt

Prompt for generating BL sentences

The Bulwer-Lytton Fiction Contest challenges participants to write an atrocious opening sentence to the worst novel never written. The whimsical literary competition honors the novelist Sir Edward George Bulwer-Lytton and his marvelously awful opening to his 1830 novel Paul Clifford:

"It was a dark and stormy night; the rain fell in torrents—except at occasional intervals, when it was checked by a violent gust of wind which swept up the streets (for it is in London that our scene lies), rattling along the housetops, and fiercely agitating the scanty flame of the lamps that struggled against the darkness."

The rules for the Bulwer Lytton Fiction Contest are childishly simple:

- Each entry must consist of a single sentence.
- Sentences may be of any length but we strongly recommend that entries not go beyond 50 or 60 words.
- Entries must be "original" (as it were) and previously unpublished.

Based upon these instructions and the original example, your goal is to write the most atrocious opening sentence to the worst novel ever written in the following genre: {genre}. Your final output should contain only the sentence, with no other text or explanations.

We sample from the following genres which were the most popular in the human-written BL dataset: Adventure, Science Fiction (100 each), Purple Prose, Romance, Crime & Detective, Vile Puns (150 each), Western, Historical Fiction, Children's

Literature, and Fantasy & Horror (50 each). The numbers for each genre were chosen as they closely followed the relative proportion of sentences from each genre in the human-written BL corpus.

We generate samples from GPT-4.1 and GPT-5 using the OpenAI API (`https://platform.openai.com`), and DeepSeek using the Together API (`https://api.together.ai`). Samples from `gpt-oss-120b` were generated locally on a workstation with 2 Nvidia RTX 6000 Ada GPUs. All responses were generated with temperature set to 1 and with thinking set to off (or low).

D Literary Devices

Figure 6 shows the literary device breakdown by category, focusing on those categories linked to at least 50 entries–for instance, 80% of entries categorized as "Vile Puns" were assigned the Puns feature and 84% of entries categorized as "Purple Prose" were assigned the Simile feature. Figure 7 shows the literary device breakdown for two additional models used to generate synthetic BL sentences, GPT-4.1 and gpt-oss 120b. We find that like with the other synthetic sentences, simile, metaphor, and onomatopoeia are exaggerated.

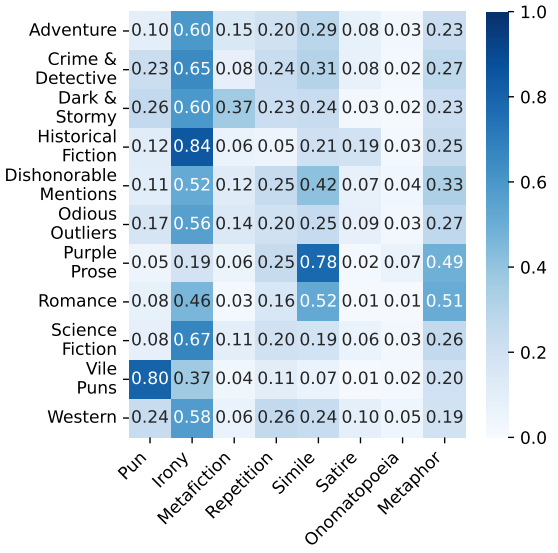

Figure 6: Literary device presence in BL sentences from different categories.

Feature Validation We conduct an intruder task similar to the task introduced by Chang et al. (2009) to evaluate the quality of the feature evaluations. Human annotators are shown a feature, its description, and five BL sentences. One sentence was not assigned the feature and the other four were; the annotator must determine which is the "intruder."

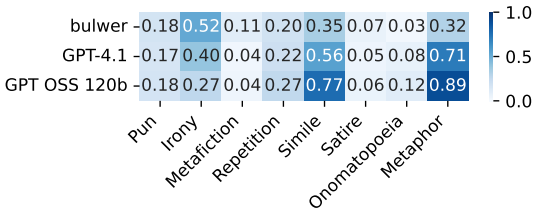

Figure 7: Literary device presence in human-written BL sentences compared to those generated by GPT-4.1 and gpt-oss 120b.

