Words that Work: Using Language to Generate Hypotheses

Rafael M. Batista* James Ross*

September 9, 2024

Latest Version Here

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. A third hypothesis posited that referring to multimedia (e.g., GIFs, videos) would influence engagement, and it does, only it significantly increases engagement in one domain while significantly decreasing it in another. 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.

Keywords:
Consumer Psychology; Consumer Language; Machine Learning; Text Analysis; Organizational Learning; Hypothesis Generation; A/B Testing

*University of Chicago Booth School of Business. For valuable comments and advice we thank Sendhil Mullainathan and Abigail Sussman. Corresponding author: rafael.batista@chicagobooth.edu.

INTRODUCTION

Language is core to consumer behavior; not only as a topic to study consumers' decision processes (Pogacar et al. 2022; Packard and Berger 2024; Berger and Packard 2023), but also as data, providing researchers and practitioners alike a rich source of insights about customers and companies (Berger et al. 2020; Humphreys and Wang 2018). This paper uses language as data to investigate what motivates consumers to engage with a message.

Many papers in marketing have explored the effects of language on engagement. Much of this research examines what features of text draw consumers' attention (e.g., Banerjee and Urminsky 2023; Bruce, Murthi, and Rao 2017; Kanuri, Chen, and Sridhar 2018; Zor, Kim, and Monga 2022; Robertson et al. 2023) and sustain it (e.g., Berger, Moe, and Schweidel 2023; Berger, Kim, and Meyer 2021; Deolankar et al. 2024).[1] However, these insights often depend on the context, platform, and population in which they are discovered, making it hard for practitioners and marketing researchers to make compelling predictions about engagement in a particular real-world setting (Banerjee and Urminsky 2023).

For a given application, the space of possible insights is vast, and discoveries take time (Chu and Evans 2021; Rzhetsky et al. 2015; Fiedler 2018). The current approach relies on both human ingenuity and trial and error to come up with and then test one hypothesis at a time. When studying consumer language, this task is particularly vexing because language encompasses many dimensions (Aka, Bhatia, and McCoy 2023; Clark 1973). Where does one even begin?

Machine learning (ML) can help uncover patterns that humans may miss (Oquendo et al. 2012; Shin et al. 2023; Wang et al. 2023), but utilizing these tools for discovery often comes at the cost of interpretability and understanding (Messeri and Crockett 2024). Therefore, marketing managers and scholars would benefit from an approach that helps uncover insights from existing data and presents those insights in a format that humans could readily

[1] While engagement can mean different things in marketing (see Table 1 in Berger, Moe, and Schweidel 2023, but also Brodie et al. 2011), we will focus primarily on attracting attention.

understand.

The goal of the present research is to advance our understanding of the effects of language on engagement while offering a framework to researchers, organizations, and policymakers that they could generate hypotheses about what drives engagement in their specific context with their customers and constituents. Similar to existing work in consumer psychology, we study how modifying the language in a message affects consumers' propensity to engage with it. For instance, does framing a message with an element of surprise, followed by a cliffhanger, make people more likely to engage with it? Does describing physical reactions make a message more engaging? How does focusing on positive aspects of human behavior affect engagement?

The main contribution of this paper, however, is not in testing these specific hypotheses (although we do that too); instead, it is in developing the data-driven process that generated them. This paper proposes a framework for generating novel and interpretable hypotheses from text using a combination of large-language models (LLMs), machine learning (ML) tools, and psychology experiments. Large-language models, such as OpenAI's GPT-4, play a crucial role in processing text data and generating coherent hypotheses (Banker et al. 2023; Demszky et al. 2023). At the same time, off-the-shelf machine-learning tools help to uncover meaningful patterns in large volumes of unstructured data (Wang et al. 2023; Ludwig and Mullainathan 2024). We integrate both technologies and validate the various steps through standard psychology experiments. The framework, therefore, consists of three steps: (1) *generating* a set of hypotheses from observations in the data, (2) *ranking* the set of hypotheses using a machine learning algorithm trained on past outcomes, and (3) *filtering* the set of hypotheses to select the ones most likely to have a meaningful effect.

We apply our framework to aggregate insights from several thousand marketing experiments and generate hypotheses for a specific application: how language affects engagement. For this application, we are particularly interested in what features of language drive consumers to click on a headline. We use the Upworthy Research Archive (Matias et al. 2021),

which contains 32,487 randomized field experiments ("A/B tests") that test 64,983 unique headlines across 150,817 experimental arms. For each experimental arm, we also see the click-through rate (CTR), which we use to measure engagement.[2]

Overview of the Framework. The first step is to *generate* hypotheses from observations in the data. We provided OpenAI's GPT with 2,100 unique pairs of headlines written for the same story to produce 2,100 hypotheses. For example, when provided the pair, "Headline A: I Thought Long And Hard About Sharing This But I Decided I Had To Because These Dogs Need Our Help" and "Headline B: Don't Click This If You're Looking For Something That'll Make You Feel Better About The Human Race" GPT responded with, "Hypothesis: Incorporating reverse psychology leads to more engagement with a message."[3] To assess the quality of these, we had 79 human participants rate a subset of hypotheses on several dimensions. Most of the hypotheses were perceived to be clear, usable, and generalizable to new contexts (see also Banker et al. 2023).

The second step is to *rank* the hypotheses by their predicted effects. We use an ML algorithm — trained to predict CTR using the high-dimensional information contained in the headlines (i.e., sentence embeddings; Song et al. 2020) — to identify which hypothesized features are likely to have an effect when applied to various messages. We start by applying each hypothesis to several different headlines, again using GPT. For example, for a given hypothesis (e.g., "Incorporating reverse psychology leads to more engagement with a message.") applied to an actual Upworthy headline (e.g., "Folks Who Work In Tipped Jobs Would Like You To Spend A Minute Looking At Something"), GPT produces an alternative, "morphed" headline (e.g., "You Probably Shouldn't Read This if You Think Tipping Is Optional").[4] For each hypothesis, we generated approximately 73 morphed headlines, each

[2]Among the benefits of this dataset is that it has been used before in consumer language research (e.g., Banerjee and Urminsky 2023; Robertson et al. 2023; Gligorić et al. 2023; Hopkins, Lelkes, and Wolken 2023; Shulman, Markowitz, and Rogers 2024); thus, allowing us to build on the work of others and benchmark our findings against existing insights extracted from this data.

[3]The complete set of hypotheses can be sampled online at bit.ly/jmp-hyp-samp

[4]The complete set of morphs can be sampled online at bit.ly/jmp-morph-samp

one paired to one actual Upworthy headline, producing a total of 250,000 morph-original pairs. We then used the ML algorithm to *predict* what the CTR would be for the morphed headline relative to the original headline. This approach to morphing and scoring produces a "predicted treatment effect" (PTE) for each morph that we can then aggregate at the hypothesis level. We use the average PTE of a hypothesis to rank-order the set. Therefore, we rank hypotheses using a measure that incorporates both the ML signal and an element of generalizability (since we apply each hypothesis to many headlines before estimating an effect).

The third step is to *filter* the hypotheses. For nearly every pair, GPT produced a hypothesis that was understandable and plausible. While most were sensible, we cannot expect them all to be discoveries. Many of them overlapped with others in the set. Some were too specific (e.g., "Hypothesis: using language that humanizes animals affects engagement with a message") and would not apply to most other messages. To narrow the set, we group similar hypotheses together using a sequential selection strategy. We go through the ordered list, starting at the top, and select unique hypotheses. When we come across a hypothesis similar to one already selected, we exclude it from the list. This grouping reduced our set down to 205 hypotheses. Finally, we calculate a test statistic for each of the remaining hypotheses to determine whether their predicted effects differ meaningfully from zero. In the end, 16 hypotheses had average predicted effects that were positive and significant ($p < .05$).

Among the hypotheses generated are six illustrative hypotheses that we selected to test out of sample.[5] Four of these hypotheses were predicted to increase engagement: 1) framing a message with an element of surprise followed by a cliffhanger, 2) incorporating a concept of parody, 3) incorporating multimedia evidence, and 4) describing physical reactions. Two were predicted to decrease engagement: 5) shortening and simplifying phrases and 6) focusing on positive aspects of human behavior.

[5]Testing these hypotheses involved humans reading through headlines and labeling them based on the hypothesized features. We hand-picked six to facilitate this process; even then, we gathered over 140,000 human labels. We could have picked a different set and in the Appendix we report the results for similar out-of-sample tests on 400 hypotheses randomly selected (only we use GPT-labels rather than human labels).

After generating the hypotheses, we test them. We take advantage of the experimental setup of the data to estimate a causal effect using standard approaches. We intentionally left some of the data untouched when training the algorithm and throughout each step of the pipeline so that we could conduct these tests out of sample. We recruited 800 participants to code 3,386 headlines from 1,693 unique pairs of original Upworthy headlines, producing more than 100,000 ratings. Each pair was from the same randomized trial, which allowed us to estimate the effect on CTR of changing a specific feature (e.g., "reference to multimedia"). Of the six hypotheses that we selected and pre-registered for testing, four had a significant effect on engagement (two $ps < .001$ and two $ps < .05$), and a fifth had a marginal effect ($p = .094$). Of these five, all showed effects in the predicted direction. To verify whether the features discovered through this process coincide with insights within the algorithm's "black box", we regress the algorithm's prediction on the human ratings for each feature. All six features were significant predictors of the ML algorithm, independently and in a multivariate model ($ps < .001$).

Finally, we explore whether these hypothesized features predict similar outcomes in other settings. Specifically, we test the same set of hypotheses using a similar modeling specification in a second dataset containing social media posts made by an online entertainment company. The messages resemble Upworthy's headlines in style and content, but the time period, platform, audience, and organizational strategy differ. Although these posts were not released as A/B tests, they still provide correlational evidence in support of four of the six hypotheses discovered using the Upworthy dataset.

For both datasets, we have reserved 40% of the data for us to analyze upon conditional acceptance of the paper. Therefore, the current manuscript serves as a registered report of our method and findings (Nosek and Lakens 2014; Chambers and Tzavella 2021; Urminsky and Dietvorst 2024).

Contributions. 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., Humphreys and Wang 2018; Berger et al. 2020; Berger and Packard 2023; Hartmann and Netzer 2023; Jackson et al. 2022). As more of our everyday language is captured — through audio, video, or online communication — there will be more data to explore. One persistent challenge with this unstructured data is interpretability (Hartmann et al. 2019; Hartmann and Netzer 2023). 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., Banerjee and Urminsky 2023; Lee, Hosanagar, and Nair 2018; Berger, Moe, and Schweidel 2023; Berger, Kim, and Meyer 2021). 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 (Moorman and Day 2016; Day 2011; Gebhardt, Carpenter, and Sherry 2006). Organizations today continuously run A/B tests to learn how various messages affect consumers' behavior (Lee, Hosanagar, and Nair 2018; Angelopoulos, Lee, and Misra 2024; Matias et al. 2021). Nevertheless, many of these tests prioritize learning *what* works (e.g., by comparing wholesale changes; Koning, Hasan, and Chatterji 2022; Azevedo et al. 2020) 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 (McGuire 1997; Ludwig and Mullainathan 2024; Banker et al. 2023; Aka, Bhatia,

and McCoy 2023; Adolphs et al. 2016). While marketing researchers are driving some of the innovation in this space (e.g., Aka, Bhatia, and McCoy 2023; Banker et al. 2023), a lot is also happening in outside disciplines such as computer science and economics (Ludwig and Mullainathan 2024; Zhou et al. 2024; Manning, Zhu, and Horton 2024). This work tries to bridge this literature and, in doing so, broaden the reach of our field (MacInnis et al. 2020).

Related Work. Our work is similar to recent work that attempts to generate interpretable hypotheses. Ludwig and Mullainathan (2024) develop a procedure for generating hypotheses from unstructured *image* data. They leave open the question of whether their procedure could be extended to other high-dimensional datasets, including text. Text, however, is quite different from images. For instance, where images are continuous, text is discrete. Changing the color of a single pixel maintains much of the image intact — it resembles the original. In contrast, changing even a letter ("run" to "ran") or removing a punctuation ("Let's eat, Grandma" to "Let's eat Grandma") can change the whole meaning of the sentence. Another important distinction is in the nature of the hypothesized features. Features contained in text are mutable in ways that features contained faces (and images, more generally) tend not to be. That is, hypotheses derived from text should be not only interpretable but also *usable* or able to be applied to new messages written by humans (or LLMs). Our work, therefore, builds on the ideas in Ludwig and Mullainathan (2024), integrating new technologies to generate hypotheses from a different data source.

In particular, we rely on LLMs to generate hypotheses similar to the work of Banker et al. (2023), Manning, Zhu, and Horton (2024), and Zhou et al. (2024). Banker et al. (2023) demonstrates how one could fine-tune LLMs using published and unpublished papers to produce novel psychological hypotheses. Researchers reviewing the hypotheses produced through this process rated them equal quality to human-generated hypotheses in published papers. However, the paper did not formally test the hypotheses as we do here. Manning, Zhu, and Horton (2024) attempts to automate the social scientific process by using LLMs

for generating hypotheses and testing them. The process in that paper relies on structural causal models to propose hypotheses, design experiments, and then test them in a simulated environment. Our process is like theirs in that it is semi-automated, requiring almost no human involvement in generating and interpreting the hypotheses. Our paper deviates from the earlier work in that it starts with real-world data to generate the hypotheses. Then, it tests these hypotheses with real-world outcomes. Zhou et al. (2024) offers an alternative approach to ours, also using the Upworthy data. Their paper introduces an algorithm that starts with an initial set of example hypotheses and leverages an LLM to update and refine the set iteratively. The algorithm employs a reward function inspired by the multi-arm bandit literature to guide the updating process. This method provides an efficient search process. We, instead, opted for an alternative step where we apply each hypothesis to new headlines and use the ML algorithm to estimate a predicted effect for each. Our ranking procedure, therefore, allows us to utilize the full capacity of machine learning algorithms to detect complex patterns in the data (Oquendo et al. 2012; Hutson 2023; Wang et al. 2023) while also taking into account the generalizability of each hypothesis.

Current Paper. 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 marketing hypotheses readily interpretable by humans.

To use it, one needs a corpus of text and access to a large-language model (e.g., OpenAI's GPT, Anthropic's Claude, or Google's Gemini). The primary prompt asks the LLM to produce an insight that captures what changed between a pair of messages and respond with a hypothesis, "Hypothesis: _____ increases [decreases] engagement with a message."[6]

[6]The specific prompts we used are available for download on https://bit.ly/headlines-osf

Notice that this step, the actual *generation* of hypotheses, can be done by nearly anyone with access to a computer today without needing any outcome variable. And although we use pairs of messages from the same A/B test in this paper, this process is not restricted to experimental data. This means that the first step of our process can be applied, for example, by researchers attempting to come up with alternative explanations for a given set of stimuli or firms hoping to learn more about a competitor's marketing strategy. Our ranking and filtering steps leverage off-the-shelf machine learning tools, which others could also use to approximate how different hypotheses, applied to one's data, could affect one's outcome(s) of interest. A company with multiple brands or diverse sets of customers can train an algorithm (or fine-tune an existing one) for each use case and then re-rank the list of hypotheses based on the predictions for each group. Together, this framework integrates existing technologies to provide others with an accessible method for discovery.

As a guide to the rest of the paper, in the next section (Section 2), we describe the application and data used in this paper. Then, we detail our approach for generating hypotheses (Section 3). After generating hypotheses, we test a subset using human coders, both in a holdout set (Section 4) and in an entirely different messaging context (Section 5). In the final section, we discuss the process, results, and implications for future research.

Data and materials related to the studies conducted with human participants are available on the Open Science Framework (OSF; see bit.ly/headlines-osf)

APPLICATION TO ONLINE NEWS HEADLINES

Value of Click-Throughs

To illustrate this procedure, we start with a concrete application: *what linguistic features of a headline lead people to engage with it?* where engagement, conditional on seeing a particular headline, is measured through click-through rates (CTR). This application has broad relevance not only for the consumption of news, but also for other domains where engagement precedes behavior (Petty and Cacioppo 1986). For example, domains such

as advertising (Lee, Hosanagar, and Nair 2018; Phillips and McQuarrie 2010), influencer marketing (Chung, Ding, and Kalra 2023; Cascio Rizzo et al. 2023), constituent services (De La Rosa et al. 2021; Linos et al. 2024), customer communication (Reiff et al. 2023; Kaul et al. 2024), and online education (Nie et al. 2024; Kizilcec, Piech, and Schneider 2013; Deolankar et al. 2024).[7]

Headlines also represent one form of text where the procedure we propose could prove particularly useful. Countless headlines are created and promoted each day. The text is relatively short, typically between 53 and 100 characters, making it easier to parse and compute possible variations. Multiple headlines could be written for the same story, allowing one to study variations while keeping the theme or topic constant. Furthermore, variations matter — different headlines drive different click-through rates; combined with a randomized-controlled trial, this reveals that something about the text influences behavior.

Upworthy Research Archive

Our specific application uses the Upworthy Research Archive (Matias et al. 2021), a dataset of 32,487 randomized trials (A/B tests) conducted by Upworthy.com between 2013 and 2015.[8] Each trial has multiple experimental arms with varying headline text, excerpt, and image. Additional details about this dataset are provided on upworthy.natematias.com.

Data Pre-Processing

2.3.1 Data Cleaning

To clean the data, we applied a few standard steps (e.g., Berger et al. 2020, Table 3). First, we removed one observation where the headline text was missing. Next, we cleaned

[7] While engagement is often a necessary pre-condition for influencing behavior, we recognize it is often not sufficient (e.g., John et al. 2017). Engagement alone cannot, for example, overcome structural barriers (Linos et al. 2022; Thaler and Sunstein 2009)

[8] On July 11, 2024, the authors published a correction to the data, noting problems with the randomization of trials between June 25, 2013 and January 10, 2014. They advise that these trials be omitted when conducting causal analysis. We report results without exclusions in the paper and replicate the main set of tests in the Appendix. See Appendix.

the raw text by removing non-visible characters (e.g., HTML tags) and replacing non-ASCII characters with ASCII equivalents. For cases where two or more treatment arms in a trial had the same headline (e.g., where the image varied), we collapsed the rows into one, summing the number of clicks and impressions.

2.3.2 Data Partitioning

The original data was released already split for exploratory, confirmatory, and testing analysis. However, because the headlines were sometimes reused across trials, the headlines found in one of these original splits sometimes appeared in another. This kind of "leakage" is problematic for machine learning applications and can lead to over-optimistic results (see Kapoor and Narayanan 2023). Therefore, we resampled the complete set into new splits by "component", which we defined so that we could group *trials* with overlapping headlines.[9] This ensured that headlines repeated within and across trials were contained within the same split. The resulting splits include:

- A *training* set (40% of trials; 12,800 trials). This set is further partitioned for training the machine learning model ($N = 11,535$) and tuning the model's hyperparameters ($N = 1,265$). This set is also used for generating hypotheses, described in Section 3.1. Note that we did not use the ML algorithm, which is different to GPT, to generate hypotheses, so we were not concerned about data leakage between these two uses.

- A *morphing* set (10% of trials; 3,366 trials). This set was used to produce counterfactual headlines, described in Section 3.2.

- A *regression* set of (10% of trials; 3,136 trials). This set was used as a validation set for testing the hypotheses we uncovered, described in the Hypothesis Testing section.

[9] It appears that sometimes headlines were reused; for example, imagine Trial 1 tested Headline A against Headline B, Trial 2 tested B against C, and Trial 3 tested C against D. In this case, even though Headline A and D never appeared in the same trial, we assume *something* about them are the same since they are "linked" by Trial 2. To minimize leakage (Kapoor and Narayanan 2022), Trials 1-3 would all be assigned the same component.

We also used this set for benchmarking initial model performance (see the below where we explore the signal in the text).

- A *lock-box* or hold-out set (40% of trials; 13,185 trials). We plan to unlock and analyze this set upon conditional acceptance.[10]

2.3.3 Defining the outcome of interest

The outcome we care about in this application is the click-through rate (CTR). For each headline, the CTR is defined as $\text{CTR} = \frac{\text{Clicks}}{\text{Impressions}}$. To account for variability in CTRs arising from trials of different sizes, we employed a shrinkage procedure toward the overall average CTR. Specifically, we adjusted each headline's CTR by adding the overall mean CTR to the numerator and 1 to the denominator. For any headline H_a, we define this as the smoothed CTR estimate:

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

where $\overline{\text{CTR}}$ was the mean CTR calculated across all headlines. This approach effectively reduced the variance of CTR estimates for headlines with limited data, leveraging the global average as a stabilizing prior. 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 any two headlines H_a and H_b from the same trial.[11] For simplicity, we refer to Smoothed CTR as CTR in the remainder of this paper.

[10] Once the paper is conditionally accepted for publication, we will have headlines from these trials labeled on the final set of hypothesized features in order to replicate our findings. The current manuscript, therefore, serves as a registered report (Nosek and Lakens 2014; Chambers and Tzavella 2021; Urminsky and Dietvorst 2024).

[11] Other reasonable approaches would include using a hierarchical Bayesian model to determine the level of mean shrinkage, or a binomial likelihood to handle trial sizes directly. While these approaches could have been used for modeling CTR, we have chose to use a strategy that we felt was easier to understand and readily generalizes to other settings.

2.3.4 Formatting

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. After cleaning, we collected all pairs of headlines H_a and H_b that appeared in the same trial. Our data partitioning 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. Because a trial with k unique headlines contains $k(k-1)$ unique pairs of headlines (independent of order, e.g., A-B and B-A are two pairs), the number of trials in the pairwise dataset does not match up precisely to the number of trials. For example, 14,729 headlines were dropped because the trial consisted of only one headline (i.e., zero *pairs* of headlines). Note that while this does constitute 45% of the trials in the entire dataset, it makes up only 16% of the headlines in the entire dataset (as these are, by definition, trials with the fewest headlines).

The pairwise dataset splits therefore contain:

- Training set (40%): 112,350 unique headline pairs from 7,048 trials.
- Morphing set (10%): 29,600 unique pairs from 1,807 trials. However, because morphing is done at the headline level, the pairwise dataset is not used for the morphing process.
- Regression set (10%): 27,206 unique pairs from $1,701$ trials. However, for the actual regression step, we will further sample to a single pair from each trial.
- Lock-box set (40%): 112,998 unique pairs from 7,202 unique trials.

Additional Features: Semantic representation, psycholinguistic features, and human labels

2.4.1 Semantic representation

We converted the raw text to its high-dimensional semantic representation or sentence embedding to analyze the text data and train our machine learning algorithm. Sentence

embeddings are vector representations of text, which are both fixed length and numeric, meaning they could be used as inputs to various downstream tasks. We extract sentence embeddings using a pre-trained MPNet model (Song et al. 2020), which converts text into a vector of length 768.[12] It produces this embedding using a transformer architecture: the text is first converted into a sequence of "tokens", each token is mapped to a numeric vector, the starting sequence of vectors are transformed into a sequence of output vectors by several transformer layers in a neural network, and a final output vector is produced by taking the mean value per index across all output vectors. In addition to training the machine learning algorithm, we used these embeddings for other tasks, such as measuring textual similarity and diversity.

2.4.2 Existing Psycholinguistic Features

Absent in the Upworthy data are the explicit hypotheses each trial intended to test. Although we cannot impute specific hypotheses from the past, we could examine how features known to affect behavior influence the click-through rate.

One advantage of starting with the Upworthy dataset is that we are not the first to use it (e.g., Banerjee and Urminsky 2023; Robertson et al. 2023; Gligorić et al. 2023; Rathje et al. 2023; Hopkins, Lelkes, and Wolken 2023; Shulman, Markowitz, and Rogers 2024; Zhou et al. 2024). In particular, Banerjee and Urminsky (2023, "BU") creates a set of features representing psychological constructs deemed relevant by previous research and maps them to each headline.

For our paper, we replicated the work of BU using their materials (posted on `osf.io/826jq` in September 2022) to check that we could reliably extract their features. We combined the outputs from LIWC (Tausczik and Pennebaker 2010), TextAnalyzer (Berger, Sherman, and Ungar 2020), and unique word lists BU compiled from past papers to reconstruct the

[12] We used a version of this model that was additionally fine-tuned as part of a HuggingFace event, see `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

feature set representing the 51 psychological constructs used in BUs analyses.[13] Notably, the set also includes features that Banerjee and Urminsky (2023) have shown affect click-through rates in this dataset, features such as *reading ease*, *numeric reference*, and the use of *visual language*.[14]

2.4.3 Human Labels

Despite the many constructs captured in BU's features, it is possible that some of what is known is not reflected in the features. To address this, we could have enumerated a set of additional constructs based on a broader read of existing research and then created new dictionaries (Humphreys and Wang 2018) or had humans code each headline on each construct. This process is, of course, expensive in both money and time and still runs the risk of not capturing implicit knowledge that humans hold in their heads but cannot articulate (e.g., Malt et al. 1999; Batista et al. 2024).

