Words that Work: Using Language to Generate Hypotheses

Rafael M. Batista* James Ross*

Abstract

In this paper, we examine how specific features of language drive consumer behavior. Our contribution, however, lies not in testing specific hypotheses; rather, it is in demonstrating a data-driven process for generating them. We devise an approach that generates interpretable hypotheses from text by integrating large-language models (LLMs), machine learning (ML), and psychology experiments. Using a dataset with over 60,000 headlines (and over 32,000 A/B tests), we produce human-interpretable hypotheses about what features of language might affect engagement. We then test a subset of these hypotheses out-of-sample using two datasets: one consisting of 1,600 A/B tests and another containing over 5,000 social media posts. Our approach indeed facilitates discovery. For instance, we find that describing physical reactions significantly increases engagement. In contrast, focusing on positive aspects of human behavior decreases it. This approach extends beyond a single application. In general, it offers a data-driven method for discovery that can convert unstructured text data into insights that are interpretable, novel, testable, and generalizable. It does so while maintaining a transparent role for both human researchers and algorithmic processes. This approach offers a practical tool to researchers, organizations, and policymakers seeking to aggregate insights from multiple marketing experiments.

1 Introduction

Language shapes people's beliefs and motivates behavior [e.g., 35, 34, 26, 15]. But what is it exactly about everyday messages that drive behavior? Many theories begin to answer this. But the space of possible insights is vast, and discoveries take time, relying on both human creativity and trial-and-error to come up with and then test one hypothesis at a time. How can we efficiently explore the space of testable hypotheses? Recent advances in machine learning (ML) can help [42], but often at the cost of interpretability and understanding [33].

We present a process for generating novel and interpretable hypotheses from text by combining large-language models (LLMs), ML, and choice experiments. Our process begins with a corpus of text and an outcome of interest and outputs a set of hypotheses that are interpretable, testable, novel, and generalizable to other contexts. Two studies test the hypotheses discovered across many A/B tests conducted by two organizations, providing evidence of their effects. In doing so, we also provide organizations and policymakers with a process for aggregating insights from several A/B tests.

This paper is about discovering what features of language drive everyday judgments and decisions. Our contribution, however, is not in testing a specific hypothesis; it is in demonstrating a data-driven process for generating one.

*University of Chicago Booth School of Business.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Method

2.1 Data preparation

We use the Upworthy Research Archive [31], public data containing headline text and click-through rate (CTR) for 32k experiments (A/B tests). To avoid overfitting, we split the data (40% training, 10% morphing, 10% regression, 40% lockbox). Given the experimental setup of the data, we decided to produce our analysis at the pair level, where each observation consists of a pair of headlines. We collected all pairs of headlines H_a and H_b that appeared in the same trial.[2]

The outcome we care about in this application is the click-through rate (CTR). For each headline, the CTR is defined as CTR $= \frac{\text{Clicks}}{\text{Impressions}}$. To account for variability in CTRs arising from trials of different sizes, we employed a shrinkage procedure toward the overall average CTR:

$$\text{Smoothed CTR}_a = \frac{\text{Clicks}_a + \overline{\text{CTR}}}{\text{Impressions}_a + 1} \tag{1}$$

where $\overline{\text{CTR}}$ was the mean CTR calculated across all headlines. Finally, we defined our outcome of interest to be the *difference* in CTR:

$$\Delta\text{CTR}_{a,b} = \text{Smoothed CTR}_b - \text{Smoothed CTR}_a. \tag{2}$$

For simplicity, we refer to Smoothed CTR as CTR in the remainder of this paper.

2.2 Modeling

To motivate our work, we build on [6]. Using this data, they extract over 50 psychological features (e.g., LIWC, TextAnalyzer) and test for their effect on CTR. For each of the 51 psychological constructs used in Bannerjee and Urminky's analyses, we take the difference in construct values between the headlines in each pair. The result is 51 features defined as the difference in a psychological construct (such as reading ease, numeric reference, or visual language). We then estimate an OLS regression of the form

$$\Delta\text{CTR}_{a,b} = \beta_0 + \sum_{i=1}^{51} \beta_i \cdot \Delta\text{Rating } i_{a,b} + \varepsilon_{a,b}, \tag{3}$$

which we will call the 'BU model.'

We then train an ML algorithm to predict $\Delta\text{CTR}_{a,b}$, to explore whether it outperforms the model from (3). We employ a Siamese network architecture [13], initializing with a pre-trained MPNet model [40] to convert each headline into a vector of length 768.[3]. We then take the difference between these vectors, and add a randomly-initialized, fully-connected linear layer, which outputs a single value. The underlying embedding model and the final regression layer are then simultaneously fine-tuned using a standard gradient descent approach, to improve the performance in predicting ΔCTR.

