THE HUMOROUS HILL: A U-SHAPED RELATIONSHIP BETWEEN HUMOR AND QUALITY

A PREPRINT

Antara Raaghavi Bhattacharya[*] **Jennifer Hu**[†] **Tomer D. Ullman**[‡]

ABSTRACT

We propose and examine a "Humorous Hill" pattern. The pattern suggests that the output of learning systems will show a U-shaped relationship between humor and quality. We tested the Humorous Hill by using language models to create novel examples of open-ended categories. The language models varied in quality and architecture. Human participants (N = 450) rated the output of the language models for quality (category membership, goodness of fit), and humor. Across models and categories, we found an inverted U-shaped relationship between humor and accuracy. We evaluate the pattern against existing theories of humor and suggest that many do not obviously account for it without modification. We propose that the Humorous Hill applies generally to learning dynamics, including those found in children.

Keywords Humor | Language models | Artificial intelligence | Categories

1 Introduction

Humor is an important part of everyday life. Despite centuries of study, there is no agreed-on explanation for what makes some things funny [Hurley et al., 2013, Warren et al., 2021]. The lack of agreement is not due to a lack of possible explanations, and much-discussed theories include incongruity [Deckers and Kizer, 1975], relief [Spencer, 1875], superiority [Gruner, 1978, Berger, 1993, Martin and Ford, 2018], benign violation [McGraw and Warren, 2010], and play [Van Hooff, 1972]. Here, we highlight and examine a robust pattern within humor, not discussed or explained by previous major theories. We call this pattern 'The Humorous Hill'. This pattern emerges when a learning agent is attempting to produce outputs to match some given idea, concept, or template. The agent is not trying to be funny, but on the way to being accurate, it inadvertently is funny.

Before detailing the hypothesis and experiments, it helps to have examples of the phenomenon. Consider for example a recent predictive algorithm trained on the Harry Potter books, and tasked with producing similar work. The result was *Harry Potter and the Portrait of what Looked like a Large Pile of Ash* [Botnik Studios, 2018]. Or consider the work of Janelle Shane, who trains recurrent neural networks on various concepts such as recipes, resulting in recipes like "Chocolate Chocolate Chocolate Cake" and "Completely Meat Circle" [Shane, 2021]. The models are not trying to produce funny outputs, but the outputs are funny. It seems likely that much of the engagement with such artificial systems is not because they are good at what they do, but because they are bad in a particular kind of way.

With the examples in mind, we propose the *Humorous Hill* pattern: the output of learning systems will show a U-shaped relationship between humor and quality. When model outputs are gibberish, they are simply inaccurate rather than funny ('Harry Potter and the zfya sgdgx' is not funny, except possibly as anti-humor). When model outputs match the target, they are simply accurate rather than funny ('Harry Potter and the Vault of Whispers' is simply plausible). However, on the way from being bad to good, outputs will be funny, and they will be funny when they match an overall expected pattern, but fall short of actually being accurate or good Ullman et al. [2016]. Figure 1 shows a sketch of the "hill" pattern. We hypothesize that this pattern is largely independent of the architecture of the learning system. We also

[*]Harvard University, antara.raaghavi@gmail.com
[†]Johns Hopkins University, jennhu@jhu.edu
[‡]Harvard University, tullman@fas.harvard.edu