Instead, we attempt to capture any remaining information by collecting human guesses (Ludwig and Mullainathan 2024). We recruited 303 participants through Prolific (www.prolific.com) and incentivized them to choose from a pair of headlines — written for the same story — which one they believed had performed better in an A/B test.[15]

Each participant completed 10 "training" rounds and 30 "test" rounds, one page at a time. In each round, participants were shown a pair of headlines written for the same story (i.e., from the same A/B test) and asked to choose which headline they thought performed better. During the training rounds, participants received feedback after each guess, where we revealed the correct answer. During the testing rounds, participants received no feedback but were incentivized to answer correctly. Specifically, participants received \$0.25 for selecting the correct answer in at least 17 out of 30 rounds plus an additional \$0.25 for *each* correct

[13] At the time of this writing, textanalyzer.org, was no longer online. However, we collected the features for the full dataset prior to this.

[14] Note that while we will be using the same set of *features*, our data splits and modeling specifications will be different. Therefore, the results in this paper may be appear inconsistent with BU's work.

[15] Half the participants were randomized into a condition that asked them to identify which headline performed *worse*, but there is no evidence this affected performance, ($t(300) = .93$, $p = .36$)

response beyond that.[16] Data and materials are available on OSF.

Participants labeled 1,693 pairs of headlines from the *regression* set, where each pair received a median of five responses (IQR: 4, 7). The interrater reliability, measured by the `ICC1k` variation (Revelle 2007), was above 99% for all labels.

The "known psychological features" gathered using BU's approach and the human labels we collected play an essential role because they provide a baseline of knowledge. We use these features both to approximate how much of the algorithm's predictions are already explained by existing literature in consumer psychology and to estimate how much of the explainable variation in CTR is still left to uncover. By comparing predictions from a model that uses these known psychological features and the human labels to one that combines these known features and our algorithm's predictions, we can evaluate the extent to which the algorithm is uncovering something new versus rediscovering features already known. We can also use the known features to verify post-hoc whether features discovered using our procedure capture any signal above and beyond the existing set.

Is there any signal in the text left to discover?

Textual cues have been shown to motivate engagement. In fact, using this dataset Banerjee and Urminsky (2023) test the effects of more than 50 psychological constructs. Therefore, it is reasonable to ask *is there anything left to discover?* And if so, how much?

For this, we need to compare the signal captured in the models that include the known features and human labels to the model that uses predictions from the machine learning algorithm trained on sentence embeddings (which we describe below).

2.5.1 Predicting click-through rates using known features and human labels

Our main set of baseline models are linear models in which we include predictors for the known psychological features extracted using BU's approach ($\widehat{BU}$), the human labels

[16] Participants were shown headlines from the training set for the training rounds and from the validation-regression set for the testing rounds to avoid leakage here too.

(G), or both combined. We report all our results out of sample to accurately compare the performance of the various models (see Table 3).[17]

We formed the BU predictor as follows. For each of the 51 psychological constructs used in BU's analyses (see Section 2.4.2), 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

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

and estimate the coefficient values on the full training set consisting of 112,350 unique headline pairs. For robustness, we also fit a non-linear model using XGBoost which can better account for complex relationships, such as interactions, between the known features. For this model, we first train the model using a subset of 99,670 pairs of headlines, and use the remaining 12,680 pairs as a tuning population for finding ideal hyperparameter values for the XGBoost model. For both of these models, we then use the estimated coefficients to extract predictions on the regression partition of the pairwise data. These predictions, which we call the "BU predictor" and the "BU predictor (non-linear)", can then be used as features in regressions on the *regression* partition not used in training any of the models.

For the human guess predictor, no additional transformations were required since the guesses collected were already at the pair level. There was also no need to estimate any coefficients on an independent training sample since only a single feature exists. Instead, the feature itself (the proportion of people guessing H_B instead of H_A for the pair of headlines) is used as the human guess predictor.

One way to examine the predictive accuracy of these models is to look at their Adjusted R^2,

[17] Out of sample (OOS) predictions are a standard way to evaluate model performance in machine learning tasks (Mullainathan and Spiess 2017). To obtain an OOS prediction, we first fit a regression on the *training* set and use that model to predict the outcome in the *regression* (validation) set. The known features model, therefore, represents a model in which we regress the CTR on the predicted outcome given the known features.

which captures the proportion of the variation in CTR explained by the predictors. The model with only the human labels has an Adjusted $R^2 = .008$. The model with only the predictor of known psychological features has an Adjusted $R^2 = .042$. The model containing both the known features and the human labels has an Adjusted $R^2 = .049$. When comparing the linear and non-linear BU models, we find that performance tends to improve, but results are qualitatively similar (see Table 3).

Another way to assess performance is to leverage the experimental setup of these data to ask: how well do humans (or a model that includes the known psychological features) pick the winning headline? All three baseline models perform modestly; each picks the winner significantly better than chance (50%). For instance, participants in our labeling study picked the better headline 53.0% of the time (95% CI [50.6%, 55.4%]), better than 50% a random guess. In contrast, the model using only the known psychological features picks the winner 56.9% (95% CI [54.5%, 59.2%]) of the time. Finally, the model with the combined human guesses and the predictor of known features selects the winner 56.8% (95% CI [54.4%, 69.2%]) of the time.

2.5.2 Predicting click-through rates using machine learning algorithm

How well does the machine learning algorithm do? To answer this, we train an ML algorithm to predict $\Delta\text{CTR}_{a,b}$. We employ a Siamese network architecture (Bromley et al. 1993), which in our case works by first transforming headlines H_a and H_b into vectors using a text embedding model (see Section 2.4.1 for an explanation of an example of such a model), then taking the difference between these vectors, and using that difference as input to a linear regression which outputs a single value. To initialize the model, we again use the pre-trained MPNet architecture as a sentence embedding model (Song et al. 2020), and use a single, randomly-initialized, fully-connected linear layer for the regression. 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

call the fully trained model m, and write $\widehat{m}$ for the algorithm's prediction.

To evaluate the performance of the ML algorithm, we again refer to the *regression* set, which contains headlines which the model has not seen during training. As we did above, we consider the proportion of variance explained (Adjusted R^2). Regressing $\Delta\text{CTR}_{a,b}$ on the algorithm's prediction, $\widehat{m}_{a,b}$, results in an Adjusted $R^2 = .130$. Treating the outcome as a binary measure, our algorithm correctly picks the winner 63.9% of the time (compared to 50% guess; 95% CI [61.5%, 66.1%]).

2.5.3 Comparing performance

Using the known features and human labels models as benchmarks, we see that the algorithm provides a significant improvement on every measure. In Table 3, we examine whether the algorithm's prediction captures any signal beyond what is known. Regressing CTR on the algorithm's prediction results in an Adjusted $R^2 = .130$. This is noticeably higher than the model of known features (Adjusted $R^2 = .042$) and human guesses alone (Adjusted $R^2 = .008$). Combining the known features, the human labels, and the algorithm's predictions lifts the Adjusted R^2 to .136, outperforming any of the models on their own.[18] In R^2 terms, the ML algorithm captures $\frac{.130}{.136} = 95.6\%$ of the predictive signal.

A similar pattern could be seen with the binary measure. A model that includes known features, human labels, and the ML algorithm correctly picks the winning headline 62.9% (95% CI [60.6%, 65.2%]). Noticeably better than the model reported above that only includes known features and human labels, but marginally worse than the ML predictions on their own.

To get a sense of how much of the ML prediction is captured by the known features, we regressed $\widehat{m}_{a,b}$ on the known features, $\widehat{BU}$, and the human labels. The Adjusted $R^2 = .197$, suggesting there is a lot in the ML predictions not accounted for in what is already known.

[18] Adding the ML predictor to a model of known features and human labels significantly increases the proportion of variance explained, $F(1, 1689) = 171.14, p < .001$. Adding known features and human labels to the ML-only model also improves the performance of the base model, $F(2, 1689) = 6.70, p = .001$.

In the next section, we explore these predictions further through a series of steps designed to uncover hypotheses from text.

DATA-DRIVEN HYPOTHESIS DISCOVERY

The algorithm is picking up signals that humans fail to see and that past research in marketing and psychology may not yet have discovered. It has, in a sense, made a discovery.

But the discovery remains unknown; the signal is uninterpretable to human researchers. For a particular class of applications where prediction is the main objective (Kleinberg et al. 2015), this may be enough. However, stopping here would leave a lot to be desired when the aim is to uncover novel insights. If the predictors were a set of pre-specified features, one could effectively "read out" the significant predictors and use these to form hypotheses (e.g., Guenoun and Zlatev 2023; Netzer, Lemaire, and Herzenstein 2019; Sheetal, Feng, and Savani 2020, but see Mullainathan and Spiess 2017). In the current approach, the algorithm is trained using embeddings that are uninterpretable to humans, even if one were to use regression methods.

This section describes the steps we devised to recover some of these insights in the form of human-interpretable hypotheses. The goal is to set up a pipeline where one could input text, and the output would be a set of hypotheses to test. Throughout the process, it should be apparent where the algorithm played a role (and where the human did). Our process consists of three steps: generating, ranking, and filtering hypotheses. For each step, we explain the procedure, the output, and any additional checks we did to validate our approach. Figure 1 presents an overview of the framework.

All the code for the steps described are available upon request and will be made public when the paper is accepted for publication. Supplemental materials, such as prompts used with the LLMs and additional figures, are described in the Appendix and posted on OSF (bit.ly/headlines-osf).

Step 1: Generating Hypotheses

Our process begins with generating hypotheses. This step is designed to mimic the behavior of a careful researcher who systematically writes down hypotheses as they comb through a dataset. Row by row, this researcher might examine a pair of messages and jot down an insight based on what they observed. Each insight forms the basis of a "hypothesis" that could later be tested.[19]

Although humans *could*, presumably, do this task, existing evidence suggests they may not do it well. For instance, humans may be limited by what they can see, noticing some changes more easily than others (Adams et al. 2021). They also tend to search for evidence consistent with their preferred hypothesis (Hartzmark, Hirshman, and Imas 2021; Bhatia 2014; Jerath and Ren 2021; Klayman and Ha 1987; Piezunka and Dahlander 2015). Third, the average person may be limited by their belief that their creativity is finite (Lucas and Nordgren 2020), thus undersampling the space of possible discoveries.

LLMs, instead, offer a way to do this task at scale while generating a diverse set of hypotheses (for an analysis comparing the semantic diversity of human-generated versus LLM-generated hypotheses, see Appendix Section 4).[20]

Procedure. We used OpenAI's GPT-4-Turbo ("GPT") to generate a total of 2,100 hypotheses from 2,100 unique pairs of headlines. To select these pairs, we started with the full set of 282,154 headline pairs, where the headlines in a pair were always from the same trial.[21] We then selected the top quartile of the absolute value of $\widehat{m}_{a,b}$ (the entire dataset was used to determine quartiles). Next, we restricted the sample to pairs from the *training*

[19] This is analogous to "divergent" or "fanning out" approaches found in the creativity literature, where the aim is to come up with an expansive list of ideas (e.g., Vanden Bergh, Reid, and Schorin 1983; Kilgour and Koslow 2009; Rosengren et al. 2020; Toubia and Netzer 2017).

[20] While we do not test whether LLMs have the same biased tendencies as humans (though see Hagendorff, Fabi, and Kosinski 2023), we do gather hypotheses from human participants using a similar approach to that used with GPT and find that human hypotheses as diverse as those uncovered by GPT. See Appendix Section 4.

[21] A trial with three arms, {A, B, C} would have six pairs, A-B, A-C, B-C, B-A, C-A, C-B. However, for nearly all our analyses, we randomly drew at most one pair per trial.

set to minimize leakage in later steps.[22] From this subset, we randomly drew one pair per "component" (which also meant at most one pair per trial; see Section 2.3).

Each pair was assigned to one of five model temperatures — .4, .6, .8, 1.0, 1.2 — and one of 288 prompt combinations, minimizing the chance our results were due to a specific prompt.[23] The set of prompts was created by combining 4 "instructions" x 9 "roles" x 8 "hypothesis formats". All prompts began with assigning a *role*; i.e., "Assume you are {`role`}..." where {`role`} was replaced with the assigned role, such as "an editorial strategist focused on digital content optimization." All prompts also included a specific *hypothesis format* that the response should be in. For instance, "produce [an] insight as a single sentence that begins and ends in this exact format..." where the format was always one of eight possible formats that started with "Hypothesis:" and ended with a reference to the feature's effect on engagement. To illustrate, one of the prompt formats reads: "Hypothesis: _____ leads to {`direction`} engagement with a message." where {`direction`} was filled in with the value of another variable — typically "more [less]" or "increases [decreases]" — depending on whether $\widehat{m}_{a,b}$ was positive or negative. Thirdly, prompts varied in the information presented as part of the *instructions*. For example, one of the prompts listed all the constructs from Banerjee and Urminsky (2023) and asked GPT to "look for patterns not yet known." Three of the four instruction templates included the pair of headlines that were meant to be used to generate a hypothesis. The fourth template did not refer to any Upworthy headlines and was included simply as a *Control* to later assess whether hypotheses generated by GPT with access to our dataset differed from those generated by GPT without any specific headlines.[24] Consistent across all prompts was a list of five criteria we wanted each hypothesis to meet. Specifically, each hypothesis should be: (i) clear, (ii) generalizable, (iii) empirically plausible,

[22] Note, this is out of abundance of caution. We use only one pair for each hypothesis. It would have been just as feasible to use one pair to generate hypotheses and use different pairs from the same set for the *ranking* step below.

[23] "Temperature" refers to a parameter, ranging from 0-2, that determines the randomness of responses. Lower values produce more consistent outputs while higher values produce responses that are more diverse or 'creative'. "Prompts" are the conversational input used query an LLM. For more on prompting see www.promptingguide.ai.

[24] 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.

(iv) unidimensional, and (v) usable (see Appendix 1 for more details and OSF for exact wording).

Output. This step produced 2,100 hypotheses, which appear to be clear and coherent. Examples of hypotheses are included in Table 4.

Additional Checks. To further assess the quality, we recruited 79 participants using Prolific (prolific.com) [June 2024] and had them rate 106 hypotheses on several dimensions (e.g., "clarity"; for details, see Appendix 3). On every dimension, the average rating for most hypotheses was above the scale's midpoint. In that survey, we also asked participants whether they believed a given hypothesis could be applied to other contexts, such as "product descriptions" or "billboard advertisements". Every hypothesis rated seemed as though it could be applied to at least one other context (M = 3.72 out of 7; Mdn = 3.76).

At least on the surface, these appear to be good quality hypotheses. Each hypothesis is based on an observation made in the dataset, but it is not clear at this point whether the feature that has been identified is representative of a general insight or specific to the single pair of headlines. Furthermore, this task was mainly a creative exercise. Past work supports our findings that LLMs are capable of producing high-quality hypotheses (Banker et al. 2023), but it remains uncertain whether these hypotheses lead to a larger discovery. The next step takes a hypothesis generated from one observation and applies it to a different set of headlines. It then scores these pairs using the ML algorithm, $\widehat{m}_{a,b}$.

Step 2: Ranking Hypotheses

This step aims to rank-order hypotheses such that insights most likely to affect the outcome (CTR) are prioritized. To achieve this, we want to leverage the ML algorithm to help identify insights that are predicted to have an effect when applied to several new messages.

Ranking is therefore done in two parts, which we refer to as "morphing" and "scoring".

Morphing *applies* each hypothesis to various headlines. Scoring uses the ML algorithm to *predict* the difference in CTR between a morphed headline and the original target headline. Aggregating these "predicted treatment effects" (PTEs) at the hypothesis level provides a measure that incorporates both the ML signal and an element of generalizability (since each hypothesis is applied to a random set of headlines). This measure is then used to rank-order the hypotheses.

3.2.1 Morphing

Procedure. We began with a set of actual headlines from the Upworthy dataset, drawing only from the *morphing* set to avoid overlapping with the headlines used to generate hypotheses. We randomly selected 120 headlines from this subset, each from a unique "component", to be morphed according to each hypothesis.[25]

To generate the morphs, we first matched each of the 2,100 hypotheses to each of the 120 original headlines sampled, producing an intermediate dataset containing 252,000 rows. We then paired each row with one of the three prompts and randomly set the model temperature to .75 or .9. Unlike the prompts used to generate hypotheses, we did not vary the role; instead, all the prompts began with "Assume you are a copywriter for an online news platform. Here are some examples of recent headlines from your company..." We then provided three example headlines randomly drawn from the same subset of headlines (excluding those that belonged to the same component as the one we were morphing) to allow for "few-shot learning" of what headlines look like in this distribution (e.g., Min et al. 2022; Brown et al. 2020). The prompts then explicitly stated, "You need to rewrite Headline A below according to the given instructions. Keep the content of the story as similar as possible. Respond by writing out Headline B." All prompts included an original headline to be rewritten and one hypothesis as the given "instruction" for how the headline should be modified (see Appendix

[25] We decided to use the same set of original headlines for all hypotheses to facilitate a more direct comparison of simulated treatment effects across hypotheses.

1 for more details). The specific wording of the different prompts is available on OSF.[26] We used OpenAI's GPT-4-Turbo to generate the morphed headlines.

Output. The output contained 252,156 new headlines.[27] We then applied a heavy filter to remove anomalous responses from this set. First, we removed 37,430 morphs (17.85 per hypothesis) that were longer than 100 characters. Upworthy seems to have used a strict character limit, so any morphed headline with more than 100 characters was considered "out of distribution" and removed from the set of morphs to be scored. Second, we removed 956 morphs corresponding to the eight *hypotheses* produced by the "Control" prompt; that is, hypotheses produced without referring to any Upworthy headlines. Third, we removed 63,119 morphs (or 30.1 per hypothesis) generated in response to the prompt that instructed GPT to "dial down" the hypothesized feature.[28]

In the end, we had an average of 72.63 morphed headlines (SD = 11.24; Med = 74) for each of the 2,092 hypotheses (excluding those generated using the "Control" prompts), where each morphed headline is associated with an original Upworthy headline and a specific hypothesis. Table 5 provides a set of examples.

Additional Checks. Although LLMs can produce new coherent text from past examples, it is unclear whether they can perform the current task well. This process assumes that morphs are not only coherent but that they also incorporate the hypothesized feature while keeping the content similar to the original. To check whether the morphs varied according

[26]Two of the three prompts produced a single headline. These prompts stated that the aim was to rewrite the headline to maximize engagement, i.e., emphasizing [minimizing] the hypothesized feature when it was hypothesized to *increase* [*decrease*] engagement. The third prompt produced two headlines, one in which the feature was supposed to be "dialed up" and the other in which it was meant to be "dialed down". Although morphed headlines generally varied the hypothesized feature, the direction of the responses was sometimes inconsistent. One reason could be that when a feature is not already present, it is easier to emphasize it than to minimize it.

[27]In generating these morphs, we encountered an error partway through the procedure, causing us to terminate the process early. We originally anticipated approximately 336,000 new headlines (since one-third of the prompts produced two headlines). Nevertheless, we randomized the order in which we generated the morphs, which resulted in a uniform sample of morphs for each hypothesis.

[28]In a post-hoc analysis, morphed headlines corresponding to the "Control" hypotheses and those meant to "dial down" a feature were predicted by the ML algorithm to be consistently worse than the matched original headlines.

to the hypothesized feature, we again used GPT, this time to label how much of each feature was present in a given headline on a scale of 0 to 7. Through this assessment, we saw that 53% of morphs were rated as having *more* of the hypothesized feature compared to the original used to create it (40% of morphs had the same value as the original headline, and only 7% had less of the hypothesized feature). In contrast, only 24% of the morphs had a higher score for the feature of interest when that feature was not one prompted to change, with 57% remaining unchanged and 19% decreasing.

The next part involves using the ML algorithm to score these morphs. However, for the algorithm's predictions to serve as a useful measure, two assumptions must be satisfied: First, the algorithm should be reliable. In this case, it should be able to predict CTR out of sample. Second, the morphs generated by GPT must be similar to those found in the Upworthy dataset on which the algorithm was trained (i.e., they should be "in-distribution").

The first assumption is supported by the results reported above, in the section where we describe the predictive performance of the algorithm. To support the second assumption, we conducted a series of comparisons to check whether the morphs were, in fact, similar to the Upworthy headlines. These checks included a computational check and three pre-registered human assessment experiments. We briefly describe these here, but see the Appendix, Section 3 for more details.

To validate whether the morphs were similar in meaning to the original used to generate them, we gathered the sentence embedding for the morphed headlines and compared them to embeddings of the original. We also compared the original headlines used to generate morphs to other original headlines found in the same trial and original headlines found in other trials. Morphed headlines were semantically more similar to the original Upworthy headline used to generate it than two original Upworthy pairs were to each other; this was true of original-original pairs from the same trial (i.e., for same story) and from separate trials (i.e., different stories).

We also conducted three pre-registered online experiments that asked participants to as-

sess a mix of original and morphed headlines. Methods and results for these experiments are provided in more detail in the Appendix (Section 3). On multiple measures of attitude ($n = 120$), participants considered the morphed headlines equivalent to the Upworthy headlines. When incentivized to identify which headlines were generated by AI ($n = 101$), participants did not think that the morphs or the Upworthy headlines were AI-generated. Finally, when incentivized to identify which headlines were produced by Upworthy ($n = 100$), participants believed writers at Upworthy.com wrote both the original headlines and the morph.[29] Together, these findings suggest that the morphs were similar to the originals used to create them, both in content and style.

3.2.2 Scoring

Procedure. Next, we want to know how each morph might perform against the original headline. We use the ML algorithm (described in Section 2.5.2) to predict the ΔCTR for the morphed headline relative to the original Upworthy headline from which it was generated. These are, in effect, *predicted treatment effects* (PTEs).[30] Here, we are primarily interested in the average effect per hypothesis, so we average PTEs at the hypothesis level.

Output. Scoring produces an average PTE for each hypothesis. This measure captures the effect a hypothesis is predicted to have *on average*; in this case, when applied indiscriminately to a random set of headlines written for different stories. This measure is likely to be a conservative estimate because, in practice, not every hypothesized feature will apply to all topics. For example, "excessive sensationalism" may not be appropriate for sensitive topics.

[29] When comparing the two sets of headlines to each other, the morphed headlines were relatively more likely to be perceived as AI generated ($p = .045$) and relatively less likely to be perceived as written by Upworthy ($p < .001$). In both cases, however, the equivalence tests were also significant given equivalence bounds of half a unit on the scale ($-.5$, $.5$; see also Lakens 2017). For more details, see Appendix Section 3.

[30] Note, that we have no "ground truth" of either CTR (y) or ΔCTR for the morphed headlines, only the ML prediction ($\widehat{m}_{a,b}$).

Additional Checks. As one might expect, not every hypothesis was predicted to have a positive effect. In fact, most hypotheses applied to a random set of headlines were predicted to have a negative effect. Across all hypotheses, the average PTE was $-.00065$ ($SD = .00086$; $Mdn = -.00070$). This result corresponds to a decrease in CTR of approximately 4% relative to the average CTR in the original dataset (see Table 2). The predicted differences at the morph-original headline pair level show that the average PTE was $-.00065$ ($SD = .00279$, $Mdn = -.00061$). Furthermore, the standard deviation of predicted differences at the morph-original headline pair level is .00279, which is 77% the size of the standard deviation in the ML predictor and 44% the size of the standard deviation in ΔCTR. This spread indicates that this process creates morphs with nearly as much variation as we see in predictions from the original dataset.

Interpreting these results is difficult because there is no equivalent measure to compare this to in the original dataset. The average PTE is an attempt to estimate the average effect if many A/B tests were conducted to test the same hypotheses applied to many stories. In the original data, each pair presumably tests something different, if not multiple hypotheses at once (Koning, Hasan, and Chatterji 2022).

Nevertheless, we conducted an additional check, a modified specification curve analysis (Simonsohn, Simmons, and Nelson 2020), where we consider the complete set of average PTEs jointly and ask, how inconsistent are these results with the null distribution? We construct the null distribution for this data by resampling under-the-null. Specifically, we randomly reshuffle the hypotheses, "assigning" them to new morph-original headline pairs, then calculate the average PTEs for each hypothesis. This reshuffling preserves features of the morph-headline pairs; however, now we know the null is true by construction since there is no link between (shuffled) hypotheses and the PTEs. Repeating this exercise many times produces a distribution of average PTEs under the null. Figure 2 shows that the distribution of average PTEs in the observed data is much steeper than those in the null distribution, which suggests that our scoring method is aggregating information about the hypotheses

beyond what we might expect by chance.

Step 3: Filtering Hypotheses

The final step narrows the set of hypotheses.[31] We implemented two data-driven techniques that others can use—the first technique involved clustering similar hypotheses, and the second tested whether the average predicted treatment effect for a given hypothesis was meaningfully different from zero. Beyond these two filters, researchers could apply others. The purpose of the filters is to encourage researchers to transparently document the dimensions they use to select and exclude hypotheses.

