Words that Work: Using Large Language Models to Generate and Refine Hypotheses from Text

Rafael M. Batista* James Ross†

October 9, 2025

Latest Version Here

Abstract

In this paper, we introduce a data-driven framework for generating and refining hypotheses from text. Our three-step approach—Hypothesize, Intervene, and Predict—integrates large-language models (LLMs), machine learning (ML), and experimentation to discover testable insights about how language drives consumer behavior. Using a dataset with over 60,000 headlines and 32,000 A/B tests, we first Hypothesize linguistic features by prompting an LLM to identify differences between headline pairs. We then use an LLM to Intervene—systematically rewriting headlines to incorporate these features—and use an ML model, trained on historical outcomes, to Predict the causal impact of these changes on engagement. The framework generates a prioritized list of hypotheses, which we validate on a hold-out set of 1,693 A/B tests. Our approach indeed facilitates discovery. For instance, we find that describing physical reactions significantly increases engagement, while focusing on positive aspects of human behavior decreases it. This approach extends beyond headlines, offering a method for converting 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, providing a practical tool for researchers, organizations, and policymakers seeking to aggregate insights from several messaging experiments.

Keywords:
Consumer Psychology; Consumer Language; Large Language Models; Digital Experiments; Hypothesis Generation; Scientific Discovery

*Princeton University, Corresponding Author: rbatista@princeton.edu
†Learning Collider

CONSUMER RELEVANCE AND CONTRIBUTION STATEMENT

How do we learn what shapes consumer behavior from the language they use and encounter every day? Language is core to consumer behavior, not only as a topic for studying decision processes but also as a rich source of data. As more of our everyday communication is captured—through audio, video, and online interactions—there is more data than ever to explore. A key challenge has been converting this vast amount of unstructured data into interpretable insights. New technologies offer new opportunities to address this.

This paper introduces a data-driven framework—Hypothesize, Intervene, and Predict—that integrates large language models, machine learning, and experimentation to systematically discover and prioritize hypotheses from text data. Applying this framework to over 60,000 headlines and 32,000 A/B tests, we demonstrate its capacity for discovery by uncovering linguistic drivers of consumer engagement. Our process allows researchers and managers to understand not just what language works, but to generate and prioritize specific, testable hypotheses about which linguistic features drive engagement.

Our contributions are threefold. The first is methodological. We offer a transparent and scalable process for hypothesis generation and refinement. While large language models now make generating hypotheses easier, our framework addresses a critical next step: once researchers have generated many hypotheses, how do they prioritize which to test? We use machine learning to predict treatment effects at scale while maintaining interpretable hypotheses that researchers can test empirically. The second is substantive. We uncover new insights about how language affects engagement. For instance, describing physical reactions significantly increases engagement while focusing on positive aspects of human behavior decreases it. Finally, our work offers a practical tool for organizations. The framework transforms accumulated A/B test data from isolated findings into systematic insights, enabling managers to aggregate learnings across experiments and develop more effective, data-driven communication strategies.

INTRODUCTION

Language is core to consumer behavior, both as a topic for studying decision processes (Pogacar et al. 2022; Packard and Berger 2024; Berger and Packard 2023) and 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.

Several 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 2024; 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 without a randomized trial (Banerjee and Urminsky 2024).

This paper advances our understanding of the effects of language on engagement while offering a scalable framework to researchers, organizations, and policymakers for generating hypotheses about what drives engagement in their specific context. 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 and refining novel and interpretable

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

hypotheses from text using a combination of large-language models (LLMs), machine learning, and psychology experiments. Large-language models, such as OpenAI's GPT, play a crucial role in processing text data and generating coherent hypotheses (Banker et al. 2024; Demszky et al. 2023; Zhou et al. 2024). 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; Shin et al. 2023). We integrate both technologies and validate the various steps through standard psychology experiments.

The framework consists of three steps: First, *hypothesize*. This step involves generating hypotheses from a pair of messages. LLMs are prompted to identify a feature that differs between the two messages which could plausibly affect engagement. 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" the LLM responded with, "Hypothesis: Incorporating reverse psychology leads to more engagement with a message." Second, *intervene*. The second step uses an LLM to rewrite a message to incorporate a hypothesized feature. Following from the earlier example, incorporating "reverse psychology" to the headline "Folks Who Work In Tipped Jobs Would Like You To Spend A Minute Looking At Something" produces an alternative, counterfactual headline that reads, "You Probably Shouldn't Read This if You Think Tipping Is Optional". Third, *predict*. Equipped with two messages, the third step leverages an ML model, trained on past messages and outcomes, to predict the effect of incorporating the hypothesized feature. When done at scale, across several messages, this approach utilizes a machine learning algorithm to detect complex patterns in the data (Oquendo et al. 2012; Hutson 2023; Wang et al. 2023) while also accounting for the generalizability of each hypothesis.