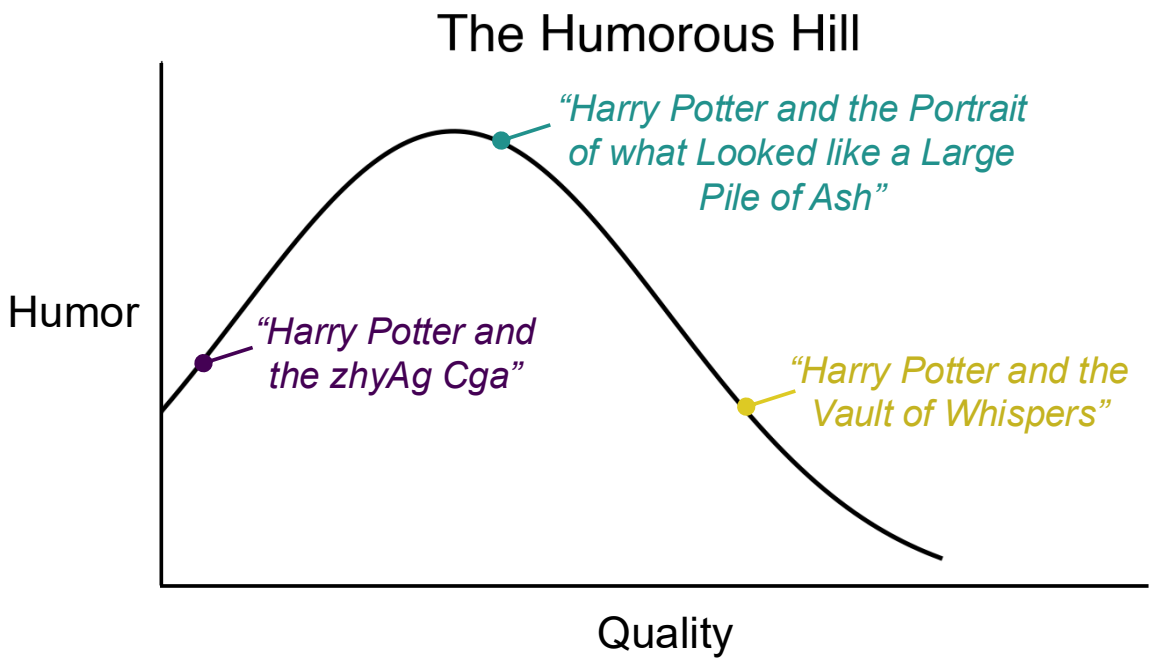

Figure 1: Sketch of the *Humorous Hill* pattern: Learning systems attempting to match a type or category will initially become more (inadvertently) humorous as they improve in quality, before humor falls off as quality improves. Things are most funny when they match an expected pattern, but fall short of being good.

propose that this form of humor underlies the amusement and engagement with many children's utterances, and much of the recent public engagement with artificial intelligence.

We studied the Humorous Hill by examining the relationship between the *quality* of a novel category exemplar (e.g. romance movie titles) and its humor. We systematically manipulated the quality of exemplars by generating them using language models (LMs) of varying capacities. The LMs ranged in architecture and sophistication, including eight n-gram models with increasing window sizes, and five increasingly advanced Transformer models within the GPT family.

We used the 13 LMs to generate 10 novel exemplars each of open-ended categories (names of paint colors, titles of romance movies, or names of towns in the United Kingdom). Models were not tasked with being funny or humorous; their only "goal" was to generate new exemplars based on exposure to real exemplars in each category. We then had human participants rate these exemplars along three dimensions: funniness, accuracy (how well the exemplar belongs to the category) and category membership (whether the exemplar seems to belong to the category at all). The last two were meant as measures of quality. We collected judgments from 25 participants for each combination of model family (n-gram or GPT), category (paint colors, romance movies, United Kingdom towns), and rating dimension (funniness, accuracy, category membership), for a total of $N = 450$ (2 model families $\times$ 3 categories $\times$ 3 dimensions $\times$ 25 participants). Participants were recruited via the Prolific platform. Participants saw half of all exemplars from the model family and category they were assigned to (e.g., n-gram-generated paint color names), and were asked to provide a binary judgment of the dimension that they were assigned to (e.g., "Is this paint color name funny?").