3.3.1 Clustered Selection

While this process can generate a large number of hypotheses, many of them are similar. For example, "an element of surprise followed by a cliffhanger" and "an element of suspense and an unexpected outcome" are two of the hypotheses in our sample. Even though each hypothesis originated from a unique pair of headlines, different pairs may have nonetheless varied on the same dimension. An organization like Upworthy may have seen that humor works in one A/B test and decided to try it again in another. It is also possible that copywriters and editors have their own style or deliberate writing strategies that may appear across multiple trials. A third reason for the overlap could be related to the task. By soliciting a single feature that could apply in other contexts, GPT is effectively "forced to choose" a salient feature amongst many that may be present.

Procedure. To account for similar hypotheses, we use a sequential selection strategy. First, we calculate a reference vector for each hypothesis by taking the difference in embedding space between a headline and its associated morph (using the embeddings derived from the fine-tuned embedding model mentioned above, see Section 2.5.2) and averaging the

[31] This is analogous to "convergent" or "fanning in" approaches commonly discussed in the creativity literature, where the aim is to reduce the set of ideas (Banathy 1996; Cropley 2006; Malaie, Spivey, and Marghetis 2024; Toubia and Florès 2007).

differences at the hypothesis level. Next, we order the list of hypotheses according to their average PTE and select the hypothesis at the top of this list. We then use each hypothesis's reference vector to calculate the pairwise distance between the selected hypothesis and every other hypothesis on the list. We exclude hypotheses further down the list if they are similar to one of the ones already selected. For instance, if the first hypothesis suggests that "Framing a message with an element of surprise followed by a cliffhanger" increases engagement, it gets selected because it is first. The second — "incorporating a personal anecdote or reaction increases engagement with a message." — also gets selected for being different from the first. The third, however, refers to an "element of suspense and an unexpected outcome," which is similar to the first, so it is not selected.

Whether or not two hypotheses are similar is determined computationally according to a distance parameter, ε, which we set at .03.[32] This process is repeated for each hypothesis until a set number of hypotheses are selected or the list is exhausted, and every hypothesis is either selected or grouped with one of the ones selected.

Output. The result is a list of hypotheses, arranged by average PTE, with a pairwise distance of at least $\varepsilon = .03$ between their reference vectors. Clustering reduced our hypothesis set from 2,092 to 205 (for selection and respective clusters, see the "hypothesis" dataset on OSF).

3.3.2 Significance Testing

Procedure. With fewer hypotheses, we turned to test whether PTEs were significantly greater than zero.[33] For each of the 205 hypotheses, we conducted a one-sample, one-sided

[32]This parameter determines the distance threshold. Lower values mean hypotheses need to be close together to be counted as the same, which results in fewer observations per cluster and, therefore, more clusters. Higher values cluster more broadly but risk grouping a truly novel or unique hypothesis with more common ones. To select .03, we tried different values and picked one that we believed was selecting a unique set of hypotheses.

[33]We used zero as the null because our algorithm was trained on a normalized outcome measure of differences. Indeed, when we look at the difference in *predicted* CTR among pairs of headlines within the validation set, the mean and median "treatment effect" are both effectively 0 ($< .000001$).

t-test. We then applied a False Discovery Rate (FDR) correction using the `stats` package in R (Benjamini and Hochberg 1995).

Output. Sixteen hypotheses had average PTEs that were positive and significantly greater than zero after correcting for the FDR ($p < .05$); seven at $p < .001$. These are displayed in Table 6.

Reading these hypotheses, it is clear that some overlap despite the clustering in the earlier step. For instance, "the utilization of multimedia elements such as gifs influences engagement with a message" is similar to "incorporating multimedia evidence in a headline results in more engagement with a message."[34] Therefore, we hand-picked four to test (from the set of seven at $p < .001$) out-of-sample, using the *regression* set that was set aside from the start.

We also selected two hypotheses with average PTEs *less* than zero for robustness. For this, we repeated the clustering step, starting from the bottom (most *negative* average PTEs) and then performed the same significance testing procedure on the 212 hypotheses that remained (testing for whether average PTE was *less* than zero). There were 114 hypotheses with average PTEs less than zero ($p < .001$). From these, we hand-picked two.

Discussion

This process has produced a set of interpretable hypotheses which, according to the ML algorithm, are predicted to affect engagement as measured through CTR. Together, these steps offer a systematic approach for aggregating insights from several marketing experiments. Importantly, these steps constitute a data-driven framework for generating hypotheses *before* any confirmatory tests are conducted. Hypothesis testing comes next.

[34] A higher ϵ could have prevented this but at the increased risk of clustering a new hypothesis in with the old.

HYPOTHESIS TESTING USING HOLD-OUT SET

To test our hypotheses — and assuage any concerns of overfitting or p-hacking (Simmons, Nelson, and Simonsohn 2021; Wicherts et al. 2016) — 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 (Kerr 1998; Schaller 2016; Landy et al. 2020). The pre-registration of this analysis is available on AsPredicted.org/S6H_ZPF (#172038).

Procedure

We followed a standard procedure for testing the hypotheses. First, we had humans code different headlines based on the hypothesized feature. Then, we estimated an OLS regression to test whether varying the feature led to a difference in engagement.

The Upworthy dataset has the advantage of being a dataset of randomized experiments. While the experiments were not originally designed to test our hypotheses explicitly, we can still assess whether the features identified in this process affect consumers' propensity to engage with a message. Furthermore, since the pairs of headlines within a trial were written for the same news story, we effectively control for various confounds due to the topic by studying the differences between headlines.

For testing, we used the *regression* set. This set contains pairs from 1,693 trials. Note that none of these trials (or headlines within the trials) overlap with the trials (headlines) used to train the ML algorithm, generate hypotheses, or generate morphs. Where we used this set before was to gather human labels (see Section 2.4.3). We decided to use the same set of 1,693 pairs (3,386 headlines) so that we could also compare whether these new features captured any signal in the humans' intuition from earlier.

We pre-registered our procedure and the six selected hypotheses, noting that while the

experimental data had already been collected, we had not conducted any coding of the hypothesized features in the regression set. Materials for this survey are available on OSF.

The plan was to recruit 800 participants. 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*.[35]

To supplement the main set of tests, we also had participants rate an additional four headlines after the 26 from the validation set. The additional four headlines were drawn from the 117 original headlines used for morphing and the 404 morphed headlines corresponding to each of the six selected hypotheses. These ratings were meant to check whether the morphing procedure morphed the headlines on the feature it was meant to. We noted in the pre-registration that this analysis was intended as exploratory and do not report that analysis here (but see Appendix 3.2.5).

Results

We recruited 800 participants (M_{age} = 41.51, SD = 13.75; 379 Male, 401 Female, 20 Self-Identified; 62.4% White, 14% Black, 9% Latin American, 5.4% Multi-racial, 9.3% all others) on Prolific. Altogether, participants provided 144,000 labels (124,800 for headlines in the regression set, 4,212 for headlines in the morph set, and 14,988 for morphed headlines). Each headline was rated on each feature a median of six times (IQR: 4, 8).

[35]Note that these were rated using a separate evaluation design, where options are presented individually (rather than paired) and evaluated separately (Hsee et al. 1999). We felt this was closer to what one might experience when seeing a headline online.

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

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

where ΔRating represents participants' mean rating of headline H_b minus the mean rating of headline H_a for each hypothesized feature, and β_r is the coefficient related to the rated feature.

Table 7 displays the estimated coefficients for each regression. We rescaled the outcome variable, ΔCTR, by dividing it by the standard deviation of CTR and normalized the hypothesized features to have unit variance. Therefore, one standard deviation increase in ΔRating in any of the hypothesized features produces an estimated change in ΔCTR equivalent to $\hat{\beta}$ times the standard deviation in CTR.

Four of the six hypothesized features were significant predictors ($p < .05$; two $p < .001$) of the outcome. A fifth had a marginal effect ($p = .094$). Of these five, all showed effects in the predicted direction. Furthermore, when we fit a regression with all the predictors included, four of the six are significant ($p < .05$; two $p < .001$), suggesting these features capture distinct signals in the text.[36]

Through this process, we have made a discovery. A question remains as to whether any of these are novel. Statistically, we estimate another set of regressions where we include the prediction from the "known features" derived from Banerjee and Urminsky (2023), which was estimated on the training partition (i.e., $\widehat{BU}$ The results of these regressions are available in Table 9. Two of the six features continue to be significant predictors, *surprise, cliffhanger* and reference to *multimedia evidence* ($ps < .01$). A similar pattern holds when we include all six features plus the $\widehat{BU}$ in a single model, where three of the six features are significant predictors ($p < .05$).

In another specification, we compare a baseline model that regresses ΔCTR on all 51

[36]Though note that *parody* is not significant on its own but is a significant predictor when controlling for the other features, albeit in the direction opposite the one predicted. In contrast, *physical reaction* is significant on its own, but not when adjusting for the other features.

features from Banerjee and Urminsky (2023) (see Equation 3) to one that also includes one of the new features. The models which included ratings for surprise, cliffhanger ($F(1, 1640) = 11.37$, $p < .001$) and multimedia evidence ($F(1, 1640) = 21.95$, $p < .001$), respectively, significantly outperformed the baseline. Together, this suggests that this process has uncovered at least two features that explain the outcome above and beyond what was known.

In a third set of regressions, we check whether these features are capturing any signal from the ML algorithm. We regress $\widehat{m}_{a,b}$ on $\Delta\text{Rating}_{a,b}$. All six features on their own were significant predictors of the ML prediction, $\widehat{m}$, at $p < .001$, with Adjusted R^2 ranging from .006 to .025. A model with all six features produced an Adjusted $R^2 = .098$. The results of these regressions are available in Table 10.

As a follow-up analysis, we consider the effect on CTR if we treat headlines within a pair as being *higher* or *lower* on a given feature. Doing so allows us to analyze the data "as-if" we had pre-tested a set of headlines written for the same story and assigned the ones with the higher feature to one group and the lower feature to another. We could then ask: what is the average effect on CTR of seeing a headline from the high feature condition compared to the low? Figure 3 displays the average change in CTR from moving from a headline lower on the feature to one that was higher. The largest effect comes from referencing multimedia evidence, which increases CTR by 4.2% on average compared to the average CTR of the "low feature" group.

Discussion

These results provide causal evidence against the null for several hypotheses. Using data from nearly 2,000 digital experiments, we find that varying the feature between the two headlines led to a significant difference in CTR in at least four of the six hypothesized features we tested (a fifth showed a marginal effect). These are, however, only six from a set of dozens of hypotheses generated. In the Appendix, we provide results for the same set of

tests conducted on a random set of 400 hypotheses.[37] The pattern of results is similar; among the top decile of hypotheses predicted through our ranking procedure to have meaningful effects on CTR, we find evidence in support of 75% of them ($ps < .05$ after FDR correction; 28% at $ps < .001$).

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 Banerjee and Urminsky (2023). 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.

We explore the generalizability of these hypotheses in the next section and return to the question of whether they are interesting in the General Discussion, where we consider striking a balance between basic theoretical insights and more applied insights.

GENERALIZING HYPOTHESES TO NEW CONTEXTS

Behavioral interventions that work in one context may not produce the same effects in another (Goswami and Urminsky 2022; Landy et al. 2020). Nevertheless, researchers are often interested in coming up with and testing hypotheses that are broadly applicable. Therefore, to investigate the generalizability of our hypotheses, we partnered with an online entertainment company to test whether the hypotheses generated in one context (Upworthy headlines) could inform interventions in another entirely different context.[38]

[37] As described in the Appendix, we used ratings collected using GPT instead of humans for this analysis. Also in the Appendix is an analysis comparing GPT to human ratings (see also Rathje et al. 2023).

[38] In the Appendix, we provide tests for a third dataset consisting of A/B tests conducted by a non-profit organization focused on progressive outreach. The context is sufficiently different from either of the two reported here that we felt it would require a longer discussion to contextualize the hypotheses and the results (Goswami and Urminsky 2020, 2022; Markowitz and Shulman 2021).

Data

The data from the online entertainment company consists of social media posts. The posts resemble the Upworthy headlines in style and content. Where Upworthy headlines tend to focus on uplifting stories, the content from this company emphasizes popular culture and entertainment, including sports, movies, music, and celebrities. 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. We partitioned the data following a similar process we used for the Upworthy data; 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 applied the same smoothing function used above (see Equation 1). More details about this dataset, including summary statistics of key variables and descriptions of the secondary outcomes we examined, are available in the Appendix.

Procedure

For this analysis, we followed a procedure similar to the one above (see Hypothesis Testing section). Again, we pre-registered our procedure and hypotheses on AsPredicted.org/FN5_CNG (#181144), keeping the same six hypotheses uncovered and tested above. We kept the direction consistent for simplicity, but note here that behavioral interventions often have different effects across different people and contexts (Goswami and Urminsky 2022; Markowitz and Shulman 2021). Our primary interest was in seeing whether the hypothesized feature(s) influenced the outcome, *click-through rate* (CTR). That is, whether hypotheses generated in one dataset could predict outcomes in another.

We planned to recruit 900 participants to rate the 5,077 social media posts. The survey format was the same as the task above. Each participant saw 30 messages, each on a separate page, randomly drawn from a set of 5,077. For each message, participants were asked to "select the level which each trait is featured in this headline, from '1 (Low)' to '7 (High)'."

Results

Demographic details for this study are reported in the Appendix.

We estimated OLS regressions following a similar specification in Equation 4 to test each of the six hypotheses. 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.

The results are displayed visually in Figure 4 together with the results from the out-of-sample Upworthy tests for comparison (for table of coefficients, see Appendix). In the social media posts, 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.

Discussion

We tested hypotheses generated in one context using a second dataset. The aim was to explore whether the hypotheses generated through our framework were specific to a time and place or whether they might generalize to other contexts. Of the six hypotheses we selected for testing, four appear to predict the effects of language in another context.

GENERAL DISCUSSION

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.

We demonstrate how to use this framework by applying it to a specific case: uncovering features of language present in a message that affect consumers' propensity to engage with it. Through our process, we produced dozens of hypotheses and selected six to test in this paper.[39] Four were predicted to increase engagement: 1) framing a message with an element of surprise followed by a cliffhanger, 2) incorporating a concept of parody, 3) incorporating multimedia evidence, and 4) describing physical reactions. Two were predicted to decrease engagement: 5) shortening and simplifying phrases and 6) focusing on positive aspects of human behavior. When we tested these hypotheses out of sample, using pairs of headlines from a hold-out set of A/B tests, we found causal evidence supporting five out of six hypotheses.

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.

Beyond the specific hypotheses uncovered for this application, this paper illustrates how marketing researchers and organizations could use these tools to generate new insights into what drives consumers' behavior. We consider a few cases here for how and when others might use these tools. In all three cases, one needs a corpus of text. In two of the three cases, one needs an outcome variable that they can statistically predict. We focused on CTR because it is a meaningful engagement metric for media companies whose revenue models often rely on page visits. However, other companies will have other metrics they care about, and our framework easily accommodates this.

The first use case looks similar to ours. It starts with a large corpus of text with a corre-

[39] See OSF for the complete set of hypotheses. See Appendix Section 5 for results of out-of-sample tests using GPT ratings.

sponding outcome of interest. Through the same steps, one could generate new hypotheses, rank them using an algorithm trained on their data, and filter the hypotheses for testing. Researchers interested in studying consumer language could also apply this approach to existing datasets to explore new research directions.[40]

As we have demonstrated, this approach is useful for aggregating insights across many messages or A/B tests. Marketing teams at companies like Netflix (2022), Uber (2022), and Upworthy (Matias et al. 2021) are continuously running A/B tests, often testing the effects of various messages on consumers' choices (Lee, Hosanagar, and Nair 2018; Angelopoulos, Lee, and Misra 2024). However, unlike academic studies that attempt to conduct carefully designed, high-powered experiments to test a hypothesis tied to a specific theory, experiments done in organizations often throw in the kitchen sink in an effort to optimize products and services quickly (Koning, Hasan, and Chatterji 2022). Moreover, many organizations adopt a "lean" strategy, running many smaller (often underpowered) experiments that allow them to explore a broader range of ideas (Azevedo et al. 2020). Experiments are often local, divorced from existing theory (Browne and Jones 2017) and independent of other teams in the organization (Tang et al. 2010; Koçak, Levinthal, and Puranam 2023; Hern 2018). This "decentralized" strategy can promote discovery through exploration but comes at the cost of aggregated learning (Fang, Lee, and Schilling 2010; Siggelkow and Levinthal 2003; Koçak, Levinthal, and Puranam 2023; Wu, Wang, and Evans 2019; Sah and Stiglitz 1986). Together, these strategies can help organizations learn "what" works, but often at the cost of learning "why." Nevertheless, it is possible that, through many messages, specific patterns emerge. Our method allows organizations and marketing researchers to leverage this data to extract plausible "whys."

The second use case involves generating and ranking hypotheses on independent datasets or disparate outcomes. Companies with multiple brands or several customer segments could

[40]Researchers looking for publicly available text data can find several sources at i) www.english-corpora.org, ii) index.quantumstat.com and iii) convokit.cornell.edu (Chang et al. 2020). For examples of text data from specific organizations, see Upworthy (Matias et al. 2021), Amazon reviews (Hou et al. 2024), and Yelp.com (www.yelp.com/dataset/).

generate hypotheses using messages in one domain (or from across them all) while ranking them separately according to each context or group. For this, they will need a set of example messages from the relevant domains for morphing and domain-specific algorithms for scoring. Notice that access to the outcome variable is not required for generating hypotheses, meaning an organization could just as easily use messages from elsewhere to generate hypotheses and rank the hypotheses using an internal algorithm (for which they do have access to the outcome). An advantage of mixing and matching datasets for different steps is that it allows for discovering insights not available in the target dataset. Data-driven approaches, such as the one proposed here, will only uncover insights present in the data. However, one can expand the space of possible discoveries by generating hypotheses in one domain and then using them to morph and score messages in another.

The third use case involves generating hypotheses on messages without access to the outcome variable. Without an outcome variable, one cannot train a machine-learning algorithm, as we did here. While this limits one's ability to rank hypotheses based on a predicted effect, it does not preclude drawing insights from existing messages in the form of hypotheses using an LLM. For example, companies looking to glean insights into a competitor's strategy could generate hypotheses from others' above-the-line marketing campaigns, such as promotional emails and social media posts. Alternatively, companies could generate hypotheses from customer complaints or product reviews. Finally, companies brainstorming new ideas for what to test can use this approach to broaden their options using data from elsewhere. These uses are similar to others where text mining is used for market research (see also Netzer et al. 2012; Hewett et al. 2016; Brand, Israeli, and Ngwe 2023; Hauser, Tellis, and Griffin 2006). In academic settings, researchers (and reviewers) could generate hypotheses using a set of stimuli to explore potential confounds or alternative explanations.

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). The benefit of data-driven approaches is that one starts with an

observation. The effect is there; subsequent questions focus on describing how general it is and examining its causes and consequences. 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; even though the feature significantly predicts the outcome in two domains, more research could help to reconcile the fact that the observed effect is in opposite directions. Furthermore, related insights run the risk of "talking past" each other without a formal theory to connect them. Science requires both (Mortensen and Cialdini 2010; Alba 2012; Lynch et al. 2012) and, in fact, in marketing, both approaches are regularly used Janiszewski and van Osselaer (2021). The framework presented here adds to the toolkit of data-driven approaches. At the same time, 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.[41] There were two dimensions we attempted to balance. The first dimension maps onto discussions about "basic", or theoretical, insights versus insights that are more "applied," or substantive (e.g., Lynch et al. 2012; Blanchard et al. 2022). The hypotheses generated through this process are more substantive than theoretical — this was intentional. We aimed to generate hypotheses that were "empirically plausible" (Ludwig and Mullainathan 2024) or able to be observed in the data without needing additional background knowledge. While off-the-shelf LLMs could conceivably draw on existing knowledge to produce more theoretically rich hypotheses (Yiu, Kosoy, and Gopnik 2023), leaning into this would increase the chance the LLMs "hallucinated" or drew insights from a world model different from our own (Vafa

[41] For some prompts, we even included these examples explicitly, for instance, "merely measuring intent will increase subsequent purchase behavior" (Morwitz, Johnson, and Schmittlein 1993) and "how people monitor their progress toward goal completion influences their motivation" (Koo and Fishbach 2012). Another prompt included definitions provided in Banerjee and Urminsky (2023) for specific constructs, e.g., "Location: About the location or position of something or about location in general."

et al. 2024). The second dimension is one of complexity. As with any pair of messages, most pairs of headlines vary several things at once. It is conceivable that the hypotheses generated could 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). However, more complex psychological hypotheses are theoretically possible (Adolphs et al. 2016; Peterson et al. 2021). An early prompt we tried produced an example of what a more complex hypothesis might look like: "begin with a positive emotional state and then transition to a negative one, depicting a journey of emotional upheaval." 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. Nevertheless, as a result, the hypotheses are simpler to read, easier to test with available data, and written at a level that others can use.

The framework in this paper is intended to be built on. Some of the technologies used in this paper are still in their infancy but are developing rapidly. The uncertainty of this development makes it difficult to predict what this will mean for the pipeline described above. However, we could comment on how our process might be extended with existing applications. Two extensions, in particular, seem ripe for future work. One extension is to our approach to generating hypotheses (Step 1). The current approach uses GPT-4-Turbo as it comes; however, one could fine-tune these models further to specific domains. Banker et al. (2023) does this, using abstracts in published and unpublished psychology papers. Extending our pipeline using a similar approach could provide theoretically richer insights while helping to identify gaps in existing knowledge (see also Sourati and Evans 2023). The second extension is to the ranking step (Step 2); the current approach uses a machine-learning prediction to score and rank-order hypotheses. Iterating through this process using an algorithm like the one used in Zhou et al. (2024) could facilitate the search process. Other methods of selection could also help, such as using a panel of everyday consumers or domain experts to evaluate the set of hypotheses (e.g., Toubia and Florès 2007; Otis 2022;

DellaVigna, Pope, and Vivalt 2019; Camerer et al. 2016; Landy et al. 2020).

In this paper, we have used relatively short bits of text, ranging from a single word to a couple of sentences. LLMs' current ability to handle much longer texts suggests one may be able to use resumes or application letters, for instance, to generate new insights into who gets hired or admitted to specific roles. However, whether the insights produced will be useful remains an open empirical question.

Like the work of those who came before us, we hope this paper engenders more innovation in methods that produce interpretable hypotheses from unstructured data sources, especially text. There is still much to learn about how language shapes behavior. This paper provides a framework to help convert language in everyday text to interpretable marketing insights. While this framework can augment the current approach to generating hypotheses, it does not preclude the need for careful testing. It would be remiss to confuse one for the other. This paper offers structure to the former.

Table 1: Counts for Headline Data

	Splits				
	Train	*Morph*	*Regression*	*Lock-Box*	**Total**
Headline-Level					
Total Headlines	36173	9434	8779	36866	**91252**
Unique Headlines	25759	6673	6282	26324	**64958**
Pair-Level					
Total Pairs	112350	29600	27206	112998	**282154**
Unique Pairs	56175	14800	13603	56499	**141077**
Unique Headlines	25520	6612	6220	26084	**64377**
Trial-Level					
Total Trials	12800	3366	3136	13185	**32487**
Total Components	3438	869	837	3609	**8753**
Average # of Headlines	2.83	2.80	2.80	2.80	**2.81**

Note: Here we have combined treatment arms within a trial that had the same headline. Trials with only one headline are dropped from the Pair-Level data.

EXHIBITS

Table 2: Summary Statistics for Headline Data

	Mean	SD	Median
Headline-Level			
CTR (Raw)	0.015	0.012	0.011
CTR (Smoothed)	0.015	0.012	0.011
Clicks	89.671	130.364	45.000
Impressions	5898.752	6015.678	3560.000
Character Count	81.024	14.902	84.000
Word Count	16.078	3.370	16.0
Absolute Value of Pair-Level Differences			
ΔCTR (Raw)	0.004	0.005	0.003
ΔCTR (Smoothed)	0.004	0.005	0.003
ΔClicks	16.542	25.295	10.000
ΔImpressions	233.771	1142.076	67.000
ΔCharacter Count	12.859	11.351	10.000
ΔWord Count	2.995	2.532	2.000
Trial-Level Averages			
Mean CTR (Raw)	0.016	0.012	0.013
SD CTR (Raw)	0.004	0.003	0.003
Mean CTR (Smoothed)	0.016	0.012	0.013
SD CTR (Smoothed)	0.004	0.003	0.003
Mean Clicks	161.458	186.665	91.000
Mean Impressions	9760.456	8404.783	6303.333
Mean Character Count	81.291	12.457	83.000
Mean Word Count	16.125	2.840	16.000

Note: Here we have combined treatment arms within a trial that had the same headline.