We compared the out-of-sample predictions of the BU model ($R^2 = 0.04$) to one that included the prediction from an ML algorithm we trained using each headline's sentence embedding ($R^2 = 0.13$). Indeed, the ML predictions improved the performance, $F(1, 1690) = 169.4$, $p < .001$. This suggests there is signal in the text to be discovered.

2.3 Hypothesis Generation

To generate interpretable hypotheses from this data, we devised a set of steps that used the headlines to extract a set of features, then sorted them using "predicted" effects obtained with the ML algorithm. Figure 1 gives a graphical overview of the steps, and the full prompts for these steps are included in the Appendix Section A.2.

[2] Our data splitting process ensures that all headlines in a trial are allocated to the same partition, and therefore, all pairs of headlines within a trial are also allocated to the same partition.

[3] We used a version of this model that was fine-tuned on additional data, see `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

The first step extracted features. We provided GPT-4 with pairs of headlines written for the same story and indicated which had the higher CTR. The prompt then elicited a feature that fits the format: "*Hypothesis: _____ [increases/decreases] engagement with a message.*" This step produced 2,100 interpretable hypotheses (which human raters, $n = 79$, also believed to be good quality). A sample of hypotheses is shown in Table 1.

The next step combined GPT and the ML algorithm to produce predicted effects for each hypothesized feature. First, we produced 252,000 counterfactual headlines ("morphs") by having GPT rewrite a set of Upworthy headlines to incorporate each feature. Each morph was based on one actual headline and one of the hypotheses. We then used the ML algorithm to predict the difference in CTR between each morph and the original headline it was based on—i.e., each pair had a predicted treatment effect, which we aggregated at the hypothesis level. By applying the hypotheses to many different headlines and predicting their effect, we get a sense of how generalizable it is (e.g., it's unlikely that a very specific feature would produce a large effect on average). This step incorporated the ML signal and an element of generalizability to rank-order the hypotheses. A sample of headline morphs is shown in Table 2.

The last step narrowed the set. First, for each hypothesis we calculated the average difference between embedding vectors for associated headline and morph pairs, producing a single embedding vector for each hypothesis. We then ranked hypotheses by the score calculated in the previous step. Finally, working from highest to lowest scores, we selected hypotheses that had a Euclidean distance greater than 0.03 from previously selected hypotheses. Finally, we tested whether the predicted treatment effects of the remaining 205 hypotheses were significantly different from 0 (after applying a correction to control for false discovery rates [8]). Sixteen hypotheses had significant, positive predicted effects ($p<.05$). With hypotheses in hand, we tested them in two studies.