2 Materials and Methods

We used language models of varying quality (eight n-gram models with n $\in \{2, ..., 9\}$ and five transformer-based models from GPT-babbage to GPT-4o) to create novel examples of items in three open-ended categories – names of paint colors, titles of romance movies, and names of towns in the UK. Categories were selected such that they would be familiar to the average person, and suitably open-ended to allow for plausible new outputs. For both model families, models were made to produce new example items in a category – they were not trained to be humorous. After data cleaning and validation, we randomly sampled 10 example outputs per category for each model. We then collected online evaluations of our example outputs from people by asking separate groups of participants to evaluate, as a binary judgment, whether each example was funny, accurate, and belonged in the respective category. For the surprisal experiment, we calculated

the surprisal of our n-gram and GPT-generated list items over the tokens of `meta-llama/Llama-2-7b-hf`. An extended discussion of our methods, as well as references to data, stimuli, and code availability, is included in the appendix. All participants provided informed consent, and the study was approved by the Harvard University Committee on the Use of Human Subjects (IRB19-1861).

3 Results

Figure 2 shows average participant judgments of humor, accuracy, and category membership for each domain and language model. We first verify that the models produce the expected *learning* pattern (getting better at producing outputs over time): we found accuracy and category membership judgments tend to increase as model quality increases along the x-axis. A linear mixed-effects model fit to our data as $\texttt{response} \sim \texttt{model\_quality} + (1 \mid \texttt{participant}) + (1 \mid \texttt{item})$ shows that a one-step increase in model quality increased the proportion of people's judgments of accuracy by 38% ($z = 12.41, p = 2 \times 10^{-16}$), and judgments of category membership by 42% ($z = 12.006, p = 2 \times 10^{-16}$).