Table 3: Out-of-sample regression performances with and without ML model predictions included as a predictor

	Adj R^2		Binary accuracy		Binary AUC	
Baseline features	No ML	Plus ML	No ML	Plus ML	No ML	Plus ML
B&U	0.042	0.133	0.569	0.636	0.596	0.688
B&U (non-linear)	0.041	0.134	0.564	0.639	0.591	0.688
Human guess	0.008	0.134	0.530	0.632	0.550	0.690
B&U + Human guess	0.049	0.136	0.568	0.629	0.608	0.692
ML only	—	0.130	—	0.639	—	0.687

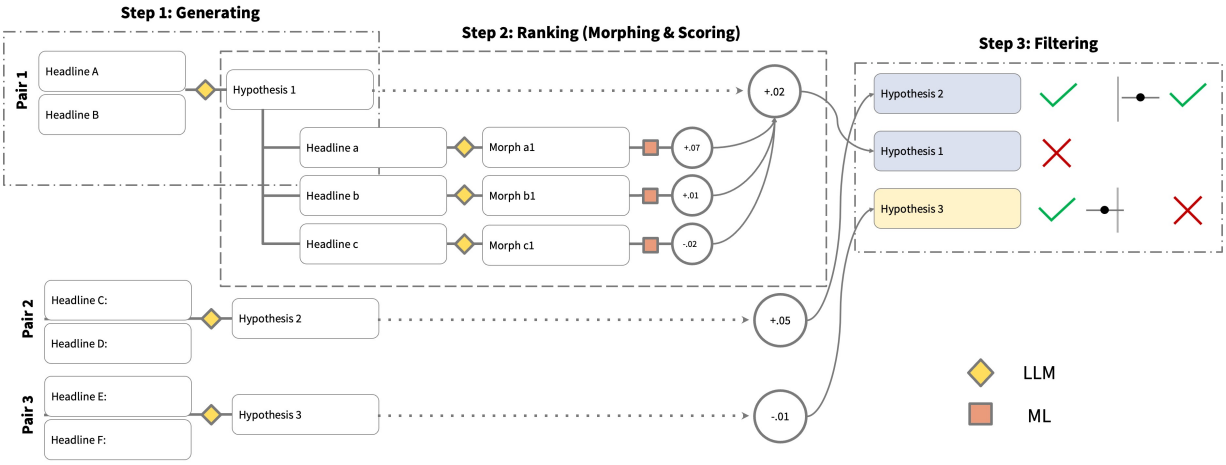

Figure 1: Overview of steps for generating and selecting hypotheses

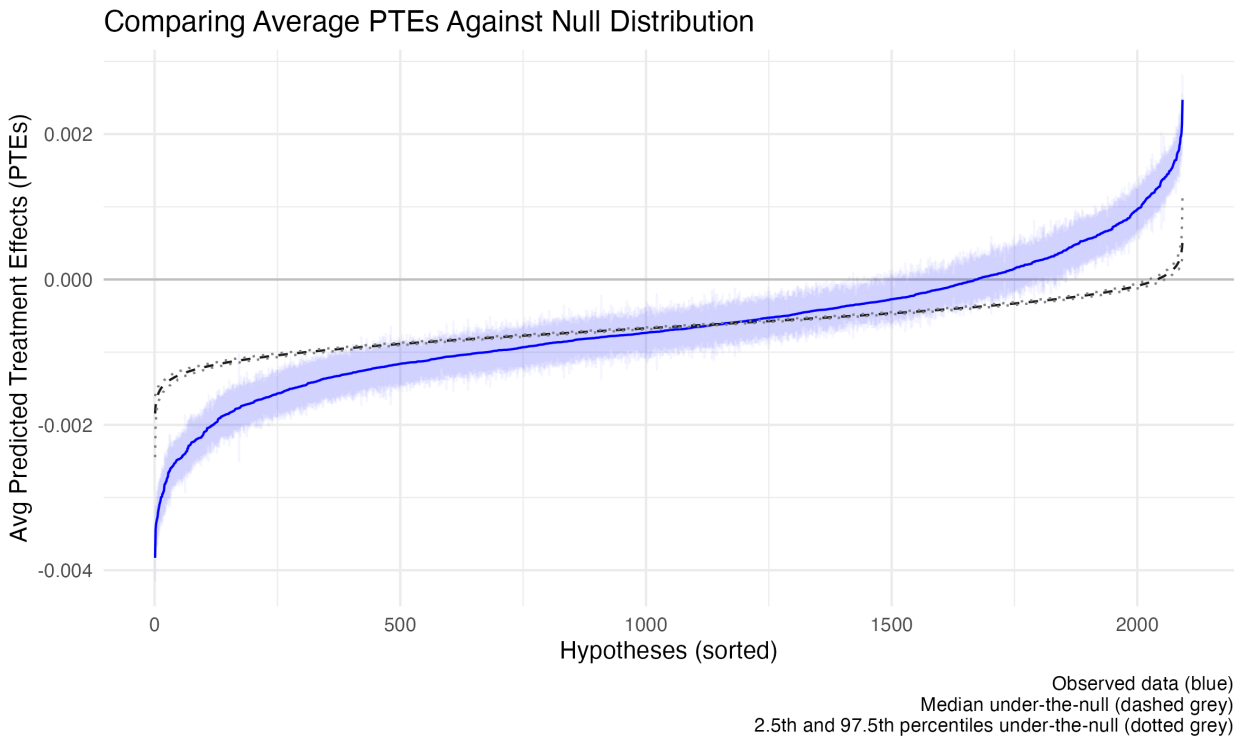

Figure 2: How do average PTEs compare to null distribution? Observed data is steeper than distribution under-the-null, suggesting the predictions conditional on specific hypotheses are more similar than what we might expect by chance.

Table 4: 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 5: 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.

Table 6: Generated Hypotheses Predicted to Have a Positive Effect on Engagement

No.	Short Name	Hypotheses	Selected
1	Surprise, Cliffhanger	Framing a message with an **element of surprise followed by a cliffhanger** makes people more likely to engage with a message.	✓
2	Surprise + Emotion	Incorporating a **narrative of surprise and emotional reaction** in messaging makes people more likely to engage with a message.	
3	Personal Anecdote	Incorporating a **personal anecdote or reaction** increases engagement with a message.	
4	Curiosity + Questions	Incorporating **elements of curiosity through direct questions or incomplete revelations** increases engagement with a message.	
5	Multimedia (e.g., GIFS)	The utilization of **multimedia elements such as GIFs** influences engagement with a message.	
6	Parody	Incorporating the **concept of parody** makes people more likely to engage with a message.	✓
7	Multimedia	Incorporating **multimedia evidence** in a headline results in more engagement with a message.	✓
8	Specific Incident + Question	Introducing a **specific incident and posing a direct question** leads to more engagement with a message.	
9	Conversational Language + Confident Prediction	Using **conversational language and making a confident prediction about audience enjoyment** leads to more engagement.	
10	Physical Reactions	Describing **physical reactions** makes a message more engaging.	✓
11	Direct & Provocative Language	Using **direct addressing and provocative language** influences engagement.	
12	Taboo Topics + Curiosity	Incorporating **taboo topics and invoking curiosity** leads to more engagement.	
13	First-Person	Using a **first-person narrative** affects engagement with a message.	
14	General Accusation to Specific Anecdote	Shifting the focus of a message **from a general accusation to a specific anecdote** affects engagement with a message.	
15	Visual Lang + Curiousity	Incorporating **visual elements and invoking curiosity** leads to more engagement with a message.	
16	Mistake + Long-Term Consequence	Using a narrative that includes **a mistake and its long-term consequences** makes people more likely to engage with a message.	

Note: Two additional hypotheses, predicted to have a negative effect, were also selected for testing. 1) *Shortening and simplifying phrases affects engagement with a message.* and 2) *Focusing on positive aspects of human behavior affects engagement with a message.*

Table 7: How well do features explain pairwise difference in click-through?

	Dependent variable: ΔCTR						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Surprise, Cliffhanger	0.055***						0.056***
	(0.014)						(0.014)
Parody		−0.014					−0.036*
		(0.014)					(0.014)
Multimedia			0.063***				0.067***
			(0.014)				(0.015)
Physical Reactions				0.029*			0.019
				(0.014)			(0.015)
Short, Simple Phrases					−0.023†		−0.024†
					(0.014)		(0.014)
Positive Human Behavior						−0.027*	−0.047**
						(0.014)	(0.014)
Constant	−0.010	−0.009	−0.010	−0.008	−0.009	−0.010	−0.012
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.013)
Observations	1,693	1,693	1,693	1,693	1,693	1,693	1,693
R^2	0.010	0.001	0.013	0.003	0.002	0.002	0.032
Adjusted R^2	0.009	0.0001	0.012	0.002	0.001	0.002	0.029

Note: †p<0.10; *p<0.05; **p<0.01; ***p<0.001. To make coefficients interpretable, we have scaled the outcome variable, ΔCTR, by dividing by the standard deviation of CTR (.0119), and scaled each of the hypothesized features to have unit variance. Hence, a one standard deviation increase in any of the hypothesized features produces a change in ΔCTR equal to $\widehat{\beta}$ times the standard deviation in CTR.

Table 8: Pairs of Upworthy headlines in which feature differences are most pronounced

Feature	Rated Lower	Rated Higher
Surprise, Cliffhanger	"It Was Hard To Hide The Bruises When She Took A Photo Every Single Day For A Year"	"She Took A Photo Every Day During The Worst Year Of Her Life. Here's What It Looked Like."
Parody	"A Real Before And After View Of A Day With A Homeless Man"	"1 Man 2 Suits. See How Clothes Really Make The Man ... Beg Harder"
Multimedia	"A story about tides that's not about global warming"	"A time lapse video that allows you to watch the earth breath. It's wonderful."
Physical Reactions	"Questions You Should Never Ask Your Biracial Bestie Unless You Want These Priceless Responses"	"Questions To Never Ask Your Biracial Bestie Unless You Like To See These Blank Faces"
Short, Simple Phrases	"Mary Engelbreit: No One Should Have To Teach Their Children This In The USA"	"4 Words That Leave Me With No Words"
Positive Human Behavior	"I Don't Know About You, But If This Fierce Mom Spoke These Words To My Face, I'd Tremble A Bit"	"A Mom Spoke About Her Baby For 4 Minutes. Over 100 World Leaders Leapt To Their Feet In Applause."

Table 9: How well do features explain pairwise difference in click-through when adjusting for the BU prediction?

	Dependent variable: ΔCTR						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Surprise, Cliffhanger	0.044**						0.043**
	(0.013)						(0.014)
Parody		−0.008					−0.027†
		(0.013)					(0.014)
Multimedia			0.057***				0.063***
			(0.013)				(0.014)
Physical Reactions				0.021			0.012
				(0.013)			(0.015)
Short, Simple Phrases					−0.006		−0.007
					(0.013)		(0.014)
Positive Human Behavior						−0.024†	−0.044**
						(0.013)	(0.014)
BU predictor (linear)	0.908***	0.939***	0.918***	0.931***	0.936***	0.938***	0.844***
	(0.109)	(0.109)	(0.108)	(0.109)	(0.110)	(0.109)	(0.110)
Constant	−0.010	−0.010	−0.011	−0.009	−0.009	−0.010	−0.012
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Observations	1,693	1,693	1,693	1,693	1,693	1,693	1,693
R^2	0.049	0.043	0.053	0.044	0.043	0.044	0.065
Adjusted R^2	0.047	0.042	0.052	0.043	0.041	0.043	0.061

Note: †p<0.10; *p<0.05; **p<0.01; ***p<0.001

To make coefficients interpretable, we have scaled the outcome variable, ΔCTR, by dividing by the standard deviation of CTR (0.0119), and scaled each of the hypothesized features to have unit variance. Hence, a one standard deviation increase in any of the hypothesized features produces a change in ΔCTR equal to $\widehat{\beta}$ times the standard deviation in CTR.

Table 10: How well do features explain the ML model predictions?

	Dependent variable: $\widehat{m}$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Surprise, Cliffhanger	0.047***						0.051***
	(0.007)						(0.007)
Parody		−0.026***					−0.047***
		(0.007)					(0.007)
Multimedia			0.042***				0.041***
			(0.007)				(0.007)
Physical Reactions				0.036***			0.038***
				(0.007)			(0.008)
Short, Simple Phrases					−0.035***		−0.036***
					(0.007)		(0.007)
Positive Human Behavior						−0.024***	−0.038***
						(0.007)	(0.007)
Constant	0.009	0.009	0.008	0.010	0.009	0.009	0.007
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Observations	1,693	1,693	1,693	1,693	1,693	1,693	1,693
R^2	0.025	0.008	0.020	0.015	0.014	0.007	0.102
Adjusted R^2	0.025	0.007	0.019	0.014	0.013	0.006	0.098

Note: †p<0.10; *p<0.05; **p<0.01; ***p<0.001

To make coefficients interpretable, we have scaled the outcome variable, ΔCTR, by dividing by the standard deviation of CTR (.0119), and scaled each of the hypothesized features to have unit variance. Hence, a one standard deviation increase in any of the hypothesized features produces a change in ΔCTR equal to $\widehat{\beta}$ times the standard deviation in CTR.

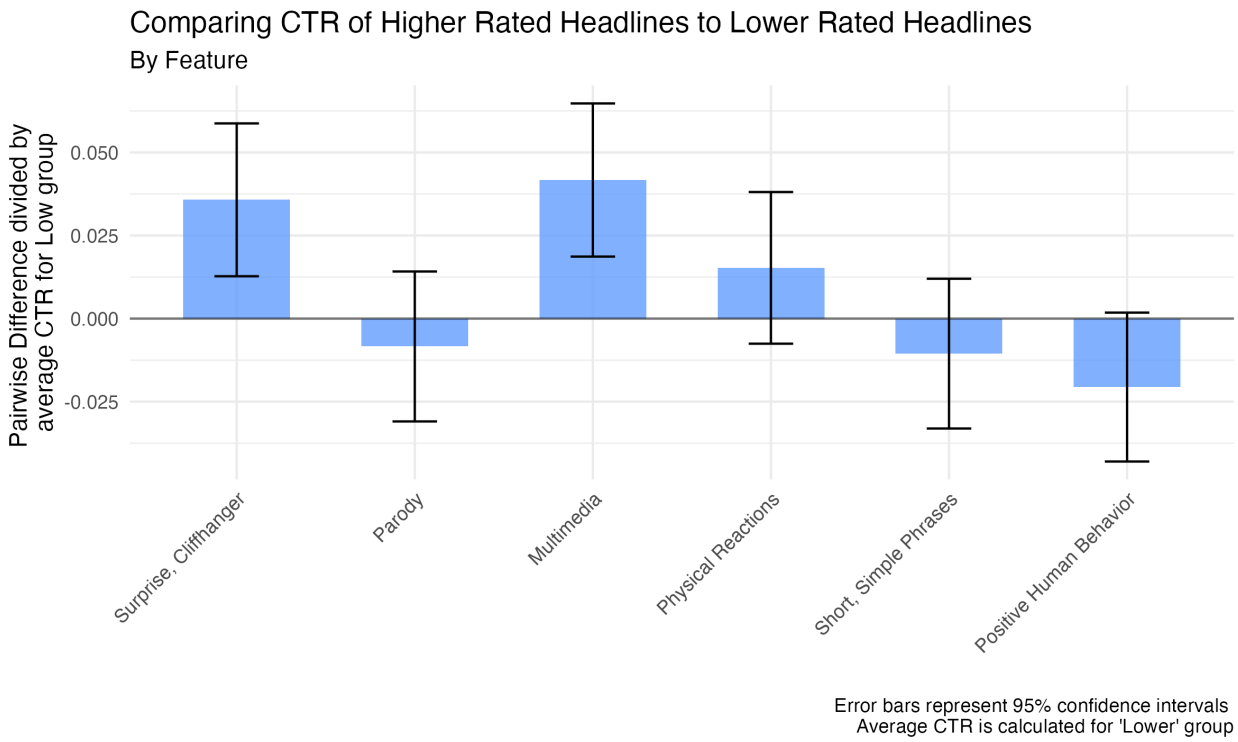

Figure 3: Mean change in CTR for headlines conditional on whether they had the larger or smaller labelled feature value within a pair, as a fraction of the average CTR of the lower-rated group. Bands represent 95% confidence intervals for the change.

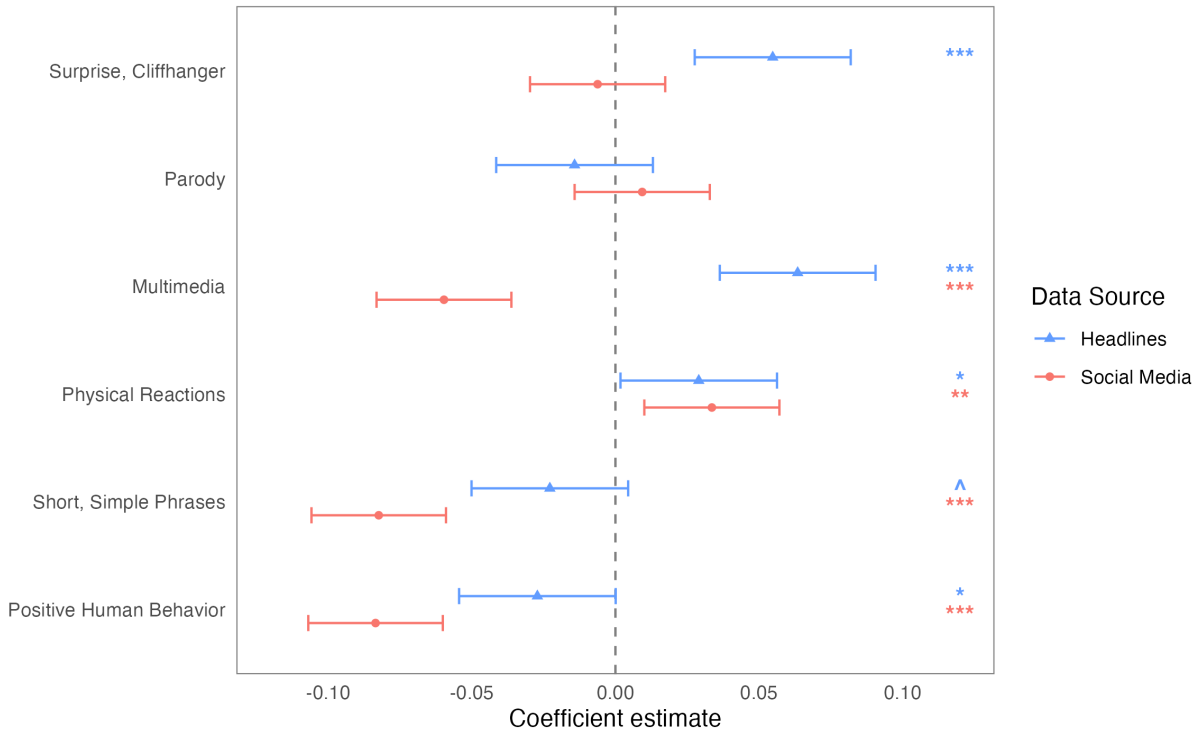

Figure 4: Coefficient estimates for all hypotheses, across two datasets, where the outcome is ΔCTR for Headlines and CTR for Social Media. Values shown are pulled from regressions, where a one unit change in variable corresponds to a $\widehat{\beta}$-unit change in the standard deviation of CTR for the respective dataset.

REFERENCES

Adams, Gabrielle S., Benjamin A. Converse, Andrew H. Hales, and Leidy E. Klotz (2021), "People systematically overlook subtractive changes," *Nature*, 592 (7853), 258–261 https://www.nature.com/articles/s41586-021-03380-y.

Adolphs, Ralph, Lauri Nummenmaa, Alexander Todorov, and James V. Haxby (2016), "Data-driven approaches in the investigation of social perception," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371 (1693), 20150367 https://royalsocietypublishing.org/doi/full/10.1098/rstb.2015.0367.

Aka, Ada, Sudeep Bhatia, and John McCoy (2023), "Semantic determinants of memorability," *Cognition*, 239, 105497 https://www.sciencedirect.com/science/article/pii/S0010027723001312.

Alba, Joseph W. (2012), "In Defense of Bumbling," *Journal of Consumer Research*, 38 (6), 981–987 https://doi.org/10.1086/661230.

Angelopoulos, Panagiotis, Kevin Lee, and Sanjog Misra "Value Aligned Large Language Models," (2024) https://papers.ssrn.com/abstract=4781850.

Azevedo, Eduardo M., Alex Deng, José Luis Montiel Olea, Justin Rao, and E. Glen Weyl (2020), "A/B Testing with Fat Tails," *Journal of Political Economy*, 128 (12), 4614–000 https://doi.org/10.1086/710607.

Banathy, Bela H. (1996), *Designing Social Systems in a Changing World* Contemporary Systems Thinking, Boston, MA: Springer US, http://link.springer.com/10.1007/978-1-4757-9981-1.

Banerjee, Akshina and Oleg Urminsky "The Language That Drives Engagement: A Systematic Large-scale Analysis of Headline Experiments.," (2023) https://dx.doi.org/10.2139/ssrn.3770366.

Banker, Sachin, Promothesh Chatterjee, Himanshu Mishra, and Arul Mishra "Machine-Assisted Social Psychology Hypothesis Generation," (2023) https://doi.org/10.31234/osf.io/kv6f7.

Batista, Rafael M., Juliana Schroeder, Aastha Mittal, and Sendhil Mullainathan "Misarticulation: Why We Sometimes Feel Our Words Don't Match Our Thoughts," (2024) https://dx.doi.org/10.2139/ssrn.4687986.

Benjamini, Yoav and Yosef Hochberg (1995), "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 https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Berger, Jonah, Ashlee Humphreys, Stephan Ludwig, Wendy W. Moe, Oded Netzer, and David A. Schweidel (2020), "Uniting the Tribes: Using Text for Marketing Insight," *Journal of Marketing*, 84 (1), 1–25 https://doi.org/10.1177/0022242919873106.

Berger, Jonah, Yoon Duk Kim, and Robert Meyer (2021), "What Makes Content Engaging? How Emotional Dynamics Shape Success," *Journal of Consumer Research*, 48 (2), 235–250 https://doi.org/10.1093/jcr/ucab010.

Berger, Jonah, Wendy W. Moe, and David A. Schweidel (2023), "What Holds Attention? Linguistic Drivers of Engagement," *Journal of Marketing*, 87 (5), 793–809 https://doi.org/10.1177/00222429231152880.

Berger, Jonah and Grant Packard (2023), "Wisdom from words: The psychology of consumer language," *Consumer Psychology Review*, 6 (1), 3–16 https://onlinelibrary.wiley.com/doi/abs/10.1002/arcp.1085.

Berger, Jonah, Garrick Sherman, and Lyle Ungar "TextAnalyzer," (2020) http://textanalyzer.org/.

Bhatia, Sudeep (2014), "Confirmatory Search and Asymmetric Dominance," *Journal of Behavioral Decision Making*, 27 (5), 468–476 https://doi.org/10.1002/bdm.1824.

Blanchard, Simon J, Jacob Goldenberg, Koen Pauwels, and David A Schweidel (2022), "Promoting Data Richness in Consumer Research: How to Develop and Evaluate Articles with Multiple Data Sources," *Journal of Consumer Research*, 49 (2), 359–372 https://doi.org/10.1093/jcr/ucac018.

Brand, James, Ayelet Israeli, and Donald Ngwe "Using GPT for Market Research," (2023) https://dx.doi.org/10.2139/ssrn.4395751.

Brodie, Roderick J., Linda D. Hollebeek, Biljana Jurić, and Ana Ilić (2011), "Customer Engagement: Conceptual Domain, Fundamental Propositions, and Implications for Research," *Journal of Service Research*, 14 (3), 252–271 https://doi.org/10.1177/1094670511411703.

Bromley, Jane, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah "Signature Verification using a "Siamese" Time Delay Neural Network," "Advances in Neural Information Processing Systems," Vol. 6., Morgan-Kaufmann (1993) https://proceedings.neurips.cc/paper_files/paper/1993/hash/288cc0ff022877bd3df94bc9360b9c5d-Abstract.html.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei "Language Models are Few-Shot Learners," (2020) http://arxiv.org/abs/2005.14165, arXiv:2005.14165 [cs].

Browne, Will and Mike Swarbrick Jones "What works in e-commerce - a meta-analysis of 6700 online experiments," Technical report, Qubit Digital Ltd (2017).

Bruce, Norris I., B.P.S. Murthi, and Ram C. Rao (2017), "A Dynamic Model for Digital Advertising: The Effects of Creative Format, Message Content, and Targeting on Engagement," *Journal of Marketing Research*, 54 (2), 202–218 https://doi.org/10.1509/jmr.14.0117.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu (2016), "Evaluating replicability of laboratory experiments in economics," *Science*, 351 (6280), 1433–1436 https://doi.org/10.1126/science.aaf0918.