Instructions Given to Annotators

Background
We are working on a project to characterize the type of humor in winners of the Bulwer-Lytton fiction contest and how it differs from other humor datasets. We have extracted a set of literary features for each sentence. The task is intended to determine the quality of the extracted features.

Your Task
You are going to see a page with a question (each question is for a specific literary feature) and five examples labeled A-E. In each case, there are going to be four positive examples that were labeled with that feature by a model and one negative example that was not labeled with that feature. For each question, you should select the answer corresponding to the example that you think is least representative of the feature (so you're identifying the negative example). This is a situation where you just have to do your best. It's possible there will be multiple examples that don't fit well or that all of the examples will fit (which indicates the model isn't doing a good job) so select whichever makes the most sense to you! Read over the sentences carefully, as they are fairly complex.

We created six instances for each of the eight features, for a total of 48 unique instances. Six annotators were randomly assigned 16 instances each, and each instance was annotated by two annotators.[3] One would expect 20% accuracy if the feature assignments did not match the text and annotators always guessed randomly. Among the annotators, accuracy ranged from $\frac{7}{16}$–$\frac{12}{16}$, with an average of 57% of instances labeled correctly.

Using Krippendorff's alpha, we found that annotators had fairly low agreement (0.51). This is unsurprising given that the sentences are fairly complex and difficult to parse. However, we believe that the accuracy indicates the usefulness of the features despite the difficulty of the annotation task.

[3] All annotators were undergraduate students and were given a $15 gift card for approximately one hour of work.

E Deviant ANs

When querying the count for ANs against the Infinigram API, we include a disjunction over lowercased and capitalized versions of the AN expression. Figure 8 plots the cumulative percentage of all unique ANs in the synthetic and human-written BL sentences as a function of their (log-scaled) frequency in the DCLM corpus: GPT-4 and DeepSeek exhibit similar distributions, while `gpt-oss-120b` and GPT5 differ significantly as they generate many more unattested ANs. Table 2 shows some rare ANs from each of our datasets.

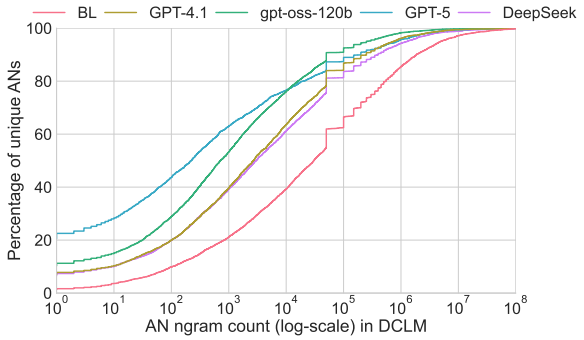

Figure 8: Cumulative distribution of percentage of ANs within a dataset against their count in the DCLM corpus.

BL	ocherous brilliance, albino apricots, helpless patrolman, indigent zucchini, antediluvian passion, crepuscular finish
combo-humor	underdressed layman, unfaithful kermit, nihilistic pencil, broke stroker, teenage fireflies, empathetic bovine
crowd-first	ghastly armoire, legged lupine, mighty arrowhead, brassy tocsin, mean loudmouths
GPT-5	brocadeed accordion, tragic omelet, moonless noon, anti-gravity mustache, overripe chandelier
GPT-4.1	cosmic cheeseboard, narcoleptic willows, dimest shadows, filmy crinolines, deafening falcon, gelatinous moons
DeepSeek	dubious grog, ghostly quadrille, enchanted paperweight, assertive sculpting, feisty seamstress, untrustworthy camels
gpt-oss-120b	emberkissed moths, amber–wreathed memories, unholy expedition, syncopated agony, soggy rumors, olated secrets

Table 2: Adjective-noun bigrams from our datasets that occur fewer than 10 times in the DCLM corpus.