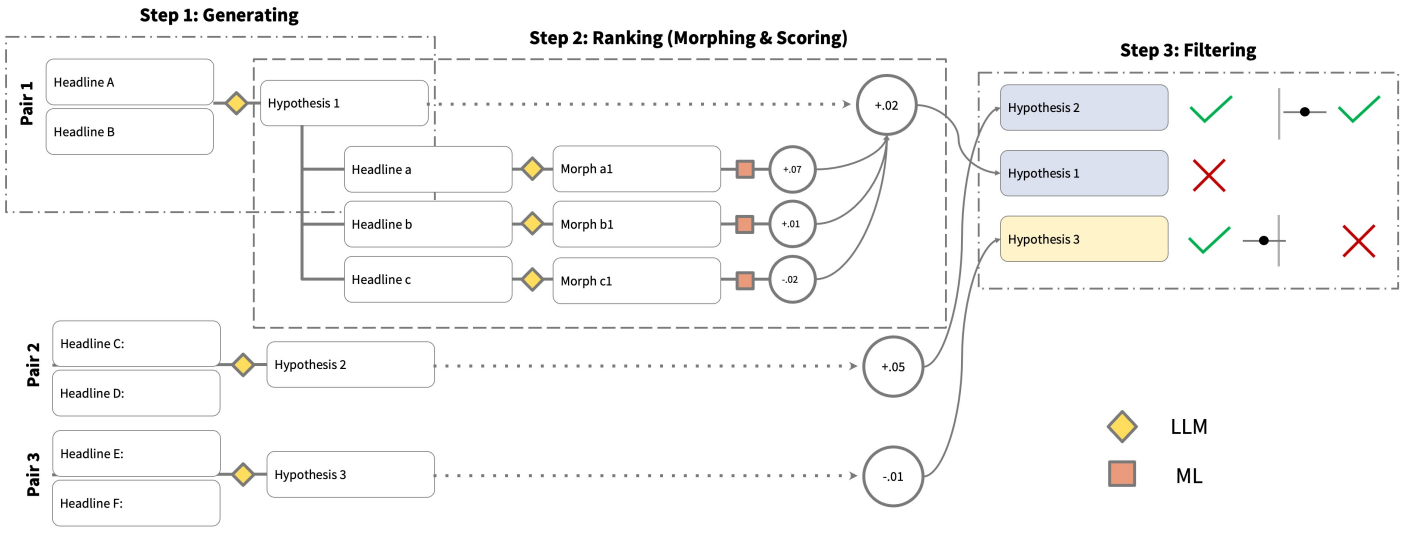

Figure 1: Overview of steps for generating and selecting hypotheses

2.4 Hypothesis Testing

To test our hypotheses — and assuage any concerns of overfitting or *p*-hacking [39, 43] — we pre-registered the six hypotheses and conducted all of our tests out of sample, on data that was intentionally left untouched in all the preceding steps for generating the hypotheses. Hypotheses were generated transparently through the process described above and pre-registered as they came, further restricting our degrees of freedom [22, 38, 24]. The pre-registration of this analysis is available on AsPredicted.org/S6H_ZPF (#172038), including the full text of hypotheses, sample sizes and regression specifications.

Study 1 : We hand-picked 4 of the 16 (+ 2 others predicted to have a negative effect). We then had 800 Prolific users code 3.4k headlines from the hold-out set on each of the 6 features. Each

participant saw 26 headlines, each on a separate page, randomly drawn from the set of 3,402. For each headline, participants were asked to "select the level which each trait is featured in this headline, from '1 (Low)' to '7 (High)'." There was also an option to select "0" to indicate the trait was not present. The traits (i.e., features) were listed by their shorthand: (i) *includes element of surprise followed by cliffhanger*, (ii) *incorporates parody*, (iii) *refers to multimedia evidence*, (iv) *describes physical reaction*, (v) *short and simple phrases*, (vi) *focus on positive aspects of human behavior*.

To test each of the six hypotheses, we estimate six OLS regressions:

$$\Delta\text{CTR}_{a,b} = \beta_0 + \beta_r \cdot \Delta\text{Rating}_{a,b} + \varepsilon_{a,b}. \tag{4}$$

Five of the 6 features had meaningful effects ($p = 0.297$; $p = 0.094$; $p = 0.046$; $p = 0.033$; $p < .001$, $p < .001$). But are these novel? Controlling for BU features, 2 of 6 had significant effects ($p < .001$).

Study 2 : To see whether these effects generalize to new contexts, we conducted the same tests with a different dataset: social media posts by an online entertainment company. The data we obtained contains a total of 553,328 different social media posts for various articles hosted on their website between July 2022 and February 2023. Here, 5,077 posts were split to test the hypotheses. Unlike the Upworthy dataset, the posts were not part of a randomized trial. Therefore, our primary outcome is the CTR (not ΔCTR), defined here as the total clicks divided by the total reach.

We had 900 users code 5,077 messages using the same survey structure as Study 1. One notable exception is that we regressed the CTR on the average rating; we did not differ in the variables as we did for Upworthy since the posts were not paired. Again, four out of the six hypothesized features were significant predictors of CTR ($ps < .01$), including (1) multimedia evidence, (2) physical reactions, (3) short, simple phrases, and (4) a focus on positive, human behavior. These are consistent with the evidence found in the Upworthy data, except for *multimedia*, for which the effect is in the opposite direction, and *surprise, cliffhanger*, for which there is a null effect.

3 Limitations

An obvious limitation of any data-driven approach is that they are inherently *data-driven* (as opposed to *theory-driven* approaches, which start from existing literature or a standard model of the world). Science requires both [37, 3, 28] and, in fact, in marketing, both approaches are regularly used [21]. The downside of data-driven approaches is that without any background knowledge, it can be hard to contextualize observed effects or generalize them to new contexts without further testing. We see an example of this in the case of the *multimedia* feature; more research could help to reconcile the fact that the observed effect is in opposite directions in different domains. For example, different consumer groups may have varied responses to a similar hypothesis. While the framework presented here adds to the toolkit of data-driven approaches, the transparency of the outputs leaves room for researchers to search through the set with an eye for theoretically relevant insights.

An open question remains regarding the right "level" of a hypothesis. In setting up the procedure, we iterated on the prompts before landing on a set where the LLM responded with a hypothesis in a format we felt resembled hypotheses found in past papers. While off-the-shelf LLMs could conceivably draw on existing knowledge to produce more theoretically rich hypotheses [44], leaning into this would increase the chance the LLMs "hallucinated" or drew insights from a world model different from our own [41].

Our pipeline assumes that data comes in the form of A/B tests, but this may be infeasible for some applications. Even with pairs of messages from the same trial, most pairs of headlines vary several things at once, making it hard to isolate sources of variation. It is conceivable that the hypotheses generated reflect this complexity; in fact, some did specify interactions (e.g., "using first-person narration *and* acknowledging personal change in beliefs leads to less engagement with a message," emphasis added). By choosing prompts for which the outputs were both empirically plausible and not overly complex, we may have shifted the distribution of hypotheses to be more substantive than theoretical.