Cascio Rizzo, Giovanni Luca, Jonah Berger, Matteo De Angelis, and Rumen Pozharliev (2023), "How Sensory Language Shapes Influencer's Impact," *Journal of Consumer Research*, 50 (4), 810–825 https://doi.org/10.1093/jcr/ucad017.

Chambers, Christopher D. and Loukia Tzavella (2021), "The past, present and future of Registered Reports," *Nature Human Behaviour*, 6 (1), 29–42 https://www.nature.com/articles/s41562-021-01193-7.

Chang, Jonathan P., Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil "ConvoKit: A Toolkit for the Analysis of Conversations," Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors, "Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue," pages 57–60, 1st virtual meeting: Association for Computational Linguistics (2020) https://aclanthology.org/2020.sigdial-1.8.

Chu, Johan S. G. and James A. Evans (2021), "Slowed canonical progress in large fields of science," *Proceedings of the National Academy of Sciences*, 118 (41), e2021636118 https://www.pnas.org/doi/abs/10.1073/pnas.2021636118, publisher: Proceedings of the National Academy of Sciences.

Chung, Jaeyeon (Jae), Yu Ding, and Ajay Kalra (2023), "I Really Know You: How Influencers Can Increase Audience Engagement by Referencing Their Close Social Ties," *Journal of Consumer Research*, 50 (4), 683–703 https://doi.org/10.1093/jcr/ucad019.

Clark, Herbert H. (1973), "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research," *Journal of Verbal Learning and Verbal Behavior*, 12 (4), 335–359 https://www.sciencedirect.com/science/article/pii/S0022537173800143.

Cropley, Arthur (2006), "In Praise of Convergent Thinking," *Creativity Research Journal*, 18 (3), 391–404 https://doi.org/10.1207/s15326934crj1803_13.

Day, George S. (2011), "Closing the Marketing Capabilities Gap," *Journal of Marketing*, 75 (4), 183–195 https://doi.org/10.1509/jmkg.75.4.183.

De La Rosa, Wendy, Eesha Sharma, Stephanie M. Tully, Eric Giannella, and Gwen Rino (2021), "Psychological ownership interventions increase interest in claiming government benefits," *Proceedings of the National Academy of Sciences*, 118 (35), e2106357118 https://doi.org/10.1073/pnas.2106357118.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt (2019), "Predict science to improve science," *Science*, 366 (6464), 428–429 https://doi.org/10.1126/science.aaz1704.

Demszky, Dorottya, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker (2023), "Using large language models in psychology," *Nature Reviews Psychology*, 2 (11), 688–701 https://doi.org/10.1038/s44159-023-00241-5.

Deolankar, Varad, Ali Goli, S. Sriram, and Pradeep K. Chintagunta "User Engagement with Online Discussion Content: Does it Affect Attrition?," (2024) https://dx.doi.org/10.2139/ssrn.4755183.

Fang, Christina, Jeho Lee, and Melissa A. Schilling (2010), "Balancing Exploration and Exploitation Through Structural Design: The Isolation of Subgroups and Organizational Learning," *Organization Science*, 21 (3), 625–642 https://doi.org/10.1287/orsc.1090.0468.

Fiedler, Klaus (2018), "The Creative Cycle and the Growth of Psychological Science," *Perspectives on Psychological Science*, 13 (4), 433–438 https://doi.org/10.1177/1745691617745651, publisher: SAGE Publications Inc.

Gebhardt, Gary F., Gregory S. Carpenter, and John F. Sherry (2006), "Creating a Market Orientation: A Longitudinal, Multifirm, Grounded Analysis of Cultural Transformation," *Journal of Marketing*, 70 (4), 37–55 https://doi.org/10.1509/jmkg.70.4.037, publisher: SAGE Publications Inc.

Gligorić, Kristina, George Lifchits, Robert West, and Ashton Anderson (2023), "Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report)," *PLOS ONE*, 18 (3), e0281682 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0281682, publisher: Public Library of Science.

Goswami, Indranil and Oleg Urminsky (2020), "No Substitute for the Real Thing: The Importance of In-Context Field Experiments in Fundraising," *Marketing Science*, 39 (6), 1052–

1070 https://pubsonline.informs.org/doi/abs/10.1287/mksc.2020.1252, publisher: INFORMS.

Goswami, Indranil and Oleg Urminsky "Why Many Behavioral Interventions Have Unpredictable Effects in the Wild: The Conflicting Consequences Problem," (2022) https://papers.ssrn.com/abstract=4199453.

Guenoun, Bushra S. and Julian J. Zlatev "Sending Signals: Strategic Displays of Warmth and Competence," (2023).

Hagendorff, Thilo, Sarah Fabi, and Michal Kosinski (2023), "Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT," *Nature Computational Science*, 3 (10), 833–838 https://www.nature.com/articles/s43588-023-00527-x, publisher: Nature Publishing Group.

Hartmann, Jochen, Juliana Huppertz, Christina Schamp, and Mark Heitmann (2019), "Comparing automated text classification methods," *International Journal of Research in Marketing*, 36 (1), 20–38 https://www.sciencedirect.com/science/article/pii/S0167811618300545.

Hartmann, Jochen and Oded Netzer "Natural Language Processing in Marketing," K. Sudhir and Olivier Toubia, editors, "Artificial Intelligence in Marketing," Vol. 20. of *Review of Marketing Research*, pages 191–215, Emerald Publishing Limited (2023) https://doi.org/10.1108/S1548-643520230000020011.

Hartzmark, Samuel M, Samuel D Hirshman, and Alex Imas (2021), "Ownership, Learning, and Beliefs*," *The Quarterly Journal of Economics*, 136 (3), 1665–1717 https://doi.org/10.1093/qje/qjab010.

Hauser, John, Gerard J. Tellis, and Abbie Griffin (2006), "Research on Innovation: A Review and Agenda for Marketing Science," *Marketing Science*, 25 (6), 687–717 https://pubsonline.informs.org/doi/abs/10.1287/mksc.1050.0144, publisher: INFORMS.

Hern, Alex (2018), "The two-pizza rule and the secret of Amazon's success," *The Guardian* https://www.theguardian.com/technology/2018/apr/24/the-two-pizza-rule-and-the-secret-of-amazons-success.

Hewett, Kelly, William Rand, Roland T. Rust, and Harald J. van Heerde (2016), "Brand Buzz in the Echoverse," *Journal of Marketing*, 80 (3), 1–24 https://doi.org/10.1509/jm.15.0033, publisher: SAGE Publications Inc.

Hopkins, Daniel J., Yphtach Lelkes, and Samuel Wolken "The Rise of and Demand for Identity-Oriented Media Coverage," (2023) https://papers.ssrn.com/abstract=4578004.

Hou, Yupeng, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley "Bridging Language and Items for Retrieval and Recommendation," (2024) http://arxiv.org/abs/2403.03952, arXiv:2403.03952 [cs].

Hsee, Chrisopher K., George F. Loewenstein, Sally Blount, and Max H. Bazerman (1999), "Preference reversals between joint and separate evaluations of options: A review and theoretical analysis.," *Psychological Bulletin*, 125 (5), 576–590 http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.125.5.576.

Humphreys, Ashlee and Rebecca Jen-Hui Wang (2018), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 44 (6), 1274–1306 https://doi.org/10.1093/jcr/ucx104.

Hutson, Matthew (2023), "Hypotheses devised by AI could find 'blind spots' in research," *Nature* https://www.nature.com/articles/d41586-023-03596-0, bandiera_abtest: a Cg_type: Na-

ture Index Publisher: Nature Publishing Group Subject_term: Machine learning, Computer science, Technology.

Jackson, Joshua Conrad, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A. Lindquist (2022), "From Text to Thought: How Analyzing Language Can Advance Psychological Science," *Perspectives on Psychological Science*, 17 (3), 805–826 https://doi.org/10.1177/17456916211004899, publisher: SAGE Publications Inc.

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

Jerath, Kinshuk and Qitian Ren (2021), "Consumer Rational (In)Attention to Favorable and Unfavorable Product Information, and Firm Information Design," *Journal of Marketing Research*, 58 (2), 343–362 https://doi.org/10.1177/0022243720977830, publisher: SAGE Publications Inc.

John, Leslie K., Oliver Emrich, Sunil Gupta, and Michael I. Norton (2017), "Does "Liking" Lead to Loving? The Impact of Joining a Brand's Social Network on Marketing Outcomes," *Journal of Marketing Research*, 54 (1), 144–155 https://doi.org/10.1509/jmr.14.0237, publisher: SAGE Publications Inc.

Kanuri, Vamsi K., Yixing Chen, and Shrihari (Hari) Sridhar (2018), "Scheduling Content on Social Media: Theory, Evidence, and Application," *Journal of Marketing*, 82 (6), 89–108 https://doi.org/10.1177/0022242918805411, publisher: SAGE Publications Inc.

Kapoor, Sayash and Arvind Narayanan "Leakage and the Reproducibility Crisis in ML-based Science," (2022) http://arxiv.org/abs/2207.07048, arXiv:2207.07048 [cs, stat].

Kapoor, Sayash and Arvind Narayanan (2023), "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, 4 (9), 100804 https://www.sciencedirect.com/science/article/pii/S2666389923001599.

Kaul, Rupali, Stephen J. Anderson, Pradeep K. Chintagunta, and Naufel Vilcassim "Call Me Maybe: Does Customer Feedback-Seeking Impact Non-Solicited Customers?," (2024) https://papers.ssrn.com/abstract=4507183.

Kerr, Norbert L. (1998), "HARKing: Hypothesizing After the Results are Known," *Personality and Social Psychology Review*, 2 (3), 196–217 https://doi.org/10.1207/s15327957pspr0203_4, publisher: SAGE Publications Inc.

Kilgour, Mark and Scott Koslow (2009), "Why and how do creative thinking techniques work?: Trading off originality and appropriateness to make more creative advertising," *Journal of the Academy of Marketing Science*, 37 (3), 298–309 https://doi.org/10.1007/s11747-009-0133-5.

Kizilcec, René F., Chris Piech, and Emily Schneider "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses," "Proceedings of the Third International Conference on Learning Analytics and Knowledge," LAK '13, pages 170–179, New York, NY, USA: Association for Computing Machinery (2013) https://dl.acm.org/doi/10.1145/2460296.2460330.

Klayman, Joshua and Young-won Ha (1987), "Confirmation, disconfirmation, and information in hypothesis testing," *Psychological Review*, 94 (2), 211–228 Place: US Publisher: American Psychological Association.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015), "Prediction

Policy Problems," *American Economic Review*, 105 (5), 491–495 https://www.aeaweb.org/articles?id=10.1257/aer.p20151023.

Koning, Rembrand, Sharique Hasan, and Aaron Chatterji (2022), "Experimentation and Start-up Performance: Evidence from A/B Testing," *Management Science*, 68 (9), 6434–6453 https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2021.4209, publisher: INFORMS.

Koo, Minjung and Ayelet Fishbach (2012), "The Small-Area Hypothesis: Effects of Progress Monitoring on Goal Adherence," *Journal of Consumer Research*, 39 (3), 493–509 https://doi.org/10.1086/663827.

Koçak, Özgecan, Daniel A. Levinthal, and Phanish Puranam (2023), "The Dual Challenge of Search and Coordination for Organizational Adaptation: How Structures of Influence Matter," *Organization Science*, 34 (2), 851–869 https://pubsonline.informs.org/doi/full/10.1287/orsc.2022.1601, publisher: INFORMS.

Lakens, Daniël (2017), "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses," *Social Psychological and Personality Science*, 8 (4), 355–362 http://journals.sagepub.com/doi/10.1177/1948550617697177.

Landy, Justin F., 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 (2020), "Crowdsourcing hypothesis tests: Making transparent how design choices shape research results.," *Psychological Bulletin*, 146 (5), 451–479 https://doi.apa.org/doi/10.1037/bul0000220.

Lee, Dokyun, Kartik Hosanagar, and Harikesh S. Nair (2018), "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook," *Management Science*, 64 (11), 5105–5131 https://doi.org/10.1287/mnsc.2017.2902, publisher: INFORMS.

Linos, Elizabeth, Jessica Lasky-Fink, Chris Larkin, Lindsay Moore, and Elspeth Kirkman (2024), "The formality effect," *Nature Human Behaviour*, 8 (2), 300–310 https://www.nature.com/articles/s41562-023-01761-z, publisher: Nature Publishing Group.

Linos, Elizabeth, Allen Prohofsky, Aparna Ramesh, Jesse Rothstein, and Matthew Unrath (2022), "Can Nudges Increase Take-Up of the EITC? Evidence from Multiple Field Experiments," *American Economic Journal: Economic Policy*, 14 (4), 432–452 https://www.aeaweb.org/articles?id=10.1257/pol.20200603.

Lucas, Brian J. and Loran F. Nordgren (2020), "The creative cliff illusion," *Proceedings of the National Academy of Sciences*, 117 (33), 19830–19836 https://pnas.org/doi/full/10.1073/pnas.2005620117.

Ludwig, Jens and Sendhil Mullainathan (2024), "Machine Learning as a Tool for Hypothesis Generation*," *The Quarterly Journal of Economics*, page qjad055 https://doi.org/10.1093/qje/qjad055.

Lynch, John G., Joseph W. Alba, Aradhna Krishna, Vicki G. Morwitz, and Zeynep Gürhan-Canli (2012), "Knowledge creation in consumer research: Multiple routes, multiple criteria," *Jour-*

nal of Consumer Psychology, 22 (4), 473–485 https://www.sciencedirect.com/science/article/pii/S1057740812000952.

MacInnis, Deborah J., Vicki G. Morwitz, Simona Botti, Donna L. Hoffman, Robert V. Kozinets, Donald R. Lehmann, John G. Lynch, and Cornelia Pechmann (2020), "Creating Boundary-Breaking, Marketing-Relevant Consumer Research," *Journal of Marketing*, 84 (2), 1–23 https://doi.org/10.1177/0022242919889876, publisher: SAGE Publications Inc.

Malaie, Soran, Michael J. Spivey, and Tyler Marghetis (2024), "Divergent and Convergent Creativity Are Different Kinds of Foraging," *Psychological Science*, page 09567976241245695 https://doi.org/10.1177/09567976241245695, publisher: SAGE Publications Inc.

Malt, Barbara C., Steven A. Sloman, Silvia Gennari, Meiyi Shi, and Yuan Wang (1999), "Knowing versus Naming: Similarity and the Linguistic Categorization of Artifacts," *Journal of Memory and Language*, 40 (2), 230–262 https://www.sciencedirect.com/science/article/pii/S0749596X98925931.

Manning, Benjamin S., Kehang Zhu, and John J. Horton "Automated Social Science: Language Models as Scientist and Subjects," (2024) https://www.nber.org/papers/w32381.

Markowitz, David M. and Hillary C. Shulman (2021), "The predictive utility of word familiarity for online engagements and funding," *Proceedings of the National Academy of Sciences*, 118 (18), e2026045118 https://www.pnas.org/doi/abs/10.1073/pnas.2026045118, publisher: Proceedings of the National Academy of Sciences.

Matias, J. Nathan, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole (2021), "The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media," *Scientific Data*, 8 (1), 195 https://www.nature.com/articles/s41597-021-00934-7.

McGuire, William J. (1997), "Creative Hypothesis Generating in Psychology: Some Useful Heuristics," *Annual Review of Psychology*, 48 (1), 1–30 https://doi.org/10.1146/annurev.psych.48.1.1.

Messeri, Lisa and M. J. Crockett (2024), "Artificial intelligence and illusions of understanding in scientific research," *Nature*, 627 (8002), 49–58 https://www.nature.com/articles/s41586-024-07146-0, publisher: Nature Publishing Group.

Min, Sewon, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?," (2022) http://arxiv.org/abs/2202.12837, arXiv:2202.12837 [cs].

Moorman, Christine and George S. Day (2016), "Organizing for Marketing Excellence," *Journal of Marketing*, 80 (6), 6–35 https://doi.org/10.1509/jm.15.0423, publisher: SAGE Publications Inc.

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

Morwitz, Vicki G., Eric Johnson, and David Schmittlein (1993), "Does Measuring Intent Change Behavior?," *Journal of Consumer Research*, 20 (1), 46–61 https://doi.org/10.1086/209332.

Mullainathan, Sendhil and Jann Spiess (2017), "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31 (2), 87–106 https://www.aeaweb.org/articles?id=10.1257%2Fjep.31.2.87&ref=ds-econ.

Netflix Technology Blog "What is an A/B Test?," (2022) https://netflixtechblog.com/what-is-an-a-b-test-b08cc1b57962.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science*, 31 (3), 521–543 https://pubsonline.informs.org/doi/abs/10.1287/mksc.1120.0713, publisher: INFORMS.

Netzer, Oded, Alain Lemaire, and Michal Herzenstein (2019), "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications," *Journal of Marketing Research*, 56 (6), 960–980 https://doi.org/10.1177/0022243719852959, publisher: SAGE Publications Inc.

Nie, Allen, Yash Chandak, Miroslav Suzara, Malika Ali, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech "The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters Exam Performances," (2024) https://osf.io/qy8zd.

Nosek, Brian A. and Daniël Lakens (2014), "Registered Reports: A Method to Increase the Credibility of Published Results," *Social Psychology*, 45 (3), 137–141 https://econtent.hogrefe.com/doi/10.1027/1864-9335/a000192.

Oquendo, M. A., E. Baca-Garcia, A. Artés-Rodríguez, F. Perez-Cruz, H. C. Galfalvy, H. Blasco-Fontecilla, D. Madigan, and N. Duan (2012), "Machine learning and data mining: strategies for hypothesis generation," *Molecular Psychiatry*, 17 (10), 956–959 https://www.nature.com/articles/mp2011173.

Otis, Nicholas "Policy Choice and the Wisdom of Crowds," (2022) https://papers.ssrn.com/abstract=4200841.

Packard, Grant and Jonah Berger (2024), "The Emergence and Evolution of Consumer Language Research," *Journal of Consumer Research*, 51 (1), 42–51 https://doi.org/10.1093/jcr/ucad013.

Peterson, Joshua C., David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths (2021), "Using large-scale experiments and machine learning to discover theories of human decision-making," *Science*, 372 (6547), 1209–1214 https://www.science.org/doi/full/10.1126/science.abe2629.

Petty, Richard E. and John T. Cacioppo "The Elaboration Likelihood Model of Persuasion," "Advances in Experimental Social Psychology," Vol. 19., pages 123–205, Elsevier (1986).

Phillips, Barbara J. and Edward F. McQuarrie (2010), "Narrative and Persuasion in Fashion Advertising," *Journal of Consumer Research*, 37 (3), 368–392 https://doi.org/10.1086/653087.

Piezunka, Henning and Linus Dahlander (2015), "Distant Search, Narrow Attention: How Crowding Alters Organizations' Filtering of Suggestions in Crowdsourcing," *Academy of Management Journal*, 58 (3), 856–880 https://journals.aom.org/doi/abs/10.5465/amj.2012.0458, publisher: Academy of Management.

Pogacar, Ruth, Alican Mecit, Fei Gao, L. J. Shrum, and Tina M. Lowrey (2022), "Language and consumer psychology.," ISBN: 1433836424 Publisher: American Psychological Association.

Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J. Van Bavel "GPT is an effective tool for multilingual psychological text analysis," (2023) https://doi.org/10.31234/osf.io/sekf5.

Reiff, Joseph, Hengchen Dai, Jana Gallus, Anita McClough, Steve Eitniear, Michelle Slick, and Charlotte Blank "When Impact Appeals Backfire: Evidence from a Multinational Field Experiment and the Lab," (2023) https://papers.ssrn.com/abstract=3946685.

Revelle, William "psych: Procedures for Psychological, Psychometric, and Personality Research," (2007) https://CRAN.R-project.org/package=psych, institution: Comprehensive R Archive Network Pages: 2.4.3.

Robertson, Claire E., Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay J. Van Bavel, and Stefan Feuerriegel (2023), "Negativity drives online news consumption," *Nature Human Behaviour*, pages 1–11 https://www.nature.com/articles/s41562-023-01538-4.

Rosengren, Sara, Martin Eisend, Scott Koslow, and Micael Dahlen (2020), "A Meta-Analysis of When and How Advertising Creativity Works," *Journal of Marketing*, 84 (6), 39–56 https://doi.org/10.1177/0022242920929288, publisher: SAGE Publications Inc.

Rzhetsky, Andrey, Jacob G. Foster, Ian T. Foster, and James A. Evans (2015), "Choosing experiments to accelerate collective discovery," *Proceedings of the National Academy of Sciences*, 112 (47), 14569–14574 https://www.pnas.org/doi/abs/10.1073/pnas.1509757112.

Sah, Raaj Kumar and Joseph E. Stiglitz (1986), "The Architecture of Economic Systems: Hierarchies and Polyarchies," *The American Economic Review*, 76 (4), 716–727 https://www.jstor.org/stable/1806069, publisher: American Economic Association.

Schaller, Mark (2016), "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 https://www.sciencedirect.com/science/article/pii/S0022103115001092.

Sheetal, Abhishek, Zhiyu Feng, and Krishna Savani (2020), "Using Machine Learning to Generate Novel Hypotheses: Increasing Optimism About COVID-19 Makes People Less Willing to Justify Unethical Behaviors," *Psychological Science*, 31 (10), 1222–1235 https://doi.org/10.1177/0956797620959594.

Shin, Minkyu, Jin Kim, Bas van Opheusden, and Thomas L. Griffiths (2023), "Superhuman artificial intelligence can improve human decision-making by increasing novelty," *Proceedings of the National Academy of Sciences*, 120 (12), e2214840120 https://www.pnas.org/doi/full/10.1073/pnas.2214840120, publisher: Proceedings of the National Academy of Sciences.

Shulman, Hillary C., David M. Markowitz, and Todd Rogers (2024), "Reading dies in complexity: Online news consumers prefer simple writing," *Science Advances*, 10 (23), eadn2555 https://www.science.org/doi/10.1126/sciadv.adn2555, publisher: American Association for the Advancement of Science.

Siggelkow, Nicolaj and Daniel A. Levinthal (2003), "Temporarily Divide to Conquer: Centralized, Decentralized, and Reintegrated Organizational Approaches to Exploration and Adaptation," *Organization Science*, 14 (6), 650–669 https://pubsonline.informs.org/doi/abs/10.1287/orsc.14.6.650.24840, publisher: INFORMS.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2021), "Pre-registration: Why and How," *Journal of Consumer Psychology*, 31 (1), 151–162 https://onlinelibrary.wiley.com/doi/abs/10.1002/jcpy.1208.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson (2020), "Specification curve analysis," *Nature Human Behaviour*, 4 (11), 1208–1214 https://www.nature.com/articles/s41562-020-0912-z, publisher: Nature Publishing Group.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu "MPNet: Masked and Permuted Pre-training for Language Understanding," (2020) https://www.microsoft.com/en-us/research/publication/mpnet-masked-and-permuted-pre-training-for-language-understanding/.

Sourati, Jamshid and James A. Evans (2023), "Accelerating science with human-aware artificial intelligence," *Nature Human Behaviour*, 7 (10), 1682–1696 https://www.nature.com/articles/s41562-023-01648-z, number: 10 Publisher: Nature Publishing Group.

Tang, Diane, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer "Overlapping experiment infrastructure: more, better, faster experimentation," "Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining," KDD '10, pages 17–26, New York, NY, USA: Association for Computing Machinery (2010) https://dl.acm.org/doi/10.1145/1835804.1835810.

Tausczik, Yla R. and James W. Pennebaker (2010), "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology*, 29 (1), 24–54 https://doi.org/10.1177/0261927X09351676.

Thaler, Richard H. and Cass R. Sunstein (2009), *Nudge: improving decisions about health, wealth, and happiness* New York: Penguin Books.

Toubia, Olivier and Laurent Florès (2007), "Adaptive Idea Screening Using Consumers," *Marketing Science*, 26 (3), 342–360 https://pubsonline.informs.org/doi/abs/10.1287/mksc.1070.0273, publisher: INFORMS.

Toubia, Olivier and Oded Netzer (2017), "Idea generation, creativity, and prototypicality," *Marketing science*, 36 (1), 1–20 ISBN: 0732-2399 Publisher: INFORMS.

Uber Engineering "Supercharging A/B Testing at Uber," (2022) https://www.uber.com/blog/supercharging-a-b-testing-at-uber/.

Urminsky, Oleg and Berkeley J Dietvorst (2024), "Taking the Full Measure: Integrating Replication into Research Practice to Assess Generalizability," *Journal of Consumer Research*, 51 (1), 157–168 https://doi.org/10.1093/jcr/ucae007.

Vafa, Keyon, Justin Y. Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan "Evaluating the World Model Implicit in a Generative Model," (2024) http://arxiv.org/abs/2406.03689, arXiv:2406.03689 [cs].