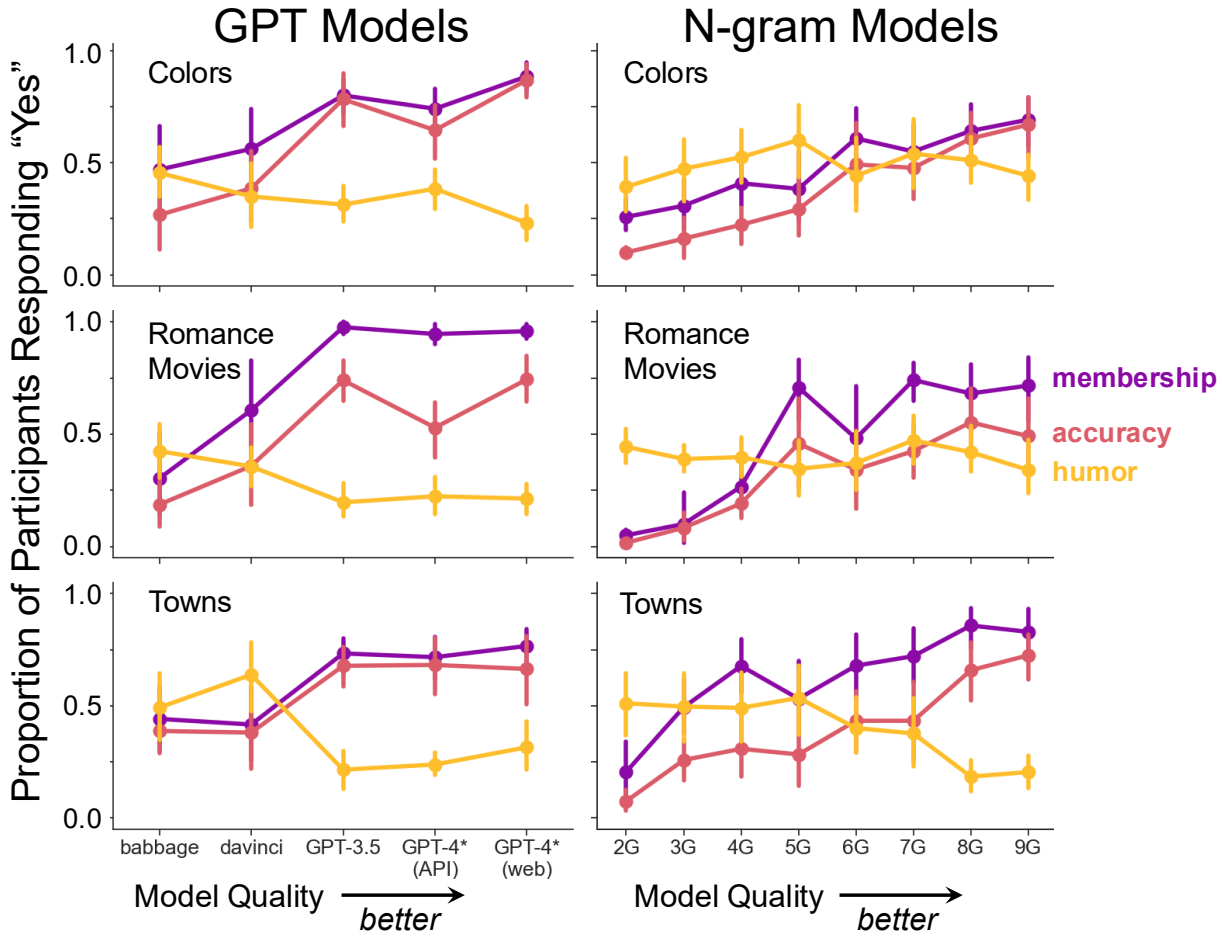

Figure 2: Participants' averaged judgments of category membership, accuracy, and humor, by domain (top to bottom rows), model architecture (left/right columns), and model quality (left-to-right within column). Error bars correspond to 95% confidence intervals. Judgments of output accuracy and category membership become more favorable with model quality, unsurprisingly.

We now turn to the main results, analyzing how judgments of humor relate to quality. Figure 3 shows the relationship between participant evaluations of how funny an output was, and two measures of quality: accuracy (left) and category membership (right). Each point represents an average across items for a single model. Across categories and models, we found an inverted U-shaped relationship between the funniness and quality of exemplars, in line with the Humorous Hill pattern. A quadratic fit was significant for the funniness-quality relationship operationalized either as accuracy (p = 0.016) or category membership ($p = 5.5 \times 10^{-5}$). A comparison of the AIC achieved by second-order (quadratic)

Funniness	vs. category		vs. accuracy	
Domain	**Linear**	**Quadratic**	**Linear**	**Quadratic**
Colors	-25.24	-31.24	-24.00	-29.24
Romance movies	-31.96	-37.82	-32.66	-34.03
UK towns	-24.42	-29.91	-27.08	-29.95

Table 1: Comparison of Akaike information criterion (AIC) achieved by second-order and linear polynomial fits, for each domain and operationalization of quality.

and linear polynomial fits confirms that the second-order fit is preferred to the linear fit in all six cases (preferring the Humorous Hill, and see Table 1).