4 Impacts

This paper is intended to help marketing researchers, organizations, and policymakers generate new insights into what drives consumer behavior. We make several significant contributions: First, we introduce a framework to convert unstructured text into marketing insights. There are several recent papers exploring how researchers can use text to study consumer behavior [e.g., 19, 9, 12, 18, 20]. One persistent challenge with this unstructured data is interpretability [17, 18]. The framework we propose utilizes various existing technologies to help address this.

Second, we generate and test actual marketing hypotheses. In doing so, we contribute to the literature studying how language affects engagement [e.g., 6, 25, 11, 10]. Using our framework, we uncover new insights, some adding to existing theories and others inspiring new questions. Although we tested a select set in this paper, our process generated dozens of hypotheses worth examining more closely in future research.

In addition, this paper adds to the literature on organizational learning [36, 14, 16]. Organizations today continuously run A/B tests to learn how various messages affect consumers' behavior [25, 4, 31]. Nevertheless, many of these tests prioritize learning *what* works [e.g., by comparing wholesale changes; 23, 5] at the cost of learning *why*, which typically requires more carefully controlled experiments. This paper demonstrates how to aggregate insights from thousands of A/B tests in the form of specific hypotheses that others can carefully test.

Finally, this paper contributes to the research on data-driven discovery and hypothesis generation [32, 27, 7, 2, 1]. While marketing researchers are driving some of the innovation in this space [e.g., 2, 7], a lot is also happening in outside disciplines such as computer science and economics [27, 45, 30]. This work tries to bridge this literature and, in doing so, broaden the reach of our field [29].

5 Discussion

The hypotheses derived from our framework have practical implications, serving as meaningful predictors of engagement as measured through click-through rates (CTR). These hypothesized features not only capture variation in CTR in the context in which they were discovered but also predict the CTR in other contexts. For instance, using social media posts from an online entertainment company, we found significant correlational evidence supporting four of the six hypotheses above. The evidence that these hypotheses extend to new contexts suggests that companies with multiple messaging channels or several brands can leverage our framework to inform a broader marketing strategy.

Whether these are novel, generalizable, and of general interest remains an open set of questions. On the question of novelty, we provide a partial answer. Statistically, at least two features — *surprise, cliffhanger* and *multimedia reference* — appear to capture information that is sufficiently different from the 51 psychological constructs derived in [6]. Nevertheless, one could argue that these features *appear* similar to insights already known. More empirical work is needed to answer this, so we leave this to future research.

6 Conclusion

The current paper produces new insights into what drives engagement. Importantly, it also offers a general framework that researchers and organizations can use to aggregate marketing insights from text. This framework can be applied whenever there is high-dimensional text data, such as text messages, emails, social media posts, brand slogans, advertising content, and customer service scripts. The data need not be structured, and the process requires little human interpretation. Nevertheless, the output is a set of hypotheses readily interpretable by humans.

This paper presents a novel framework marketers could use to generate hypotheses from text data. Our approach integrates large-language models, machine-learning tools, and psychology experiments to produce hypotheses that are both novel and interpretable. By starting with unstructured data such as text messages, emails, social media posts, or headlines, our framework outputs hypotheses that are interpretable, novel, testable, and generalizable to other contexts.

References

[1] Ralph Adolphs, Lauri Nummenmaa, Alexander Todorov, and James V. Haxby. Data-driven approaches in the investigation of social perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693):20150367, May 2016.

[2] Ada Aka, Sudeep Bhatia, and John McCoy. Semantic determinants of memorability. *Cognition*, 239:105497, October 2023.

[3] Joseph W. Alba. In Defense of Bumbling. *Journal of Consumer Research*, 38(6):981–987, April 2012.

[4] Panagiotis Angelopoulos, Kevin Lee, and Sanjog Misra. Value Aligned Large Language Models, April 2024.

[5] Eduardo M. Azevedo, Alex Deng, José Luis Montiel Olea, Justin Rao, and E. Glen Weyl. A/B Testing with Fat Tails. *Journal of Political Economy*, 128(12):4614–000, December 2020.

[6] Akshina Banerjee and Oleg Urminsky. The Language That Drives Engagement: A Systematic Large-scale Analysis of Headline Experiments., September 2023.

[7] Sachin Banker, Promothesh Chatterjee, Himanshu Mishra, and Arul Mishra. Machine-Assisted Social Psychology Hypothesis Generation, February 2023.

[8] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[9] Jonah Berger, Ashlee Humphreys, Stephan Ludwig, Wendy W. Moe, Oded Netzer, and David A. Schweidel. Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing*, 84(1):1–25, January 2020.

[10] Jonah Berger, Yoon Duk Kim, and Robert Meyer. What Makes Content Engaging? How Emotional Dynamics Shape Success. *Journal of Consumer Research*, 48(2):235–250, August 2021.