Vanden Bergh, Bruce G., Leonard N. Reid, and Gerald A. Schorin (1983), "How Many Creative Alternatives to Generate?," *Journal of Advertising*, 12 (4), 46–49 https://doi.org/10.1080/00913367.1983.10672863, publisher: Routledge _eprint: https://doi.org/10.1080/00913367.1983.10672863.

Wang, Hanchen, 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 (2023), "Scientific discovery in the age of artificial intelligence," *Nature*, 620 (7972), 47–60 https://www.nature.com/articles/s41586-023-06221-2, publisher: Nature Publishing Group.

Wicherts, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen (2016), "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking," *Frontiers in Psychology*, 7 https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.01832/full, publisher: Frontiers.

Wu, Lingfei, Dashun Wang, and James A. Evans (2019), "Large teams develop and small teams disrupt science and technology," *Nature*, 566 (7744), 378–382 https://www.nature.com/articles/s41586-019-0941-9, number: 7744 Publisher: Nature Publishing Group.

Yiu, Eunice, Eliza Kosoy, and Alison Gopnik (2023), "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 https://doi.org/10.1177/17456916231201401, publisher: SAGE Publications Inc.

Zhou, Yangqiaoyu, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan "Hypothesis Generation with Large Language Models," (2024) http://arxiv.org/abs/2404.04326, arXiv:2404.04326 [cs].

Zor, Ozum, Kihyun Hannah Kim, and Ashwani Monga (2022), "Tweets We Like Aren't Alike: Time of Day Affects Engagement with Vice and Virtue Tweets," *Journal of Consumer Research*, 49 (3), 473–495 https://doi.org/10.1093/jcr/ucab072.

Appendix: Using Language to Generate Hypotheses

September 9, 2024

CONTENTS

WEB APPENDIX 1: PROMPTING MATERIALS

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.

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. The basic prompt format is shown Section 1.2.1.

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.[1]
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).

1.1.1 Prompt format

```
Assume you are {preamble}.

Below are two headlines. Assume that both are alternative
   headlines for the same news story.

Your task is to identify what has changed from Headline A in
   order to produce Headline B. Focus on the generalizable
   insight that can be applied in other contexts. Ignore
   things that are specific to this story. Do not make
   references this story they may not be for others.

{examples}

Come up with an insight the captures the sort of change
   observed moving from A to B.

Produce this insight as a single sentence that begins and ends
   in this exact format:

{hypothesis structure}


Please make sure that the hypothesis is:
```

[1] 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.

```
i. clear (i.e., precise, not too wordy, and easy to
   understand);
ii. generalizable to novel situations (i.e., they would make
   sense if applied to other headline experiments or other
   messaging contexts);
iii. empirically plausible (i.e., this is a dimension on which
   messages can vary on);
iv. unidimensional (i.e., avoid hypotheses that list multiple
   constructs so if there are many things changing, pick one);
v. usable (i.e., a human equipped with this insight could use
   it to improve another headline in a similar way)

{known contrasts}

Headlines to Assess:
Headline A: {H_A}

Headline B: {H_B}
```

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.

1.2.1 Prompt format

```
Assume you are a copywriter for an online news platform. Here
   are some examples of recent headlines from your company:

Example 1: {example_1}

Example 2: {example_2}

Example 3: {example_3}
```

```
You need to rewrite Headline A below according to the given
   instructions. Keep the content of the story as similar as
   possible. Respond by writing out Headline B.

The aim is to rewrite the headline such that it maximizes
   engagements. Therefore, Headline B should either emphasize
   or minimize the feature mentioned according to the
   hypothesized direction. Specifically, when the feature is
   thought to increase engagement, dial that feature up in
   Headline B. When the feature is thought to decrease
   engagement, dial that feature down in Headline B. If there
   is no clear direction hypothesized, emphasize the feature.

Headline A: {H_A}

Instruction: {D}

Headline B:
```

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.

1.3.1 Prompt format

```
Assume you are a communication scientist researching the
   effects of linguistic framing on reader perception. Your
   task here is to evaluate a given headline on a specific
   dimension. Use a scale from 0 to 7, where lower values
   means the feature is weakly present and higher values mean
   it is strongly present. 0 means the dimension is not
   present.

Your response should therefore be numeric, between 0 and 7.

Headline: {H}

Rate the headline on the following dimension: {D}

Rating:
```

WEB APPENDIX 2: FORMALIZING THE STEPS OF HYPOTHESES GENERATION FRAMEWORK

Step 1: Generating Hypotheses

Assume that we have some dataset $\mathcal{D}$ composed of observations and outcomes, (x_i, y_i). In the Upworthy dataset for example, each observation x is a pair of headlines (H_a, H_b), and y is $\Delta\text{CTR}_{a,b}$. The goal of this step is to come up with a set of hypotheses about the dataset and the outcome of interest. Here, by *hypothesis* we mean a statement that links a feature about $\mathcal{D}$ to a measurable impact on y. For example, one hypothesis may be: "using ambiguous or obscure cultural references makes people less likely to engage with a message." The measurable feature is the level of ambiguous or obscure cultural references within the headline. The measurable impact is a decrease in engagement, which in our case is measured by ΔCTR. Our procedure for generating hypotheses is to take a single data point x and to ask a large language model to come up with a plausible reason for why x has a large or small value of y. In addition, we specify a strict output format for the language model, in order to force the output to be a valid hypothesis. (Details of this step, along with quality checks, are given below.) The output of this step will be a large set of hypotheses, which we call $\mathcal{H}$.

Step 2: Ranking Hypotheses

Morphing. For this step, we assume that we have the same dataset $\mathcal{D}$, along with a set of hypotheses, $\mathcal{H}$. The goal of this step is to come up with a set of morphs, which are generated data points that resemble original data points from $\mathcal{D}$, while varying features outlined in hypotheses from $\mathcal{H}$. That is, for a given data point x and a given hypothesis $h \in \mathcal{H}$, we define the *morph* of x given h to be a new data point x' which satisfies two requirements. Firstly, x' should appear as similar as possible to x. Secondly, x' should exhibit the feature from h more strongly than x exhibits the feature. Note that this definition implies x' should vary features *not* mentioned in h as little as possible, since it should appear as similar possible to x. Our procedure for generating morphs is to take a single data point x and a single hypothesis h, and ask a large language model to write a headline that is both as similar as possible to x while making adjustments to increase the feature from h. In addition, we provide some examples of other headlines, to help the language model produce a morph that is consistent in style with $\mathcal{D}$. Examples of morphs from applying this procedure to the Upworthy dataset are shown in the main text. The output of this step will be a large set of morphs, which we will call $\mathcal{M}$:

$$\mathcal{M} = \{(x, h, x') \mid x \in \mathcal{D}, h \in \mathcal{H}\}.$$

We also suggest that the data points used in the creation of $\mathcal{M}$ should be independent of the data points used to train the machine learning model, in order to avoid biased estimates in later steps.

Scoring. For this step, we assume that we have the set of morphs $\mathcal{M}$ generated in the previous step, along with a machine learning model m which estimates $\mathbb{E}[y \mid x]$. For each hypothesis $h \in \mathcal{H}$, we can then score pairs of original and morphed headlines, by averaging

over morphs in $\mathcal{M}$ that used h. We call this the *predicted treatment effect* (PTE), which can be expressed as follows:

$$PTE(h) = \mathbb{E}\left[m(x, x') \mid (x, h, x') \in \mathcal{M}\right], \tag{1}$$

where the expectation is a sample average calculated over actual morphs. In the Upworthy data, we use the model $\widehat{m}$, and note that this is why we make sure that the *training* and *morphing* partitions of the data are kept independent.

Step 3: Filtering Hypotheses

Clustered Selection. For this step, we assume that we have the set of hypothesis $\mathcal{H}$ and the predicted treatment effect function $PTE : \mathcal{H} \to \mathbb{R}$. We also require some similarity measure d between pairs of hypotheses, where $d(h, h')$ is small when two hypotheses h and h' are very similar, and is large when they are very different. Our goal is to collect a subset of hypotheses from $\mathcal{H}$ that are both highly diverse, and have a high PTE. We define this set as follows: fix the value of some $\varepsilon > 0$, then define

$$\mathcal{H}' = \{h \in \mathcal{H} \mid d(h, h') > \varepsilon \text{ for all } h' \text{ such that } PTE(h') > PTE(h)\}.$$

For a given $\varepsilon > 0$, the set $\mathcal{H}'$ is uniquely defined, and can be constructed using a sequential selection strategy outlined below.

WEB APPENDIX 3: QUALITY CHECKS FOR GPT TASKS

Quality of hypotheses

3.1.1 Hypothesis Quality Rating Task

Participants. 79 Prolific users (Age: M = 37.91, SD = 12.63; Gender Identity: 39 Female, 38 Male, 2 Self-Identified; Race and Ethnic Identity: 60.8% white, 13.9% Black, 7.6% Latin American, 10.1% Multi-Racial, 7.6% All others) completed the labeling survey conducted in June 2024. The median participant completed the survey in 17.0 minutes. Each participant consented to participating in the study.

Procedure. After consenting, participants were told that they would be reading eight hypotheses and were asked to "imagine these hypotheses being applied to messages you might see in the world, such as online newspaper headlines or political campaign text messages or email subject lines from your favorite charity".

As part of the instructions, participants learned of the two parts to the task and then proceeded to complete them.

The first part asked participants to rate a hypothesis based on the following ***traits***:

- **Clarity** (i.e., whether the hypothesis is easy to understand)
- **Face-Value** (i.e., whether the hypothesis seems logical or if it's something that could be observed without complex analysis)
- **Generalizability** (i.e., whether the hypothesis could extend to multiple contexts where messages are sent)
- **Usability** (i.e., whether the hypothesis could be used by a human to change a given message)
- **Overall Impression** (i.e., is this a good hypothesis?)

Responses ranged from "1 (Low)" to "7 (High)".

Participants were also asked to select which ***contexts*** they could imagine the hypothesis being applied to. The set of contexts included:

- Online newspaper headlines
- Product descriptions
- Emails from a doctor's office
- Political campaign text messages
- Emails from charities
- Billboard advertisements
- Social media posts

- None of the above

The second part asked participants to make a prediction into how the "insight" might affect other ***outcomes***. For example, for the *hypothesis* "using humor increases engagement with a message", the *insight* is "using humor".

The list of outcomes included:

- Making a donation (of any amount, i.e., assuming the message was asking for donations, would applying this insight affect how many people chose to donate)
- Amount donated (i.e., for those who might've made a donation anyway, would this change how much they donated or the total amount fundraised)
- Registering to vote (i.e., might applying this insight to a message intended to get people registered lead more / less people to actually register)
- Voting in an election (i.e., might applying this insight to a 'Get Out the Vote' message change how many people went and voted)
- Opening an email or message (i.e., clicking on the message to view its contents)
- Responding to an email or message (e.g., this could mean writing a reply or comment or simply clicking on the link to take some action suggested like RSVPing to invitation, making an appointment, signing up for something)
- Unsubscribing from future messages (e.g., after receiving an email, the person chooses to unsubscribe)
- Blocking the messenger (e.g., blocking them on social media; blocking the number texting you; blocking the emails)
- Sharing the message (e.g., reposting on social media; forwarding email)
- Clicking on the content (e.g., clicking on a story in a news website, clicking on a social media post)
- Scanning a QR code (e.g., in a magazine, on a billboard, flyer, etc)

Participants were asked "If you had to guess, what effect do you think it would have on the following outcomes?" Response options included "Large Decrease", "Small Decrease", "No Meaningful Effect", "Small Increase", "Large Increase", and "Not Applicable".

Each participant saw eight hypotheses, drawn randomly from a set of 106.[2] Hypotheses were shown on separate pages.

Participants would see a hypothesis on one page, along with the rating questions and the context question. On the next page, they would see the "insight" alongside the outcomes questions.

[2]These 106 hypotheses consisted of 100 hypotheses randomly drawn from full set of hypotheses generated using GPT. Specifically, we randomly selected 10 hypotheses per decile of simulated treatment effects (see main paper). The six additional hypotheses were the six selected in the paper to be tested out of sample.

At the end of the survey, participants were asked their age, gender, education, and ethnic identity. We also asked them if they used GPT or another LLM to assist with this task and whether they were familiar with Upworthy.com.

Results. Overall, participants perceived hypotheses to be of high quality. Each hypothesis was rated by a median of 6 participants (Min: 4), which we then averaged across. On each trait, ratings were above the scale's midpoint of 4 (see Figure 1).

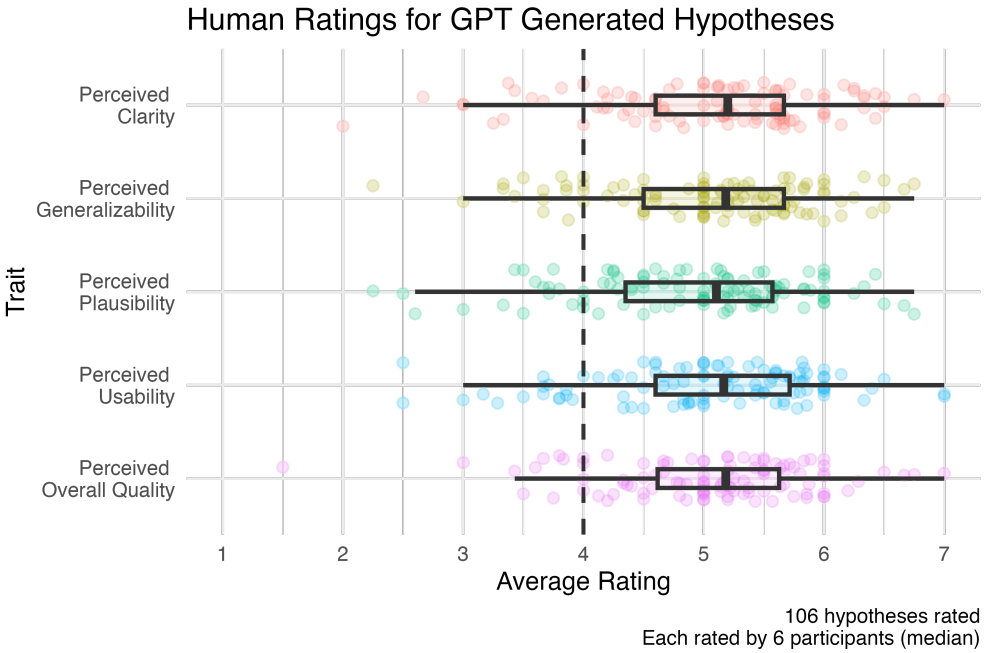

Figure 1: Human ratings for GPT-Generated Hypotheses.

Looking next at the number of contexts participants, on average, expected all hypotheses to apply to at least one other context (Prop Selecting None of the Above: 0). Most hypotheses could be applied to social media (Prop: .892), political SMS messages (.642), emails from charities (.547) and headlines (.575). Few were seen as applicable to emails from a doctor's office (Prop: .094). See Table 1.

For forecasted effects on the different outcomes, the median hypothesis was expected to have a positive effect on nearly every measure, other than "Unsubscribe" and "Block" which both received a median rating of 0 ("No Meaningful Effect"). The largest anticipated outcome was on "Clicks" for which the median rating was 1 (on a scale that ranged from -2 to $+2$.

Table 1: Summary of Participant Forecasts of Contexts and Outcomes

Raters			
Variable	**Mean**	**SD**	**Median**
Number of Raters	5.96	2.23	6
Contexts			
SMS Political	0.642		
Product Description	0.340		
Email Charity	0.547		
Headlines	0.575		
Social Media Posts	0.896		
Billboard Advertisement	0.434		
Email from Doctor's Office	0.094		
No Additional Contexts	0		
Outcomes			
Donation	0.568	0.665	0.732
Donation Amount	0.472	0.629	0.6
Register to Vote	0.563	0.569	0.667
Vote	0.580	0.579	0.667
Open	0.684	0.685	0.866
Respond	0.479	0.702	0.5
Share	0.478	0.728	0.586
Unsubscribe	0.0707	0.556	0
Block	-0.0491	0.521	0
Click	0.775	0.763	1
Scan	0.303	0.643	0.4

Note: For contexts, we report the proportion of hypotheses for which more than half the raters selected that context.

Quality of morphing procedure

Our first checks are aimed at confirming the quality of morphs.

3.2.1 Are morphs in distribution?

We first provide two checks to provide us confidence that morphed headlines are, in fact, "within-distribution". This is important, since for the ML model to make valid predictions, the morphs should come from the same data generating distribution, and for the morphs to be good quality, they should be good quality. These checks are designed to confirm that morphs are generally similar to the headlines they are based on, are high-quality, and do in fact manipulate the feature of interest.

For our first check, we use the sentence embedding model (described in the main text) to measure similarity between pairs of headlines. For reference, headlines from different trials have a median pairwise distance of 1.78 (using Euclidean distance between embedding vectors). By comparison, headlines from the same trial have a median pairwise distance of 1.04, suggesting that headlines from the same trial are more alike than those from different trials. Finally, the median pairwise distance between headlines and their associated morphs is just .469, suggesting that morphs are indeed very similar to the original headline they are based on.

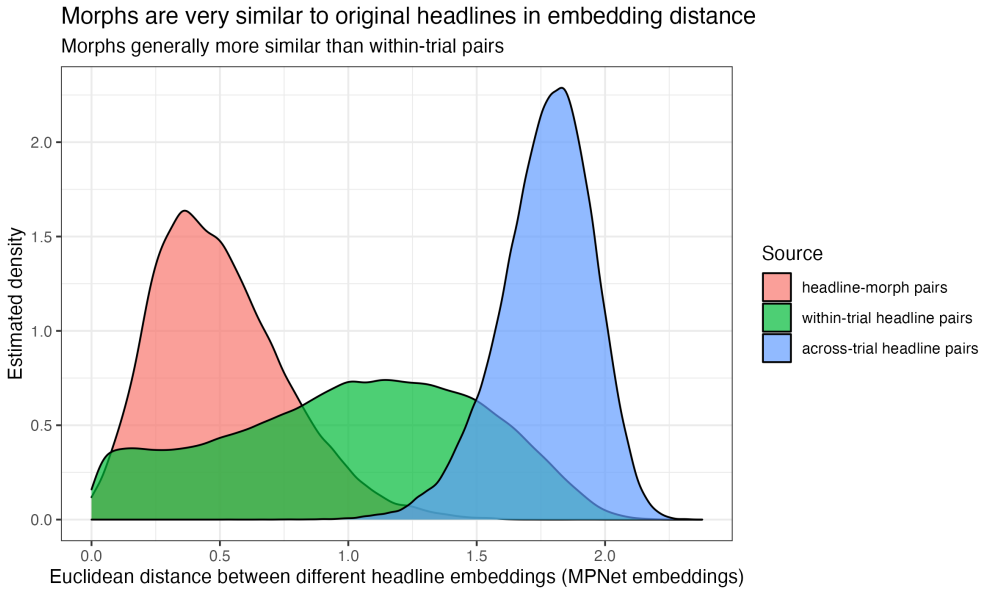

Figure 2: Average Euclidean distance between pairs.

3.2.2 Experiment 1: How do humans' attitudes towards morphed headlines compare to original?

This study aimed to test whether headlines generated using large language models (GPT-4) are perceived differently to those written by humans. In particular, we sought to test whether GPT headlines are perceived as worse. This study was pre-registered on AsPredicted.org (#177783).

Participants. 120 Prolific users (Age: M = 40.33, SD = 14.64; Gender Identity: 59 Female, 59 Male, 2 Self-Identified; Race and Ethnic Identity: 60.8% white, 12.5% Black, 8.3% Latin American, 8.3% Multi-Racial, 10.0% All others) completed the labeling survey conducted in June 2024. The median participant completed the survey in 11.12 minutes. Each participant consented to participating in the study.

Procedure. After consenting, participants were told that they would be reading twenty headlines and would be asked to rate their level of interest.

Each participant then saw twenty headlines randomly drawn from a set of 300. This set contained 150 original Upworthy headlines and 150 morphs.

Each headline was displayed on a single page along with the set of five questions:

- **Interest**: How *interested* would you be in reading this article? Response options ranged from "0: Not interested" to "6: Extremely interested".
- **Likelihood to Click**: Imagine you came across this headline online, how likely would you be to *click* on it? Response options ranged from "0: Not likely at all" to "6: Extremely likely"
- **Own Impression**: Based only on this headline, what is *your* overall impression of the article? Response options ranged from "−3: Extremely unfavorable" to "3: Extremely favorable"
- **Other Impression**: Based only on this headline, what do you think the overall impression is of *other* Prolific users? Response options ranged from "−3: Extremely unfavorable" to "3: Extremely favorable"
- **Quality**: What is the *quality* of this headline? Response options ranged from "−3: Extremely bad" to "3: Extremely good"

After completing the main rating task, participants answered a reading check to confirm they understood the instructions they were meant to be following and then ended the survey with a set of demographic questions.

Results. Our primary analyses included a series of two-sided t-tests comparing the average rating of morphed headlines to original headlines.

We find no detectable difference between attitudes towards morphs versus actual Upworthy headlines (see Table 2). The results are qualitatively similar when we account for participant fixed effects and headline-level fixed effects. And when we control for outliers. Furthermore, the equivalence test was significant ($p < .001$) for all measures, given equivalence bounds of half a unit on the scale ($-.5$, $.5$; see also Lakens 2017).

Table 2: Quality of Morphs: Mean Attitude Ratings

Measure	Morph	Upworthy	t-Statistic	Cohen's d (95% CI)	P-Value
Interest	2.62 (1.91)	2.62 (1.92)	0.01	0.00 [−0.08, 0.08]	0.992
Click	2.64 (1.97)	2.64 (1.97)	0.06	0.00 [−0.08, 0.08]	0.955
Self Impression	0.10 (1.68)	0.02 (1.67)	1.22	0.05 [−0.03, 0.13]	0.224
Other Impression	0.13 (1.58)	0.07 (1.64)	0.90	0.04 [−0.04, 0.12]	0.368
Quality	0.02 (1.82)	−0.07 (1.85)	1.21	0.05 [−0.03, 0.13]	0.228

Note: Mean and standard deviation reported for each.

3.2.3 Experiment 2: Can humans accurately detect which headlines were AI generated?

This study aimed to test whether human raters could accurately detect headlines generated using large language models (GPT-4). In particular, we sought to test whether GPT headlines were perceived as "AI generated". We also tested whether GPT headlines were perceived as relatively *more* AI generated (*less* "human generated") than Upworthy headlines written by humans. This study was pre-registered on AsPredicted.org (#177785).

Participants. 101 Prolific users (Age: M = 38.92, SD = 12.64; Gender Identity: 47 Female, 51 Male, 3 Self-Identified; Race and Ethnic Identity: 58.4% white, 13.9% Black, 9.9% Latin American, 6.9% Multi-Racial, 10.9% All others) completed the survey conducted in June 2024. The median participant completed the survey in 6.15 minutes. Each participant consented to participating in the study.

Procedure. After consenting, participants were told that they would be reading twenty headlines and would be asked decide whether each headline was written by a human or AI.

Each participant saw 20 headlines randomly drawn from a set of 300 (150 morphs; 150 original Upworhty headlines). This set was identical to the one used in the attitudes experiment above. Each headline was presented on a separate page, along with the question "Was this headline written by a human or AI?". The 7-point scale ranged from "+3: Definitely Human" to "+3: Definitely AI" where the midpoint was "0: Unsure".

Participants were incentivized to answer according to their true beliefs. Specifically, they earned and lost points depending on whether they were categorically (in)correct and how confident they were. For example, if they rated a headline as ""+3: Definitely Human" and the headline was an Upworthy headline, they earned 3 points. If, in fact, that headline was GPT-generated, they lost 3 points. In the end, the points were summed up and each participant was paid \$0.05 for every positive point. The maximum bonus one could earn was therefore \$3.00.

After completing the main rating task, participants answered a reading check to confirm they understood the instructions they were meant to be following and then ended the survey with a set of demographic questions.

Results. Our primary analysis was a one-sided t-test comparing the average rating of morphed headlines to 0.5. As pre-registered, if the mean is significantly less than 0.5, we would reject the null that GPT-generated hypotheses were perceived as AI generated. On average, morphed headlines were rated as .11, significantly less than 0.5, *one*-sided $t(1028) = -5.91$, $p < .001$.

Comparing the mean to the midpoint of the scale, zero, we see a marginal difference, *one*-sided $t(1028) = 1.59$, $p = .056$.

As a secondary analysis, we compared ratings of the morphed headlines to the ratings of Upworthy headlines. Morphed headlines ($M = .11$, $SD = 2.14$) were seen as relatively more AI generated (less human generated) then Upworthy headlines ($M = -.09$, $SD = 2.20$), $t(2018) = 2.01$, Cohen's $d = .09$, 95% [.00, .18], $p = .045$. When we account for participant-level and headline-level fixed effects, the effects are similar, $p = .057$. Nevertheless, the

equivalence test was significant ($p < .001$) given equivalence bounds of half a unit on the scale ($-.5$, .5; Lakens 2017), suggesting these differences may not be meaningfully different.