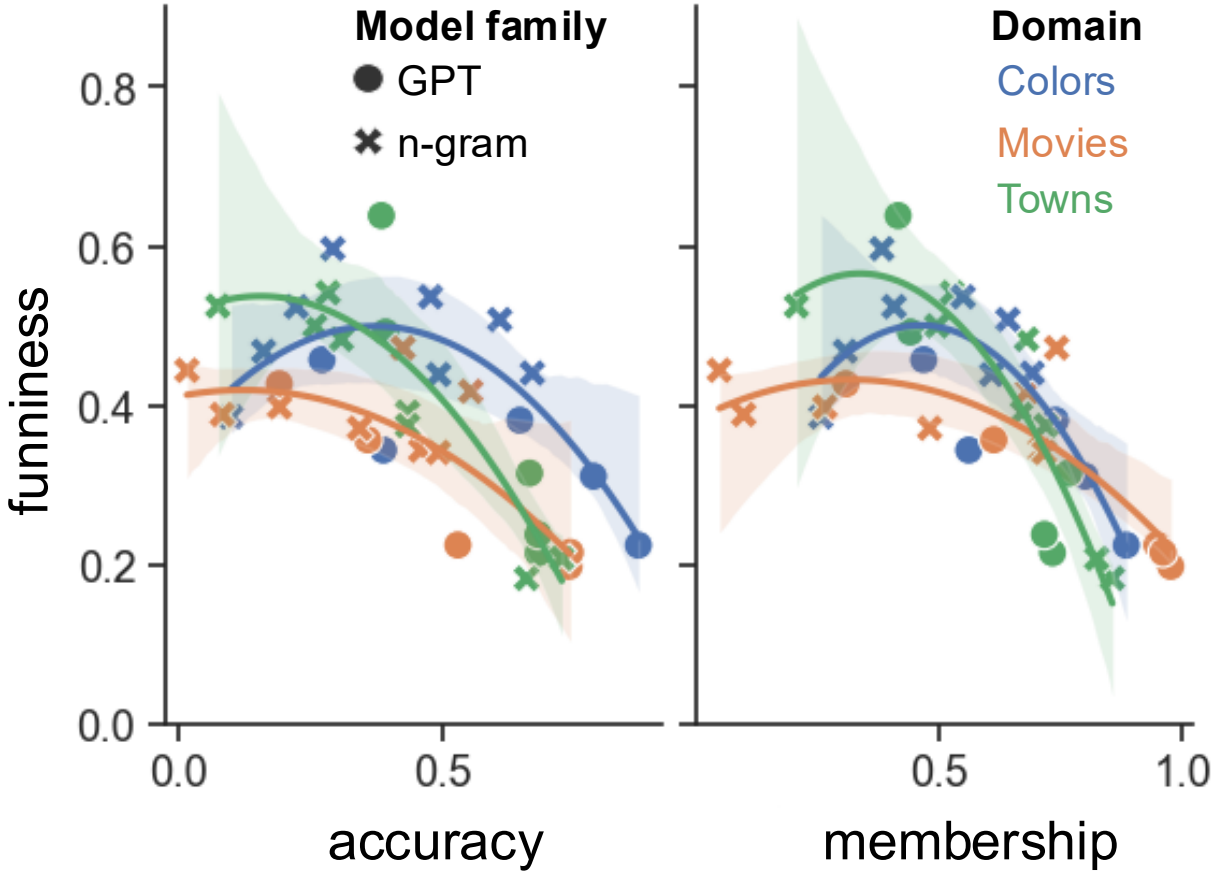

Figure 3: Mean participant judgments of funniness vs. quality, averaged across exemplars for each model and domain. Left: quality as accuracy. Right: quality as category membership. Shaded areas represent 95% confidence intervals. Curves show second-order polynomial fits to the data, pattern in line with the Humorous Hill.

Having found the pattern of interest for different model architectures and domains, we investigated whether humor judgments could be predicted by several current theoretical accounts of humor. We did not consider relief theory and taboo theory *a priori*, as they do not factor into our experimental conditions. To test whether a surprisal-based theory could explain our data, we computed the log-probability of each exemplar using Llama 2 Touvron et al. [2023], an open-source language model that was trained on Internet-scale data, able to capture many of the expectations that people likely bring to categories such as those tested in our experiments (we note that Llama 2 was not itself used to generate any of the exemplars). We found that the log-probability of exemplars, as estimated by Llama 2, was indeed a significant predictor of people's judgments of funniness. But, we note this is a linear relationship between funniness and the summed log-probability of an exemplar ($p = 0.015$) rather than a quadratic relationship (not significant). It may be

that surprisal is also capturing information about the *quality* of the exemplars. In support of this view, we found that the log-probabilities of exemplars were correlated with people's judgments of accuracy (Pearson $r = 0.470, p < 0.0001$) and category membership (Pearson $r = 0.456, p < 0.0001$).

4 Discussion

We identified a pattern within systems designed to output accurate, novel examples of a given category, type, or template. We term this pattern 'The Humorous Hill'. In this pattern, inaccurate output is not seen as funny, but gradually becomes funny as it matches an overall template while still being short on quality, then becomes less funny again as quality improves. While our exploration included a necessarily limited set of domains and architectures, the domains and architectures are different enough, and the pattern general enough, for us to suggest it may hold across systems and domains.