[11] Jonah Berger, Wendy W. Moe, and David A. Schweidel. What Holds Attention? Linguistic Drivers of Engagement. *Journal of Marketing*, 87(5):793–809, September 2023.

[12] Jonah Berger and Grant Packard. Wisdom from words: The psychology of consumer language. *Consumer Psychology Review*, 6(1):3–16, 2023.

[13] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature Verification using a "Siamese" Time Delay Neural Network. In *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1993.

[14] George S. Day. Closing the Marketing Capabilities Gap. *Journal of Marketing*, 75(4):183–195, July 2011.

[15] Wendy De La Rosa, Eesha Sharma, Stephanie M. Tully, Eric Giannella, and Gwen Rino. Psychological ownership interventions increase interest in claiming government benefits. *Proceedings of the National Academy of Sciences*, 118(35):e2106357118, August 2021.

[16] Gary F. Gebhardt, Gregory S. Carpenter, and John F. Sherry. Creating a Market Orientation: A Longitudinal, Multifirm, Grounded Analysis of Cultural Transformation. *Journal of Marketing*, 70(4):37–55, October 2006. Publisher: SAGE Publications Inc.

[17] Jochen Hartmann, Juliana Huppertz, Christina Schamp, and Mark Heitmann. Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38, March 2019.

[18] Jochen Hartmann and Oded Netzer. Natural Language Processing in Marketing. In K. Sudhir and Olivier Toubia, editors, *Artificial Intelligence in Marketing*, volume 20 of *Review of Marketing Research*, pages 191–215. Emerald Publishing Limited, January 2023.

[19] Ashlee Humphreys and Rebecca Jen-Hui Wang. Automated Text Analysis for Consumer Research. *Journal of Consumer Research*, 44(6):1274–1306, April 2018.

[20] Joshua Conrad Jackson, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A. Lindquist. From Text to Thought: How Analyzing Language Can Advance Psychological Science. *Perspectives on Psychological Science*, 17(3):805–826, May 2022. Publisher: SAGE Publications Inc.

[21] Chris Janiszewski and Stijn M. J. van Osselaer. The Benefits of Candidly Reporting Consumer Research. *Journal of Consumer Psychology*, 31(4):633–646, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcpy.1263.

[22] Norbert L. Kerr. HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3):196–217, August 1998. Publisher: SAGE Publications Inc.

[23] Rembrand Koning, Sharique Hasan, and Aaron Chatterji. Experimentation and Start-up Performance: Evidence from A/B Testing. *Management Science*, 68(9):6434–6453, September 2022. Publisher: INFORMS.

[24] Justin F. Landy, Miaolei (Liam) Jia, Isabel L. Ding, Domenico Viganola, Warren Tierney, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Charles R. Ebersole, Quentin F. Gronau, Alexander Ly, Don Van Den Bergh, Maarten Marsman, Koen Derks, Eric-Jan Wagenmakers, Andrew Proctor, Daniel M. Bartels, Christopher W. Bauman, William J. Brady, Felix Cheung, Andrei Cimpian, Simone Dohle, M. Brent Donnellan, Adam Hahn, Michael P. Hall, William Jiménez-Leal, David J. Johnson, Richard E. Lucas, Benoît Monin, Andres Montealegre, Elizabeth Mullen, Jun Pang, Jennifer Ray, Diego A. Reinero, Jesse Reynolds, Walter Sowden, Daniel Storage, Runkun Su, Christina M. Tworek, Jay J. Van Bavel, Daniel Walco, Julian Wills, Xiaobing Xu, Kai Chi Yam, Xiaoyu Yang, William A. Cunningham, Martin Schweinsberg, Molly Urwitz, The Crowdsourcing Hypothesis Tests Collaboration, and Eric L. Uhlmann. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 146(5):451–479, May 2020.

[25] Dokyun Lee, Kartik Hosanagar, and Harikesh S. Nair. Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. *Management Science*, 64(11):5105–5131, November 2018. Publisher: INFORMS.

[26] Elizabeth Linos, Jessica Lasky-Fink, Chris Larkin, Lindsay Moore, and Elspeth Kirkman. The formality effect. *Nature Human Behaviour*, 8(2):300–310, February 2024. Publisher: Nature Publishing Group.

[27] Jens Ludwig and Sendhil Mullainathan. Machine Learning as a Tool for Hypothesis Generation*. *The Quarterly Journal of Economics*, page qjad055, January 2024.

[28] John G. Lynch, Joseph W. Alba, Aradhna Krishna, Vicki G. Morwitz, and Zeynep Gürhan-Canli. Knowledge creation in consumer research: Multiple routes, multiple criteria. *Journal of Consumer Psychology*, 22(4):473–485, October 2012.