3.2.4 Experiment 3: Can humans accurately detect which headlines were produced by Upworthy?

This study aimed to test whether actual Upworthy headlines are perceived as Upworthy headlines, more than headlines generated by GPT. This study was pre-registered on AsPredicted.org (#177786).

Participants. 10 Prolific users (Age: $M = 40.43$, $SD = 13.14$; Gender Identity: 47 Female, 50 Male, 2 Self-Identified, 1 NA; Race and Ethnic Identity: 62.6% white, 13.1% Black, 8.1% Latin American, 6.1% Multi-Racial, 4.0% East Asian, 6.1% All others) completed the survey conducted in June 2024. The median participant completed the survey in 7.58 minutes. Each participant consented to participating in the study.

Procedure. After consenting, participants were told that they would be reading twenty headlines and would be asked decide whether each headline was produced by Upworthy.com or not. Participants were told that Upworthy.com was a well-known news platform alongside a link to the website in case they wanted to view it. They were also provided 10 examples headlines written by Upworthy between 2014 and 2016.

Each participant then saw 20 headlines randomly drawn from a set of 300 (150 morphs; 150 original Upworhty headlines). This set was identical to the one used in the attitudes experiment and morph detection experiment above. Each headline was presented on a separate page, along with the question "Was this headline written writers at Upworthy.com?". The 7-point scale ranged from "−3: Definitely Not Upworthy" to "+3: Definitely Upworthy" where the midpoint was "0: Unsure".

Like the study above, participants were incentivized to answer according to their true beliefs. Specifically, they earned and lost points depending on whether they were categorically (in)correct and how confident they were. For example, if they rated a headline as ""+2: Very Likely Upworthy" and the headline was *not* an actual Upworthy headline, they lost 2 points. In the end, the points were summed up and each participant was paid \$0.05 for every positive point. The maximum bonus one could earn was \$3.00.

After completing the main rating task, participants answered a reading check to confirm they understood the instructions they were meant to be following and then ended the survey with a set of demographic questions.

Results. Our primary analysis involved a two-sided t-test comparing the average rating of original headlines to morphed headlines. Headlines written by Upworthy writers ($M = .50$, 1.92) were thought more likely to have been written by Upworthy writers than morphed headlines ($M = .21$, $SD = 1.95$), $t(1978) = 3.31$, Cohen's $d = .15$, 95% CI [.06, .24]. This effect appears more pronounced when adjusting for participant-level and headline-level fixed effects, $p = .003$. However, the equivalence test was also significant ($p = .007$), given equivalence bounds of half a unit on the scale ($-.5$, .5; Lakens 2017), suggesting these ratings may not be meaningfully different.

Furthermore, morphed headlines were also perceived to be written by Upworthy writers, with an average rating significantly greater than the midpoint of zero, *one*-sided $t(994) = 3.37$, $p < .001$.

3.2.5 Do morphs manipulate the feature of interest?

We now consider a separate check, to confirm that morphs are manipulating our feature of interest. Our strategy here will be to collect labels for a variety of headline-morph-hypothesis triplets, and measuring the change in label values for hypotheses from which the morph was generated (for which the change should be large) and the change in label values for hypotheses unrelated to the morph (for which the change should be small).

Because of the large number of labels required for this exercise, we use GPT to emulate the labeling task outlined in the main text. We first select a subset of 87 hypotheses by taking the six hand-selected hypotheses (see main text) and uniformly sampling from the remaining hypotheses. We then combine the 120 original headlines and 10,351 unique morphed headlines into a single set of headlines. For each headline, we then label it on each of the 87 hypotheses, using a GPT task. The GPT task uses a prompt instructing the model to rate a headline on a given dimension, on a scale of 0 to 7, where 0 indicates the feature is not present. For the full prompt format, see above.

We find that 53% of morphs have a higher score for the feature of interest than their original headline when the feature of interest is the one on which the morph is generated. For that feature, 40% of morphs have the same value as the original headline, and only 7% decrease the label value for the morphed feature. By comparison, only 24% of morphs have a higher score for the feature of interest when that feature is not one which was being morphed, with 57% remaining unchanged, and 19% decreasing. The mean change in label value is 0.89 (on the eight-point scale) for morphed features, and 0.09 for unrelated features. Hence, we see that morphing does a reasonable job of increasing the value of the morphed feature, while holding other features constant.

Quality of labeling exercises: GPT versus human ratings

Since we use GPT to label features for a variety of sub-tasks, we are interested in checking the quality of our GPT ratings (see also Rathje et al. 2023). For this exercise, we compare the mean of human ratings and to the GPT ratings, for which we have ratings for (i.e., the six hypotheses of interest on headlines used in the regression set). We compare the strength of agreement between the human and GPT ratings.

For our first check, we calculate Krippendorff's alpha coefficient for each headline label based on the suggestion by Humphreys and Wang (2018). Because we compare ratings at the headline level, we have ratings for 3,400 headlines for each label. For the human rating, we use the mean human label value, without any other adjustment. For the GPT rating, we use the result of the GPT labeling task, without any other adjustment. To calculate the 'raw' Krippendorff's alpha, we use the ordinal metric when calculating alpha values. The results of this test are included in Table 3. We find that the agreement is very poor, and in some cases very negative. However, Krippendorff's alpha is sensitive to changes in scale and location of the rating of interest. Since we are typically interested in the relative

size of labels, we standardize each of the GPT and human labels, and repeat the above exercise. The results of this are shown as a 'Z-score' alpha in Table 3, where we see that this substantially improves the alpha coefficient, although the values still range from model (0.600 for positive human behaviour) to poor (0.111 for short and simple).

Feature name	Krippendorff's alpha	
	Raw label	Z-score
Surprise w/ Cliffhanger	-0.690	0.235
Parody	0.078	0.426
Multimedia	0.238	0.422
Phsyical Reactions	0.373	0.427
Short & Simple	-0.256	0.111
Positive Human Behavior	0.440	0.600

Table 3: Krippendorff's alpha coefficient for the six hypotheses of interest, comparing the mean of human ratings with the GPT ratings.

As a further check of GPT label quality, we consider how well a "marginal" human coder's rating agrees with both GPT and the mean human label. For this check, we randomly sample a single rating for each headline and hypothesis, and reserve this as the marginal rater. We then aggregate the remaining human ratings to form a jackknifed mean human rating. We then look at the correlation between the GPT rating against the marginal human rating, versus the jackknifed mean human rating against the marginal human rating. We report the Pearson correlation coefficient values, but find similar results using the Spearman rank correlation coefficient. The results of this exercise are in Table 4. Firstly, we see that the marginal human rating has a similar correlation to both the jackknifed mean human rating and the GPT rating for each label, suggesting that the jackknifed mean human rating and the GPT rating are both appropriate as label sources. Secondly, we see that the strength of association between the GPT label and the mean human ratings (either the jackknife mean or the full mean) is much stronger than the marginal human rating. This gives further evidence that the GPT rating is a good approximation for the mean human rating (up to a linear transformation). While GPT's rating are not perfectly correlated with the *average* human rating, they are no worse than a single human is to the average.

Feature name	Human (marginal) vs		GPT vs	
	Jackknife	GPT	Jackknife	Human (full)
Surprise, Cliffhanger	0.179	0.121	0.218	0.235
Parody	0.194	0.227	0.401	0.426
Multimedia	0.254	0.234	0.386	0.422
Physical Reactions	0.205	0.243	0.401	0.427
Short, Simple Phrases	0.120	0.088	0.078	0.111
Positive Human Behavior	0.395	0.381	0.578	0.600

Table 4: The left two columns show correlations between the marginal human rating to the jackknifed mean human rating and the GPT rating, in order to compare how closely each source resembles an additional human rater. The right two columns show correlations between the GPT rating and the jackknifed mean human rating and the full human rating, in order to assess how closely the two sources agree.

WEB APPENDIX 4: DIVERSITY OF HYPOTHESES

In this section, we consider the motivation for having GPT produce hypotheses on a pair-by-pair basis. We will collect hypotheses from both humans and GPT, and from a strategy that has a participant (or language model) generate a hypothesis from a single pair, or based on multiple pairs.

To collect human hypotheses, we conducted two studies (data and materials available on OSF). The first gathered human hypotheses from pairs of headlines. 104 Prolific users completed a study that asked them to read a pair of headlines and fill in a hypothesis in the format "Hypothesis: _____ increases [decreases] engagement with a message." The specific format was randomly drawn from the same set used in the LLM prompts. Participants in the pairwise study saw two pairs of headlines each and wrote two hypotheses.

The second study gathered human hypotheses after seeing many pairs of headlines. This was the same study used to collect human guesses. 303 participants first completed the guessing task described in the body of the paper, which consisted of 40 trials. Each trial displayed a pair of headlines written for the same story and participants were incentivized to select the headline that performed better in an AB test. After completing the main set of tasks, participants were asked to fill in a hypothesis in the format "Hypothesis: _____ increases [decreases] engagement with a message." The specific format was randomly drawn from the same set used in the LLM prompts.

For the population of GPT hypotheses based on a single pair of headlines, we use the same population of 2,100 hypotheses as those collected in the main text. For the population of GPT hypotheses based on multiple pairs of headlines, we run a separate hypothesis collection exercise, that is otherwise as similar as possible to the process reported in the main text. We use a prompt that is a slightly modified one from the prompt shown above, which replaces the single pair of headlines with 20 consecutive pairs of headlines, and makes minor adjustments to the instructions accordingly. To maximise the diversity of generated hypotheses, we draw a single headline pair from each component and assign that pair to exactly one prompt. In line with the preceding hypothesis generation process, we use only the training partition of the Upworthy dataset. This ensures that each prompt draws 20 pairs of headlines from trials not used in any other prompt. Because this process limits us to only 133 unique prompts, we also sample 10 hypotheses from each prompt. We apply the same cleaning procedure as used for pairwise hypotheses. Despite an instruction to produce only one single insight, 574 of these draws produced multiple hypotheses, which we remove from the sample.[3] This leaves a final sample of 756 hypotheses generated from the aggregate GPT exercise.

We now have a dataset containing hypotheses from two sources, humans and GPT, generated using two strategies, either from single headline pairs or from multiple headline pairs. We measure semantic diversity within each of these four groups by finding the pairwise distance between embedding vectors for hypotheses. The results of this are shown in Table 5. The results show that when presenting either humans or GPT with multiple pairs of headlines, the resulting hypotheses tend to have reduced semantic diversity. The results are stronger for headlines created from GPT.

[3] repeating the following analysis on these excluded hypotheses, we find the same conclusion: that GPT-generated hypotheses from groups of headline pairs have less diversity.

Source	Strategy	Mean	Median	75th percentile
GPT	Aggregate	0.597	0.600	0.722
GPT	Pairwise	0.857	0.851	1.025
human	Aggregate	0.768	0.747	0.897
human	Pairwise	0.873	0.864	1.016

Table 5: Pairwise distances between embedding vectors for hypothesis, generated from various strategies.

WEB APPENDIX 5: HOW STRONG IS THE AVERAGE PTE SIGNAL?

Sensitivity of results to average PTE

We have seen compelling evidence that using average PTE to prioritize hypotheses worthy of testing helps to identify features that do indeed predict our outcome of interest. In this section, we test for whether the *predicted* treatment effect produced in the generating stage correlates to the *estimated* treatment effect in the hold-out set. For this, we first use GPT to emulate the human labeling task from the main text. Then we use these labels in a series of regressions with the same specification as the Hypothesis Testing section from the main text. Finally, we compare how the coefficient value of the hypothesized features to the respective average PTE values gathered in the ranking step.

We select a random sample of hypotheses for testing by first dividing the range of observed average PTE values into 10 intervals of equal width (winsorizing the top and bottom 1% of values to account for sparsity at extreme ranges of the distribution). From each interval, we then sample 40 hypotheses uniformly at random. We then repeat the label collection strategy outlined above in Section 3.2.5, except that here the headlines we collect labels for are from the *regression* partition of our dataset. We follow a similar process to the human labeling task used for Hypothesis Testing in the main text, only this time we needed many more ratings so we used GPT (see also Section 3.3 above). Unlike the human rating task, the scale in the prompt for GPT did not include "0" as an option. The result is a dataset containing $1,360,800$ labels, exhausting all combinations of the 400 sampled hypotheses and $3,402$ headlines (from the regression set). For each label, we then run a regression using the same specification as used in hypothesis testing in the main text, predicting ΔCTR without any other covariates, and extract both the coefficient estimate $\widehat{\beta}_r$ and the p-value for $\widehat{\beta}_r$.

Figure Figure 3 displays a summary of the results where we average the $\widehat{\beta}_r$ coefficient values from the regressions per stratum. This suggests that hypotheses with higher average PTEs are identifying features that produce larger effects. Figure 4 aggregates the p-values for these coefficients per stratum. Here we see that hypotheses with higher average PTEs produce more significant results out-of-sample.

It is important to keep in mind in interpreting these results is that although these are a diverse set of hypotheses, there is still a lot of overlap, both on the surface, in the description of the features, but also likely in the underlying psychology. While we adjust for the False Discovery Rate in this analysis, it is possible that two hypotheses, randomly chosen for this exercise, are picking up a similar feature. Another thing to note in interpreting these values is that there may be an asymmetry in the *quality* of hypotheses across the various stratum. Higher average PTEs indicate that the hypotheses, when applied to several random headlines, resulted in morphed headlines that were *predicted* to perform better than their original. This is a function of both the ML algorithm, the hypothesis quality, and the morph quality. It is difficult to imagine these being high by chance, but the reverse is possible. That is, lower average PTEs could be due to many low-quality morphs or morphs that happened to out-of-distribution or it could be due to the hypothesis, either being too specific or too complex or non-sensical (therefore producing odd morphs). The latter would also be harder to rate, resulting in inconsistent effects when estimating the regression. We see some suggestive evidence of this by the fact that the proportion of hypotheses in the lowest decile (most

negative average PTEs) is not as high as those in the top decile — we suspect it is because features likely to produce negative effects are mixed in with hypotheses of lower quality.

Nevertheless, the results extend the findings in the main text. When testing with human raters, we found evidence for four out of six hypotheses, a higher proportion than we might expect at $\alpha = .05$. Note that these were selected to be independent hypotheses and therefore needed no further adjustments. Here we find a similar rate of significant effects (out of sample) for hypotheses with the highest average PTEs (calculated in the ranking step), further validating our framework for discovery.

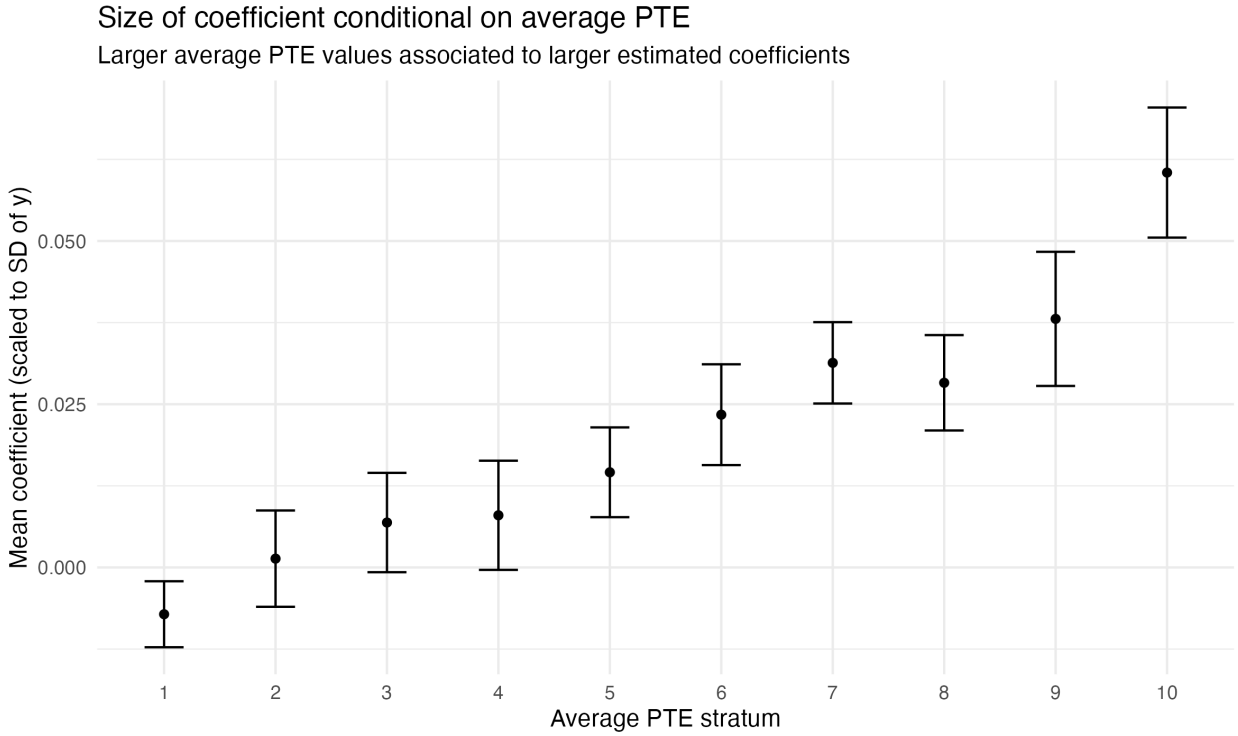

Figure 3: Coefficients tend to increase as a function of average PTE. Error bars show 2 standard errors of the coefficient values within each stratum.

WEB APPENDIX 6: GENERALIZING HYPOTHESES TO NEW CONTEXTS

Social Media Partner Data

6.1.1 Data

We partnered with an organization that produces articles on popular culture, lifestyle and sport. This organization shared a dataset of 553,328 different posts for various articles on a large social media platform between July 2022 and February 2023. The dataset contains the message of the post, along with the URL of the page being linked on the organization's website, a categorization of the page being linked into one of 66 categories. A total of 1442 rows are dropped, 1245 for missing the message information and 204 for having zero total

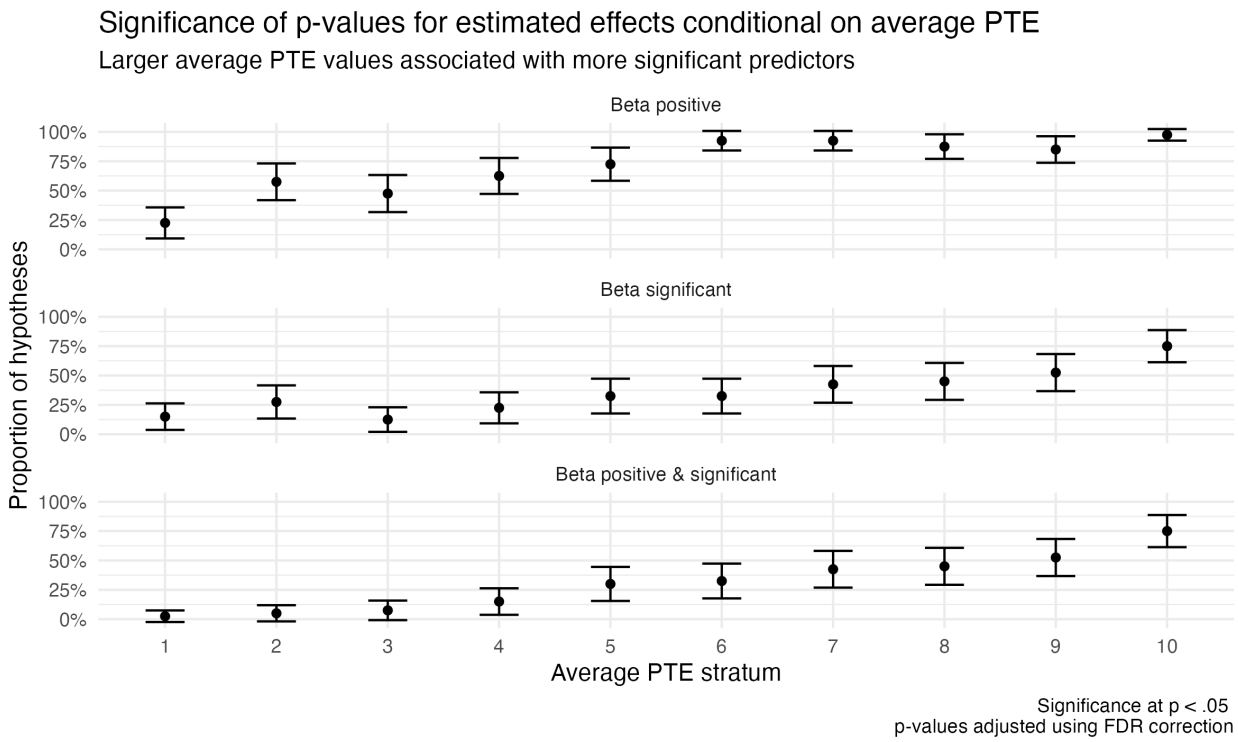

Figure 4: The significance of predictors increases as a function of average PTE. Error bars show 2 standard errors of the shown proportion within each stratum.

reach (i.e., no viewers), leaving 551,886 rows, containing 21,922 unique URLs and 35,693 unique messages. The message ranges from empty to 1,024 characters long, although 95% of messages fall between 22 and 221 characters long. It also includes several outcomes of interest: the total reach of the post, the number of link clicks it received, as well as the number of likes, comments, and shares the post received. From these measures we calculate the ratio of clicks to total reach (the *click-through-rate*), along with the ratio of comments, likes, and shares to total reach (the *comment rate*, *like rate*, and *share rate*, respectively). There are some additional outcomes, such as the number of different reaction types, but these are generally very small and excluded from our analysis (the most frequent reaction occurs at a rate less than 0.2%). Because of its similarity to the outcome used throughout the Upworthy dataset, we consider the *click-through-rate* to be our primary outcome of interest.

This dataset is not organized into any kind of valid experimental unit. There are generally a much higher number of posts linking to the same URL, and many posts that share a common outcome. However, since these are not conducted as a valid A/B trial, we neither combine posts with common messages, nor do we create a pairwise dataset comparing two different posts. Instead, for this dataset we conduct analysis at the post level, and treat *click-through-rate* as our main outcome of interest. We also repeat the dataset partitioning strategy applied to the Upworthy dataset: we form 'components' that group trials which share any common message. The resulting post-level dataset contains the following splits:

- A training dataset with 8,708 unique messages across 133,470 posts, for 5,371 different URLs

- A regression set with 8,970 unique messages across 133,739 posts from 5,475 unique URLs
- A morphing set with 3,492 unique messages across 56,851 posts from 2,123 unique URLs
- A lock-box set with 14,523 unique messages across 227,826 posts from 8,953 unique URLs

Some summary statistics for these splits are included in Table 6.

Table 6: Counts for Social Media Data

	Splits				
	Training	*Regression*	*Morphing*	*Lock-Box*	**Total**
Post-Level					
Total Subject Lines	133470	133739	56851	227826	**551886**
Unique Subject Lines	8708	8970	3492	14523	**35693**
URL-Level					
Total URLs	5371	5475	2123	8953	**21922**
Total Components	5057	5116	2007	8108	**20288**
Average # of Posts	24.85	24.43	26.78	25.45	**25.17**
Average # of Unique Messages	1.69	1.71	1.70	1.71	**1.71**

Note: Here we do not combine posts with common messages, since posts are not valid A/B trials.

6.1.2 Procedure

We follow a similar procedure as described in the main text. Again, we pre-registered our procedure on AsPredicted.org (#181144). We used the same six hypotheses uncovered and tested above. We kept the direction consistent for simplicity, but note here that behavioral interventions often have different effects across different people and different contexts (Goswami and Urminsky 2022). Our primary outcome was the *click-through-rate* (CTR). However, we were also interested in examining whether the hypothesized features might affect other outcomes too; in particular, the *like rate*, *share rate*, and *comment rate*. Together, the aim was to understand whether hypotheses generated in one dataset could predict various outcomes in another time and place.

We planned to recruit 900 participants. Each participant saw 30 subject lines, each on a separate page, randomly drawn from a set of 5,077. For each subject line, participants were asked to "select the level which each trait is featured in this subject line, from '1 (Low)' to '7 (High)'." There was also an option to select "0" to indicate the trait was not present. The traits 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*.

6.1.3 Results

In the end we recruited 900 participants (M_{age} = 37.74, SD = 13.04; 448 Male, 435 Female, 17 Self-Identified; 60.6% white, 14.4% Black, 11.7% Latin American, 5.0% Multiracial, 3.7% East Asian, 4.7% all others) through Prolific. Altogether, participants provided 162,000 labels.