The pattern of the humorous hill is not restricted to artificial systems, and may apply to humans as well. Much of the fascination and delight of sharing children's amusing utterances comes about not because the children are purposely trying to be funny. The children are trying their best to match a given template, response, or category, but they fall short of the mark. When we laugh at a child's attempted 'joke', we're laughing at them, not with them. This holds also for common comedic tropes in which a character is similar enough to humans to produce human-like behavior and speech, but falls short of the mark of actual understanding, including aliens and robots. On this last issue, we suggest that a sizable amount of the cultural engagement with the output of artificial systems is not necessarily because they are good, but because they are entertaining.

The existence of the humorous hill does not directly negate the main existing theories of humor, but they do not directly predict it. Let us consider the different theories in turn, starting as expected with surprisal and incongruity. Decreasing surprisal lends itself easily to a learning story: An entity learning to come up with examples of a category starts by producing unlikely examples, and moves to producing more likely ones. But this would be a monotonic story. To the degree that incongruity and surprise are related to humor, such a proposal should predict things going from more funny to less. But as detailed in our experiments, 'surprisal' does not account for the findings. As for superiority, the humorous hill presents similar challenges for this theory as other previous domains it has difficulty explaining, specifically word-play or things that are funny while also lacking in victims or flaws to feel superior towards. What flaw is being pointed out, or what victim is falling into misfortune by 'Harry Potter and the Portrait of what Looked like a Large Pile of Ash'? It might be argued the superiority is felt towards the entity coming up with the examples, as in 'ha ha, this stupid machine/child can't even come up with a proper Harry Potter title'. But this again would predict a monotonic effect: the machine starts out seeming very dumb and gets more clever, so we should find its output initially more funny, then less so. For relief theory, it is difficult to see why a learner would produce more tension-inducing entities in the middle of its learning curve (or why a mid-strength model would produce more tension-inducing entities than weak or strong ones). Benign violation would need to specify what the potentially aggressive or unsafe interpretation in this dynamic is such that it is particularly salient (or not) in the middle regimes. All of this is not to say the various theories of humor *couldn't* explain or predict the Humorous Hill, if given sufficient chance to defend or adapt themselves. The point is not to rule them out definitively, but to indicate they have work to do to account for an empirical phenomenon.

Certainly *some* aspects of the Humorous Hill are in line with general previous understandings of humor. Various theories emphasize an incongruity with a template, and learners tasked with coming up with examples within a category go through a learning of an overall template. The incongruity in the center of the hill is relatively easy to explain as funny, as is the decreasing humor as examples match the category more and more. The question is why the even less congruous examples at the start of the dynamic are not monotonically funnier. The added idea to account for this would then be that to even be considered as a violation of a template, an example still needs to at least minimally match the template in some way. The starting point of a learning dynamic (or the examples of very poor models) is so outside a given template that their examples cannot even be considered proper candidates for incongruity. Such examples are more like categorical errors or type errors [Magidor, 2009, 2017, Sosa and Ullman, 2022]. These examples are 'not even wrong', and so might not be candidates for any further evaluation (except perhaps as anti-humor or meta-humor). To put the argument by analogy: forget the assessment of humor and consider a different general quality like 'satisfaction'. One might say that the ending of the Odyssey is very 'satisfying', or that it is 'more satisfying' than the ending of the Iliad. But it would be difficult to assess how 'satisfying' the number three is. The number three simply doesn't have the features needed to be evaluated in this way. If one were forced to answer 'is three satisfying?', one might reply 'no', but not because three is frustrating or disappointing. To be a frustrating ending, something still has to be an ending at all.

Acknowledgments We thank members of the CoCoDev lab at Harvard University for their comments, as well as the Kempner Institute for the Study of Natural and Artificial Intelligence for its support.

References

David Aerne. Color names. https://github.com/meodai/color-names, 2024.

Arthur Asa Berger. An anatomy of humor. *An anatomy of humor.*, pages xviii, 192–xviii, 192, 1993. ISSN 1-56000-086-4 (Hardcover). Place: Piscataway, NJ, US Publisher: Transaction Publishers.

Collective Botnik Studios. Harry potter, 2018. URL https://botnik.org/content/harry-potter.html.