[29] Deborah J. MacInnis, Vicki G. Morwitz, Simona Botti, Donna L. Hoffman, Robert V. Kozinets, Donald R. Lehmann, John G. Lynch, and Cornelia Pechmann. Creating Boundary-Breaking, Marketing-Relevant Consumer Research. *Journal of Marketing*, 84(2):1–23, March 2020. Publisher: SAGE Publications Inc.

[30] Benjamin S. Manning, Kehang Zhu, and John J. Horton. Automated Social Science: Language Models as Scientist and Subjects, April 2024.

[31] J. Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media. *Scientific Data*, 8(1):195, August 2021.

[32] William J. McGuire. Creative Hypothesis Generating in Psychology: Some Useful Heuristics. *Annual Review of Psychology*, 48(1):1–30, 1997.

[33] Lisa Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, March 2024. Publisher: Nature Publishing Group.

[34] Katherine L. Milkman, Linnea Gandhi, Mitesh S. Patel, Heather N. Graci, Dena M. Gromet, Hung Ho, Joseph S. Kay, Timothy W. Lee, Jake Rothschild, Jonathan E. Bogard, Ilana Brody, Christopher F. Chabris, Edward Chang, Gretchen B. Chapman, Jennifer E. Dannals, Noah J. Goldstein, Amir Goren, Hal Hershfield, Alex Hirsch, Jillian Hmurovic, Samantha Horn, Dean S. Karlan, Ariella S. Kristal, Cait Lamberton, Michelle N. Meyer, Allison H. Oakes, Maurice E. Schweitzer, Maheen Shermohammed, Joachim Talloen, Caleb Warren, Ashley Whillans, Kuldeep N. Yadav, Julian J. Zlatev, Ron Berman, Chalanda N. Evans, Rahul Ladhania, Jens Ludwig, Nina Mazar, Sendhil Mullainathan, Christopher K. Snider, Jann Spiess, Eli Tsukayama, Lyle Ungar, Christophe Van den Bulte, Kevin G. Volpp, and Angela L. Duckworth. A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences*, 119(6):e2115126119, February 2022.

[35] Katherine L. Milkman, Mitesh S. Patel, Linnea Gandhi, Heather N. Graci, Dena M. Gromet, Hung Ho, Joseph S. Kay, Timothy W. Lee, Modupe Akinola, John Beshears, Jonathan E. Bogard, Alison Buttenheim, Christopher F. Chabris, Gretchen B. Chapman, James J. Choi, Hengchen Dai, Craig R. Fox, Amir Goren, Matthew D. Hilchey, Jillian Hmurovic, Leslie K. John, Dean Karlan, Melanie Kim, David Laibson, Cait Lamberton, Brigitte C. Madrian, Michelle N. Meyer, Maria Modanu, Jimin Nam, Todd Rogers, Renante Rondina, Silvia Saccardo, Maheen Shermohammed, Dilip Soman, Jehan Sparks, Caleb Warren, Megan Weber, Ron Berman, Chalanda N. Evans, Christopher K. Snider, Eli Tsukayama, Christophe Van den Bulte, Kevin G. Volpp, and Angela L. Duckworth. A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118(20):e2101165118, May 2021.

[36] Christine Moorman and George S. Day. Organizing for Marketing Excellence. *Journal of Marketing*, 80(6):6–35, November 2016. Publisher: SAGE Publications Inc.

[37] Chad R. Mortensen and Robert B. Cialdini. Full-Cycle Social Psychology for Theory and Application. *Social and Personality Psychology Compass*, 4(1):53–63, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9004.2009.00239.x.

[38] Mark Schaller. The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology*, 66:107–115, September 2016.

[39] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. Pre-registration: Why and How. *Journal of Consumer Psychology*, 31(1):151–162, 2021.

[40] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for Language Understanding. September 2020.

[41] Keyon Vafa, Justin Y. Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. Evaluating the World Model Implicit in a Generative Model, June 2024. arXiv:2406.03689 [cs].

[42] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, August 2023. Publisher: Nature Publishing Group.

[43] Jelte M. Wicherts, Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, 7, November 2016. Publisher: Frontiers.

[44] Eunice Yiu, Eliza Kosoy, and Alison Gopnik. Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet). *Perspectives on Psychological Science*, page 17456916231201401, October 2023. Publisher: SAGE Publications Inc.

[45] Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis Generation with Large Language Models, April 2024. arXiv:2404.04326 [cs].

A Appendix / supplemental material

A.1 Examples

Below is a sample of outputs produced by GPT-4, both for the hypothesis generation and morphing steps of the pipeline.