To test each of the six hypotheses, we estimated OLS regressions following a similar specification to the one used for testing hypotheses in the Upworthy data. Notably, we regressed CTR on Rating rather than taking the difference of each since these posts were not part of a randomized experiment. We also dropped an additional two rows from the data, for having a total reach of zero viewers.

The estimated coefficients for each of our main regressions (outcome: CTR) are displayed in Table 7. We find that four of the hypothesized features have significant association with CTR ($ps < 0.01$). We find that *physical reactions* has a positive association, which accords with its originally hypothesized association, while both *short and simple phrases* and *positive human behavior* have a significant negative relationship, in accordance with their original setting. On the other hand, *multimedia* has a strong negative association.

Table 7: How well do features explain click through rates, for social media organizational partner's data?

	Dependent variable: CTR						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Surprise, Cliffhanger	-0.006						-0.002
	(0.012)						(0.012)
Parody		0.009					0.019
		(0.012)					(0.013)
Multimedia			-0.060***				-0.069***
			(0.012)				(0.012)
Physical Reactions				0.034**			0.059***
				(0.012)			(0.013)
Short, Simple Phrases					-0.082***		-0.078***
					(0.012)		(0.012)
Positive Human Behavior						-0.084***	-0.077***
						(0.012)	(0.012)
Constant	0.618***	0.618***	0.618***	0.618***	0.618***	0.618***	0.618***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Observations	5,056	5,056	5,056	5,056	5,056	5,056	5,056
R^2	0.000	0.000	0.005	0.002	0.010	0.010	0.027
Adjusted R^2	0.000	0.000	0.005	0.001	0.010	0.010	0.026

Note: †p<0.10; *p<0.05; **p<0.01; ***p<0.001

To make coefficients interpretable, we have scaled the outcome variable, CTR, by dividing by the standard deviation of CTR (.0202), and scaled each of the hypothesized features to have unit variance. Hence, a one standard deviation increase in any of the hypothesized features produces a change in CTR equal to $\widehat{\beta}$ times the standard deviation in CTR.

We also analyzed the relationship on the remaining outcomes. Figure 5 shows the coefficient estimates are significantly different from zero ($p < .05$) for several features and outcome combinations. In addition to the four features that show strong associations to the CTR, five show strong associations to the *like rate*, five to the *share rate*, and one to the *comment rate*, $p < .05$. Worth noting is the fact that the effect of many of the features go in opposite directions depending on the outcome. As it happens, in this dataset, the rate of comments, shares and likes are negatively correlated with the the CTR.

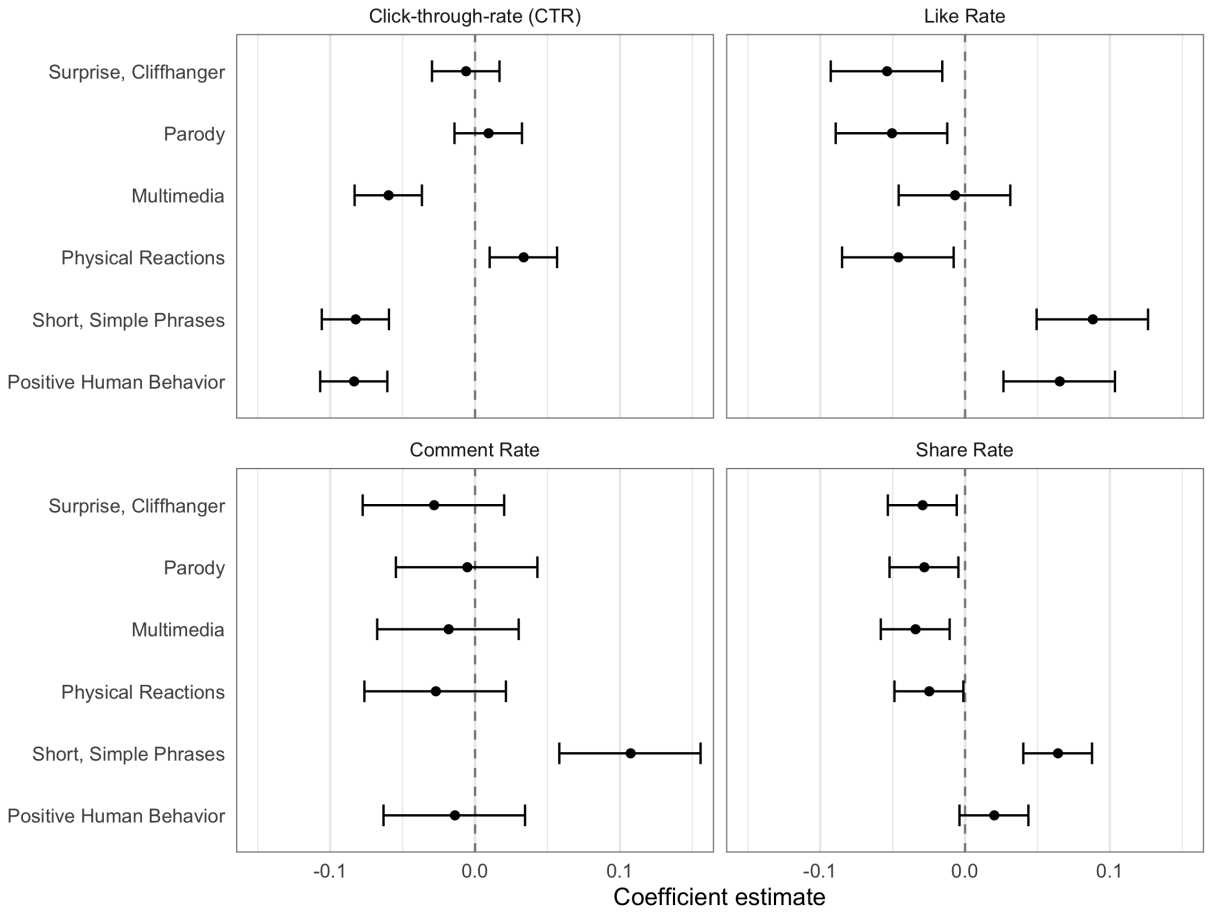

Figure 5: The coefficients for a range of outcomes and hypothesized features. Shown is the coefficient estimates with 95% confidence intervals for each outcome and feature combination.

Progressive Outreach Data

As a further check of the generalizability of our hypotheses, we partnered with a progressive outreach organization to test whether the hypotheses could generalize further to this context.

6.2.1 Data

We partnered with an organization that seeks to advance the interests of domestic workers across America. This organization shared a dataset of 1,653 email campaigns sent to its members and supporters between September 2020 and May 2024. The dataset contains the subject line of the email (which ranges from a single word to roughly 130 characters, but is between 16 and 86 characters for 95% of subject lines), the date sent, and several outcomes of interest: the number of recipients, the proportion of recipients who opened the email (*open rate*), the proportion of recipients who clicked on a link in the email (*click-through-rate*), the proportion of recipients who "took an action" (*conversion rate*; such as completing a survey, signing a pledge, or making a donation), the proportion who made a donation through the link (*contribution rate*), and the amount donated (*average contribution amount*), the proportion of recipients who unsubscribed after receiving the email (*unsubscribe rate*), and the number of emails that were 'bounced' (rejected by the email server) (*bounce rate*).

This dataset is also organized into trials, although notably, most trials contain only a single headline: of the 1,211 trials, 1,064 of them contain only one subject line, leaving 147 trials with more than one subject line and 364 unique subject lines between them. We apply the same data partitioning strategy as for Upworthy, to ensure that no subject lines are repeated across splits, and no trials are distributed across different splits. However, we use use different split sizes, in order to produce a small partition for exploratory analysis, and a larger partition for running regressions. The resulting pairwise dataset contains the following splits:

- A exploratory dataset with 138 pairs from 30 trials
- A regression set with 506 unique headline pairs from 117 unique trials

Some summary statistics for these splits are included in Table 8.

Table 8: Counts for Progressive Outreach Partner Data

	Splits		
	EDA	*Regression*	**Total**
Message-Level			
Total Subject Lines	291	1151	**1442**
Unique Subject Lines	235	947	**1182**
Pair-Level			
Total Pairs	138	506	**644**
Unique Trials	30	117	**147**
Unique Pairs	69	253	**322**
Unique Subject Lines	75	289	**364**
Trial-Level			
Total URLs	243	968	**1211**
Total Components	189	769	**958**
Average # of Messages	1.20	1.19	**1.19**

Note: Here we do not combine posts with common messages, since posts are not valid A/B trials.

Because this data has also been organized into pairs, we once again consider the difference in rates (for each of the outcomes outlined above) as our key dependent measure. Moreover, while we planned to analyze effects across all outcomes, we chose the CTR as our primary outcome since it most resembled the outcome in the Upworthy data.

6.2.2 Procedure

We follow a similar procedure as above (also reported in main text). Again, we pre-registered our procedure on AsPredicted.org (#178928). We used the same six hypotheses uncovered and tested above. We kept the direction consistent for simplicity, but note here that behavioral interventions often have different effects across different people and different contexts (Goswami and Urminsky 2022). Our primary outcome was the *click-through-rate* (CTR). However, we were also interested in examining whether the hypothesized features might affect other outcomes too; in particular, the *conversion rate*, *contribution rate*; *average contribution amount*, and *unsubscribe rate*.[4] Together, the aim was to understand whether hypotheses generated in one dataset could predict various outcomes in another time and place.

We planned to recruit 100 participants. Each participant saw 30 subject lines, each on a separate page, randomly drawn from a set of 300. For each subject line, participants were asked to "select the level which each trait is featured in this subject line, from '1 (Low)' to '7 (High)'." There was also an option to select "0" to indicate the trait was not present. The traits 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*.

6.2.3 Results

In the end we recruited 101 participants ($M_{age} = 37.5$, $SD = 11.2$; 48 Male, 50 Female, 3 Self-Identified; 66.3% white, 11.9% Black, 6.9% Latin American, 5.9% Multi-racial, 8.9% all others) through Prolific. Altogether, participants provided 18,180 labels.

To test each of the six hypotheses, we estimated OLS regressions following the specification in the main text.

The estimated coefficients for each of our main regressions (outcome: CTR) are displayed in Table 9. In a regression with two-way clustered standard errors (clustering on the subject line ID within each pair), physical reactions has a significant and positive association with CTR ($p < .05$) and short and simple has a significant and negative association with CTR ($p < 0.05$).

[4]We also planned to look at the *open rate* but this measure is particularly noisy since Apple introduced its Mail Privacy Protection policies (e.g., Kaczanowski 2021; Mask 2021).

Table 9: How well do features explain pairwise difference unique clicks, for progressive outreach partner's data?

	Dependent variable: ΔCTR						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Surprise, Cliffhanger	0.001						-0.023
	(0.018)						(0.022)
Parody		0.019					0.004
		(0.019)					(0.023)
Multimedia			0.020				0.009
			(0.019)				(0.023)
Physical Reaction				0.059*			0.061*
				(0.024)			(0.025)
Short, Simple Phrases					-0.056*		-0.054*
					(0.023)		(0.021)
Positive Human Behavior						0.025	-0.008
						(0.028)	(0.030)
Observations	506	506	506	506	506	506	506
R^2	-0.002	0.000	0.008	0.021	0.019	0.002	0.042
Adjusted R^2	-0.002	0.000	0.001	0.021	0.019	0.002	0.033

Note: †p<0.10; *p<0.05; **p<0.01; ***p<0.001

Standard errors shown in parentheses are clustered on the ID of both the left and right ID of the pair. To make coefficients interpretable, we have scaled the outcome variable, ΔCTR, by dividing by the standard deviation of CTR for experiments with at least two trials (.0101), and scaled each of the hypothesized features to have unit variance. Hence, a one standard deviation increase in any of the hypothesized features produces a change in ΔCTR equal to $\widehat{\beta}$ times the standard deviation in CTR for multi-arm trials.

We also analyzed the relationship on the remaining outcomes. Figure 6 shows the coefficient estimates are significantly different from zero ($p < .05$) for several features and outcome combinations. In addition to the two features that show strong associations on the CTR, one appears to predict the total open rate, two the conversion rate, two the contribution rate, and three the unsubscribe rate.

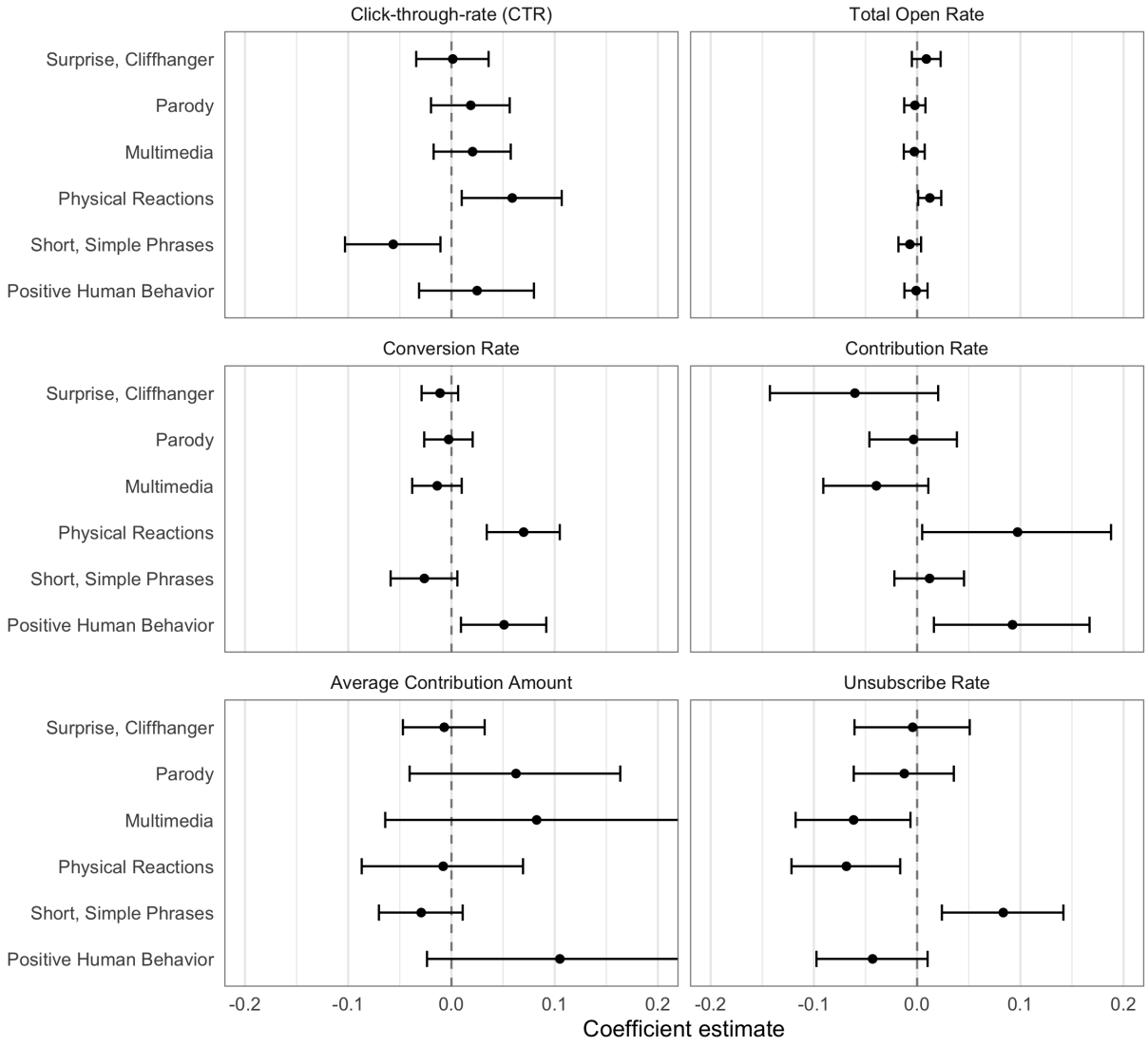

Figure 6: Most features have a non-significant relationship to most outcomes of interest. Shown is the coefficient estimates with 95% confidence intervals for each outcome and feature combination.

WEB APPENDIX 7: EXCLUDING NON-RANDOM TRIALS

On 11 July 2024, Matias et al. (2021) published a correction to the data where they acknowledge "problems with the randomization of the tests between June 25, 2013 and

January 10, 2014. A total of 7,004 A/B tests or 22% of experiments may have been affected" (see also Eckles 2024). They go on to encourage "researchers to treat these tests as not randomized... researchers conducting causal analysis [are encouraged] to omit all experiments from June 25, 2013 through the end of January 10, 2014."

Of primary concern for us is in *testing* the hypotheses generated. Table 10 estimates the primary regressions from the main text, omitting trials from June 25, 2013 to January 10, 2014.

The results are largely consistent with the results reported in the paper. While the evidence against the null is weaker for some features (e.g., physical reactions), it appears stronger for others (e.g., parody; positive human behavior).

As an additional check, we also train a new ML model to predict CTR, excluding rows of the data covered by the correction but holding all other decisions constant. We then compare the predictions of the refit model to those of the original model on the regression dataset used throughout the paper, also excluding those rows impacted by the non-randomness problem; this results in a dataset of 1337 valid headline pairs out of the 1693 pairs from the previous dataset. Firstly, we find that the two models produce similar predictions: they have a correlation of 91.7% on this dataset. Secondly, the performance is comparable: the original model has Adjusted $R^2 = .122$ on this dataset (slightly worse than the value reported in Table **??** for the larger dataset), and the refit model has Adjusted $R^2 = .126$ on the dataset excluding rows with randomization problems.

WEB APPENDIX 8: ADDITIONAL FIGURES

Table 10: How well do features explain pairwise difference in click-through? (Excluding non-randomized trials)

	Dependent variable: ΔCTR						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Surprise, Cliffhanger	0.055***						0.062***
	(0.012)						(0.013)
Parody		−0.028*					−0.047***
		(0.013)					(0.013)
Multimedia			0.039**				0.045***
			(0.013)				(0.014)
Physical Reactions				0.018			0.020
				(0.013)			(0.014)
Short, Simple Phrases					−0.018		−0.018
					(0.013)		(0.013)
Positive Human Behavior						−0.031*	−0.048***
						(0.013)	(0.014)
Constant	−0.001	−0.001	−0.001	0.0002	−0.0002	−0.0002	−0.003
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Observations	1,337	1,337	1,337	1,337	1,337	1,337	1,337
R^2	0.015	0.004	0.007	0.002	0.002	0.004	0.040
Adjusted R^2	0.014	0.003	0.006	0.001	0.001	0.004	0.036

Note: †p<0.10; *p<0.05; **p<0.01; ***p<0.001. To make coefficients interpretable, we have scaled the outcome variable, ΔCTR, by dividing by the standard deviation of CTR (.0119), and scaled each of the hypothesized features to have unit variance. Hence, a one standard deviation increase in any of the hypothesized features produces a change in ΔCTR equal to $\widehat{\beta}$ times the standard deviation in CTR.

Table 11: Human Intelligence Tasks

Common Name	Survey #	Short Description	Final Dataset	Prolific Users	Pre-Registration	Additional Notes
Human Guess Labeling Task	1	Participants are presented with 40 pairs of headlines (10 training; 30 testing), and instructed to select which headline performed better (worse) in an A/B test. They are given feedback after each selection in first 10 trials (so they can learn to identify patterns). Particiapnts $0.25 for selecting the correct answer in at least 17 out of 30 testing rounds plus an additional $0.25 for each correct response beyond that. Pairs of headlines were always from the same A/B test (written for same story). See Application ¿ Additional Features section.	Contains 12,080 human guesses for 1,793 headline pairs (100 train; 1,693 regression set). Each pair in the regression set was rated an average of 5.35 times (IQR: 4, 7).	303	—	Data and materials available on OSF.
Hypothesized Feature Labeling Task (Upworthy)	2	Participants label 30 headlines on sliders for (i) element of surprise followed by cliffhanger, (ii) parody, (iii) reference to multimedia evidence, (iv) description of physical reaction, (v) short and simple phrases, (vi) focus on positive aspects of human behavior. See Hypothesis Testing section.	Contains 144,000 labels (124,800 for headlines in the regression set, 4,212 for headlines in morph set, and 14,988 for morphed headlines)	800	AsPredicted.org (#172038)	First 26 trials were used for hypothesis testing and included only regression set headlines. The final 4 trials were used for exploratory analysis and included mix of morphs and morph-set headlines. Data and materials available on OSF.
Hypothesized Feature Labeling Task (Social Media)	3	Same as above	Contains 162,000 labels for 5,077 unique social media posts.	900	AsPredicted.org (#181144)	Survey follows an identical format to #2. Data is available on OSF; materials contain proprietary information, so are not yet available.
Hypothesized Feature Labeling Task (Progressive Outreach)	4	Same as above	Contains 18,180 labels for subject lines in the regression set.	101	AsPredicted.org (#178928)	Survey follows an identical format to #2 but replaces "headlines" with "subject lines" throughout. Data is available on OSF; materials are not yet available since they contain proprietary information.
Hypothesis Generation Task (Pairwise)	5	Participants provide a hypothesis in the format "Hypothesis: _____ increases [decreases] engagement with a message." where they write in a response to fill in the blank after seeing a single pair of headlines. Each participant sees two pairs and writes two hypotheses.	Contains 204 hypotheses written by humans.	104	—	Participants were randomly assigned to an "increase" (vs. "decrease") set in which Headline B always performed better (worse) than Headline A. Hypotheses formats reflected this difference. Formats were randomly drawn from same set used in LLM prompts. Half of participants were also randomly assigned to see four example hypotheses (vs. no examples). Data and materials available on OSF.
Hypothesis Generation Task (Aggregate)	6	Participants provide a hypothesis in the format "Hypothesis: ____ increases [decreases] engagement with a message." where they write in a response to fill in the blank after seeing 40 pairs of headlines (see #1).	Contains 303 hypotheses written by humans.	303	—	This survey is the same as #1. Hypotheses were always filled in after completing the main trials. Hypotheses formats were randomly drawn from same set used in LLM prompts.
Hypothesis Quality Rating	7	Participants rate LLM-generated hypotheses on whether they are clear, empirically plausible, generalizable, and usable. They also provide overall impressions, select which additional contexts the hypotheses might apply to, and forecast various outcomes.	Contains 3,160 labels for 106 hypotheses. Each hypotheses was rated by an average of 5.96 human raters (IQR: 4, 7).	79	—	See Quality of Hypotheses section in Appendix.
Morph Quality #1: Attitudes	8	Participants read 20 headlines and rate them based on interest, likelihood of clicking, own overall impression, others overall impression, and general quality.	Contains 12,000 ratings for 299 headlines (150 morphs; 149 original).	120	AsPredicted.org (#177783)	
Morph Quality #2: AI Detection	9	Participants read 20 headlines and assess whether they are AI or human generated.	Contains 2,020 ratings for 300 headlines (150 morphs; 150 original).	101	AsPredicted.org (#177785)	Participants were incentivized to report their true beliefs, earning and losing points based on accuracy and confidence (1 pt = $0.05).
Morph Quality #3: Upworthy Detection	10	Participants read 20 headlines and assess whether they are written by writers at Upworthy.com.	Contains 1,980 ratings for 299 headlines (150 morphs; 149 original).	100	AsPredicted.org (#177786)	Participants were provided examples of Upworthy headlines and were incentivized to report their true beliefs, earning and losing points based on accuracy and confidence (1 pt = $0.05).

REFERENCES

Eckles, Dean "Pervasive randomization problems, here with headline experiments," (2024) https://statmodeling.stat.columbia.edu/2024/06/20/pervasive-randomization-problems-here-with-headline-experiments/.

Goswami, Indranil and Oleg Urminsky "Why Many Behavioral Interventions Have Unpredictable Effects in the Wild: The Conflicting Consequences Problem," (2022) https://papers.ssrn.com/abstract=4199453.

Humphreys, Ashlee and Rebecca Jen-Hui Wang (2018), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 44 (6), 1274–1306 https://doi.org/10.1093/jcr/ucx104.

Kaczanowski, Rob "How Apple's Mail Privacy Changes Affect Email Open Tracking," (2021) https://postmarkapp.com/blog/how-apples-mail-privacy-changes-affect-email-open-tracking.

Lakens, Daniël (2017), "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses," *Social Psychological and Personality Science*, 8 (4), 355–362 http://journals.sagepub.com/doi/10.1177/1948550617697177.

Mask, Clate (2021), "Three Ways Apple's Privacy Changes Will Impact Your Business," *Forbes* https://www.forbes.com/sites/forbestechcouncil/2021/10/12/three-ways-apples-privacy-changes-will-impact-your-business/, section: Innovation.

Matias, J. Nathan, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole (2021), "The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media," *Scientific Data*, 8 (1), 195 https://www.nature.com/articles/s41597-021-00934-7.

Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J. Van Bavel "GPT is an effective tool for multilingual psychological text analysis," (2023) https://doi.org/10.31234/osf.io/sekf5.