Lambert Deckers and Philip Kizer. Humor and the Incongruity Hypothesis. *The Journal of Psychology*, 90(2):215–218, 1975. doi:10.1080/00223980.1975.9915778. URL https://doi.org/10.1080/00223980.1975.9915778.

Charles R. Gruner. *Understanding laughter: The workings of wit and humor*. Nelson-Hall, Chicago, 1978. ISBN 0-88229-186-6.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, 2015. doi:10.1145/2827872.

Ben Hughes. Uk towns list, 2025. URL https://github.com/bwghughes/badbatch/tree/master/data/uk-towns-list.

Matthew M Hurley, Daniel C Dennett, and Reginald B Adams Jr. *Inside jokes: Using humor to reverse-engineer the mind*. MIT press, 2013.

Ofra Magidor. Category mistakes are meaningful. *Linguistics and Philosophy*, 32(6):553–581, December 2009. ISSN 1573-0549. doi:10.1007/s10988-010-9067-0. URL https://doi.org/10.1007/s10988-010-9067-0.

Ofra Magidor. Category mistakes and figurative language. *Philosophical Studies*, 174(1):65–78, January 2017. ISSN 1573-0883. doi:10.1007/s11098-015-0575-1. URL https://doi.org/10.1007/s11098-015-0575-1.

Rod A. Martin and Thomas E. Ford. Classic Theories of Humor. In Rod A. Martin and Thomas E. Ford, editors, *The Psychology of Humor (Second Edition)*, pages 33–69. Academic Press, January 2018. ISBN 978-0-12-812143-6. doi:10.1016/B978-0-12-812143-6.00002-3. URL https://www.sciencedirect.com/science/article/pii/B9780128121436000023.

A Peter McGraw and Caleb Warren. Benign violations: Making immoral behavior funny. *Psychological science*, 21(8): 1141–1149, 2010.

Janelle Shane. Ai recipes are bad (and a proposal for making them worse), May 2021. URL https://www.aiweirdness.com/ai-recipes-are-bad-and-a-proposal-20-01-31/.

Felix Anthony Sosa and Tomer Ullman. Type theory in human-like learning and inference. In *Beyond Bayes: Paths Towards Universal Reasoning Systems Workshop at the International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2210.01634.

Herbert Spencer. The physiology of laughter. *Illustrations of universal progress: A series of discussions*, pages 194–209, 1875. doi:10.1037/12203-004. Place: New York, NY, US.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. URL https://arxiv.org/abs/2307.09288.

Tomer D. Ullman, Max Siegel, Joshua B. Tenenbaum, and Samuel J. Gershman. Coalescing the Vapors of Human Experienceinto a Viable and Meaningful Comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 38, 2016. URL https://escholarship.org/uc/item/5f64z7d7.

JARAM Van Hooff. A comparative approach to the phylogeny of laughter and smiling. *Nonverbal communication*, pages 209–241, 1972.

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

A Materials

A.1 Datasets

Lists were obtained from the "best of" subset of the `https://github.com/meodai/color-names` repository of color names [Aerne, 2024], UK towns were obtained from `https://github.com/bwghughes/badbatch/blob/master/data/uk-towns-list/uk-towns.csv` [Hughes, 2025], and movies from the Movielens dataset `https://grouplens.org/datasets/movielens/` [Harper and Konstan, 2015], which we filtered for romance movies in English.

A.2 Code

Our code and anonymized survey response data is available at `https://github.com/antara-raaghavi/humorous-hill`.

B Methods

B.1 Manipulating model quality

We used language models of varying **quality** to create new examples in our defined open-ended categories. The idea of model quality was operationalized in two ways, resulting in two model families.

We first manipulated a model's inductive bias by training eight n-gram language models, with $n \in \{2, 3, 4, 5, 6, 7, 8, 9\}$ on lists of items in each category obtained from web databases (see supplementary information for details). List items were tokenized at the character-level. We randomly subsampled 10 outputs from the 20 output samples generated from each model (output sampling was capped at 20 because for larger values of n, getting novel items that were not duplicates of items in the input list was taking prohibitively long).