Table 1: Examples of sampled headline pairs and generated hypotheses

Headline A	**Headline B**	**Hypothesis**
A Holocaust Survivor's Compassionate Message To The German Population	A 90-Second Message From A 90-Year-Old Holocaust Survivor	Specifying the length of content in the headline results in more engagement with a message
These Kids Don't Pass Go And They Don't Collect $200.	Behind These Numbers Sit Really Sad Truths About Our Justice System - And Some Really Young People	Incorporating emotional language results in more engagement with a message.
It's Probably Your 2nd Favorite Thing To Do And Now Science Wants You To Do More Of It	If You Think It Feels Great, You Should See What Else It's Doing To You	Framing a message to highlight unexpected benefits increases engagement with a message.
I Used To Think Adaptation Was A Good Thing Until I Realized How Humans Do It	Baby Polar Bear: 'What Use Is All This Fur If There's No Ice?' Mama Bear: 'Hush Up And Adapt'.	Personifying animals in the messaging affects engagement with a message.
She Wanted To Make Sure Everyone Knew That Her Baby Was A Boy. So She Dressed Him In Pink.	She Wants Everyone To Know That She's A Proud Mother Of A Boy, So She Dresses Him In Pink	Using past tense instead of present tense decreases engagement with a message.
Elizabeth Warren Forced To Lecture Bank Regulator Like He's A Child Who Did Something Awful	Elizabeth Warren Teaches A Bank Regulator How To Do His Job Like A Big Boy	Using a condescending tone decreases engagement with a message.

Note: To view more examples, visit `https://bit.ly/jmp-hyp-samp`. Complete set available on OSF.

Table 2: Examples of hypotheses, original headlines and the associated morphs

Hypothesis	Original Headline	Morphed Headline
Incorporating emotional triggers and a geographic reference into a headline affects engagement with a message.	That Cheap Stuff I Just Bought At Walmart? Turns Out, It Cost Me $6000 More Than I Thought	Local Man's Walmart Bargain Turns Nightmare: Hidden Costs Rack Up $6000!
Personalizing a message by focusing on an individual's story or reaction makes people more likely to engage with a message.	11 Tweets That Sum Up The Horror In North Carolina	North Carolina Resident's Heart-Wrenching Reaction Captures the Horror in 11 Tweets
Excessive sensationalism and vague phrasing leads to less engagement with a message.	An 11-year old ate a burger with a surprise ingredient. It was fatal, but ok according to the FDA.	11-Year-Old's Fatal Reaction to FDA-Approved Burger Ingredient Sparks Outrage
Introducing a narrative arc and highlighting societal themes leads to more engagement with a message.	A woman shares some thoughts on why 'being normal' isn't all it's cracked up to be.	A Brave Woman's Journey From Conforming to Defying Society: Why Rejecting 'Normal' Opens the Door to True Self-Discovery
Introducing a sense of mystery or unresolved tension affects engagement with a message.	A Haunting Photo Of Martin Luther King Jr. Plus His Immortal Audio Clip	Discover the Mystery Behind Martin Luther King Jr.'s Last Haunting Photo and Immortal Words
Introducing an element of surprise and emphasizing the impact of unawareness leads to more engagement with a message.	Food Stamps Cannot Be Used To Buy Weapons. Except In Alaska.	You Thought Food Stamps Were Just for Groceries? Guess Again, Especially in Alaska!

Note: To view more examples, visit `https://bit.ly/jmp-morph-samp`. Complete set available on OSF.

A.2 Prompts

We use large language models to generate hypotheses, produce morphs, and rate different pieces of text on various hypotheses. For each of these tasks, we require a prompt, to guide the language model's output. In order to minimize the dependence of any results on a particular prompting approach, we also introduce some randomization in the prompting process. In this section, we include a full base prompt for each task, and outline the variations applied to the base prompt. The full materials will be made available through the OSF: `https://osf.io/d5xvb/?view_only=301ca63ed1004401adb697a625ff8d61`. In particular, we highlight the `prompts.yaml` file in the OSF, which includes the raw text of all prompt formats.

A.2.1 Generating hypotheses

Our prompt for generating hypotheses takes a pair of headlines, H_A and H_B, from the same A/B test as input. It specifies that the language model should identify a feature that changed moving from H_A to H_B. In addition, it provides additional context by specifying a role for the language model and a structure for the hypothesis. We also impose some requirements on quality, to ensure that the resulting hypothesis satisfied our goals of clarity, generalizability, empirical plausibility, unidimensionality, and usability.

Within this format, we then varied multiple elements. Firstly, we randomized the role, including an editor or communication scientist for example. Secondly, we varied the hypothesis structure by providing different specific endings. Thirdly, we included more or less information for GPT by possibly giving examples of previous hypotheses, examples of "known constructs" which GPT was instructed to avoid, or removing the example headlines (to serve as a control). Below, we include some examples or an excerpt from each type of randomization.

- **Preamble**: One of nine different preambles was selected, to encourage analytical thought. Examples include:

1. *an editor of a top marketing journal such as the Journal of Consumer Research or the Journal of Marketing*,
2. *a communication scientist researching the effects of linguistic framing on reader perception*, and
3. *a consumer psychology expert specializing in persuasive messaging*.