We then manipulated the model's quality by testing increasingly sophisticated variants of transformer-based models within the GPT family, ranging from GPT-babbage to GPT-4o. We accessed babbage-002, davinci-002, gpt-3.5-turbo, gpt-4-0613, and gpt-4o-2024-08-06 via the API, additionally using the web chat interface of the latter two models. These models were prompted to generate new items within the given category.

For both model families, models were made to produce new example items in a category – they were not trained to be humorous. After data cleaning and validation, we randomly sampled 10 outputs per category for each model.

B.2 Human participant ratings

We collected online evaluations of our example outputs from human participants through the Prolific survey platform, restricting our participant pool to native English speakers located in the US. Participants were first shown a consent form and instructions about the experiment, then given two practice trials before being shown 40 (for n-gram) or 25 (for GPT) model outputs in the main experimental trials. Trials were structured as binary yes/no questions about items from "AI-generated lists" of names for paint color swatches, titles of romance movies, and names for towns in the UK. We operationalized model quality in two ways, as category membership and accuracy. For each item, we asked people the following questions to obtain evaluations of funniness and quality:

- **Funniness:** Is this x funny?
- **Accuracy:** Is this a good x?
- **Category membership:** Could this be a x? (NB: here participants were explicitly instructed that it did not matter if the item was a *good* example of a category member)

for $x \in \{$paint color name, romance movie title, name for a town in the UK$\}$.

We administered the funniness, accuracy, and category membership surveys to separate groups of participants to control for cross-survey effects. Surveys were counterbalanced to ensure an even spread of participants across different categories and list items. We filtered outlier survey responses with extreme durations (more than 3 standard deviations away from the mean trial completion time).

B.3 Surprisal experiment

We calculated the surprisal of our n-gram and GPT-generated list items over over the tokens of `meta-llama/Llama-2-7b-hf` (a model that was not used to generate the list items), accessing the model through HuggingFace. Specifically, we computed the log probability of each item as

$$\log P(\text{ITEM"}|\text{'Here is a name for a CATEGORY: ' "})$$

including quotation marks so as to make the model treat each item as a full name for something in a given domain (e.g. paint swatch color) to understand whether surprisal predicted participants' humor ratings for each item.

B.4 Details of GPT model querying

Babbage and Davinci We used the prompt "Need a unique x? Here is a unique y that doesn't exist in real life:" for $x \in \{$name for a color, title for a romance movie, name for a town in England$\}$ and $y \in \{$color name, romance movie title, name for a town in England$\}$ since it resulted in the most coherent output. We collected 50 outputs (querying independently each time). After filtering, we randomly sampled 10 outputs from the remaining list.

GPT-3.5-Turbo, GPT-4 and GPT-4o. We used the prompt "Hi! Could you give me a list of 20 x for y that don't exist?". for $x \in \{$paint color names for a color, title for a romance movie, name for a town in England$\}$ and $y \in \{$colors, romance movies, towns in the UK$\}$.

Due to duplicate responses within and across samples, we ran each query five times for a total of 5 x 20 = 100 outputs that were filtered for duplicates, from which we subsampled 10 outputs.

We used the most recent existing GPT model for the outputs gathered the chat interface. At the time of collecting the prompts for colors/romance movies, this was GPT-4; for UK towns, this was GPT-4o.

B.5 Data cleaning and validation

N-grams. To maintain consistency across the n-grams, we lowercase all samples before training our n-gram models. We use an 80-20 train-test dataset split. When we generate output from the models, we filter out all outputs that are duplicates from the input list, so that all generated outputs shown to query participants are novel.

GPT models. We lowercase all outputs obtained from GPT models when presenting them to survey participants. For the babbage and davinci models, which often produce irrelevant text after responding to the query we provide, we truncate the outputs at the first punctuation mark to cut off irrelevant text. We filter outputs with irrelevant / offensive text and discard all blank outputs.