- **Hypothesis structure**: One of eight different hypothesis structures was selected, to force a format for the output hypothesis that was compatible with later analysis. The `{direction}` key was filled in with the "more [less]" or "increases [decreases]" depending on whether $\widehat{m}_{a,b}$ was positive or negative. Examples include:
  1. Hypothesis: _____ leads to `{direction}` engagement with a message.
  2. Hypothesis: _____ makes people direction likely to engage with a message.
  3. Hypothesis: _____ influences engagement with a message.
- **Variations**: We also created three additional variations to the base prompt.
  1. **Control**: This variation did not refer to any Upworthy headlines and was included to later assess whether hypotheses generated by GPT with access to our dataset differed from those generated by GPT without any specific headlines.[4]
  2. **Examples**: In this variation, we included some examples of ideal hypotheses. This included *"taking photos with the intention to share will induce self-presentational concern and generate disutility, thus actually decreasing enjoyment of the current experience"* and *"perception of moving at faster speed results in more abstract mental representation and choices consistent with desirability"*, for example.
  3. **Known constructs**: In this variation, we included some known constructs, sourced from the BU analysis. This included *Reading Ease: Simpler and easier to read and understand* and *Common Words: Contains more simple or common words*, for example.

The complete set of prompts was made by taking the base prompt format, sampling one of the 9 preambles, one of the 8 structures, and one of the 4 variations (the three listed, plus the possibility of no variation).

A.2.2 Generating morphs

Our prompt for generating morphs takes three examples of headlines from Upworthy, a single headline, H, and a hypothesis, D. When sampling examples and headlines, we ensure that all four headlines come from different trials. The prompt then includes instructions to rewrite headline H according to the given instructions D, while keeping the content of the story as similar as possible.

In addition to the base prompt for morphing, we introduced two variations. The first instructed GPT to produce two variations as output: one that increased the feature of interest by 75%, and another that decreased the feature of interest by 75%. The second variation specified that the morph should be as similar to the original headline in nearly every way except for the feature being changed.

A.2.3 Labeling

Our prompt for labeling headlines takes a single headline, H, and a hypothesis, D, as input. It specifies that the language model should evaluate the given headline on the given hypothesis on a scale of 0 to 7.

[4] Since we planned to exclude these from the rest of the pipeline, prompts that had Control instructions were undersampled before being matched to a pair.

A.3 Online survey materials

The full survey text is also available through the OSF: `https://osf.io/d5xvb/?view_only=301ca63ed1004401adb697a625ff8d61`. In particular, the `.qsf` files are Qualtrics survey exports, that allow others to fully recreate the survey. This includes the IRB-approved consent, instructions, compensation, and survey design (including ramdomization). We additionally note that all surveys came with compensation above the federal minimum wage, in addition to any bonus compensation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claim is to have a process for generating interpretable hypotheses. This process is outlined in the body, along with significant results for the outputs of the process.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Multiple limitations, considerations, and analysis are provided in Section 3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not provide theoretical results in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We outline both a process and experimental results in our paper. The Upworthy data is publicly accessible and the process is described in detail that should be sufficient for reproducing the entire pipeline, and the experimental results are described in sufficient results for a statistical replication. Note however that the secondary dataset is not publicly available, and replications do require collecting data from new Prolific experiments until we release the data collected already.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The Upworthy data, required for the replication of the main results, is publicly accessible. The code for this project is available on OSF (linked in Appendix Subsection A.2), along with material for the prompts, the online surveys, data cleaning, and additional experiments not included in the current submission. Unfortunately, the dataset used in Study 2 is not publicly avaiable, and is not able to be released.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Information about all data preparation and model preparation are included. Some minor details have been omitted, but are generally not critical to the overall hypothesis generation procedure: the learning rate, for example, is of limited importance to the overall generation procedure. Details of the experiments are included, along with relevant code on OSF.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Significance estimates are provided on all key quantities of interest. In particular, p-values for all regression coefficient estimates are included, along with F-statistics for model comparisons.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: None of the algorithm steps or experimental results are compute intensive and can be completed in a reasonable time (within an hour) for all training and evaluation, so we have omitted these details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All research adhers to IRB requirements and the NeurIPS Code of Ethics. Human participants were paid fairly, anonymity has been preserved, and there is minimal risk of harmful consequences from this research.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have included a discussion of potential impacts for work in Section 4.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our specific application (click-through on news headlines) poses minimal risk of misuse from release, especially since the dataset is already publicly available.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: For the main dataset, we cite the original Upworthy publication [31], which contains the details of release. For the secondary dataset, we preserve the anonymity of the provider at their request. Model architectures and pre-training coefficients are cited appropriately.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We are not currently releasing any assets with this paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: All survey materials are available on OSF, linked in Appendix Section A.3.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: All participants were given informed consent according to the IRB requirements at the University of Chicago. These are visible in the supplemental resources for online experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.