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 con-

sumers 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.[2] For each experimental arm, we also see the click-through rate (CTR), which we use to measure engagement. Using this data, we uncover dozens of unique human-interpretable hypotheses, select and test six out-of-sample, and find causal evidence supporting five. For instance, we discover that describing physical reactions significantly increases engagement, while focusing on positive aspects of human behavior decreases it.

Data-Driven Hypotheses

Researchers have traditionally generated hypotheses through insight and creativity (McGuire 1997; Abbott 2004). Hypotheses were then tested against data that one had collected. This approach takes time (Chu and Evans 2021; Rzhetsky et al. 2015; Fiedler 2018) and can be susceptible to human biases (Glaeser 2006; Klayman and Ha 1987). Recent advances in artificial intelligence are beginning to change this, offering new tools to augment how researchers discover and test ideas about human behavior (De Freitas, Nave, and Puntoni 2025; Berger, van Osselaer, and Janiszewski 2025).

Ludwig and Mullainathan (2024) introduced one approach for how researchers might start with unstructured data to generate interpretable hypotheses. Their method begins with a dataset of images matched to a set of outcomes. This allows them to train an ML algorithm to predict outcomes based on the information contained in the image. Their innovation, however, is in how they extract the hypotheses. As part of their process, they also use a generative model to create new images—images that closely resemble the original barring a specific feature. The images were of the same face and, yet, when provided to the a prediction algorithm trained on past images, the prediction was starkly different. What exactly was

[2]Among the benefits of this dataset is that it has been used before in consumer language research (e.g., Banerjee and Urminsky 2024; Robertson et al. 2023; Gligorić et al. 2023; Hopkins, Lelkes, and Wolken 2023; Shulman, Markowitz, and Rogers 2024; Aubin Le Quéré and Matias 2025); thus, allowing us to build on the work of others and benchmark our findings against existing insights extracted from this data.

different about the two images? The pairs of images were shown to human participants who were asked to verbalize whatever difference was apparent to them. By having humans label the differences between morphed image pairs (with different predicted outcomes), this approach effectively translates algorithmic patterns into interpretable insights. While this approach works for images, text data is fundamentally different. For instance, where images are continuous, text is discrete. Changing a single pixel maintains much of an image intact, but changing even a letter ("run" to "r**a**n") or removing punctuation ("Let's eat**,** Grandma" to "Let's eat Grandma") can alter the meaning of the text entirely. Moreover, features in text are mutable in ways that image features typically are not—hypotheses derived from text must be not only interpretable but also *usable*, applicable to new messages written by humans or LLMs.

A more recent approach tackles the challenge that comes with working with text data. Movva et al. (2025) use sparse autoencoders on text embeddings to systematically discover latent features across entire text datasets, applying statistical methods to identify predictive features before employing LLMs to interpret these patterns in natural language. Both Ludwig and Mullainathan (2024) and Movva et al. (2025) start with identifying statistical relationships and later "translate" these to natural language. Our approach, instead, begins with interpretable features and then uses computational methods to refine the set based on their predicted effects. These approaches are not mutually exclusive; indeed, one could generate hypotheses using Movva et al. (2025)'s approach and refine the set with the method we describe below. However, one advantage of starting with interpretable features, like we propose, is that the set of hypotheses can be interrogated from the start. Researchers can sample a distribution of hypothesized features before deploying any statistical tools.

Other research in this space uses LLMs to generate hypotheses directly from text data. Banker et al. (2024) 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 hy-

potheses in published papers. Zhou et al. (2024) developed a system that prompts LLMs to generate and score hypotheses. For a given pair of messages, the LLM is provided a hypothesis and prompted to predict the winning message. The researchers then calculate a score based on how accurately the LLM predicted the winning message when using that hypothesis. When hypotheses produce incorrect classifications, those examples are used to prompt the LLM for additional hypotheses. Inspired by bandit algorithms, the system is designed to efficiently explore the space of hypotheses. Liu et al. (2025) extended this framework by incorporating summaries of selected research papers into the prompting process, alternating between prompts that include paper summaries and prompts that focus on data patterns.

Together, these papers offer different data-driven approaches to generating hypotheses from unstructured data. As these methods mature and become more accessible, a natural question emerges: Once we have generated multiple hypotheses, how do we systematically refine and prioritize them for empirical testing?

Our approach addresses this challenge by introducing a novel refinement process. We generate initial hypotheses using straightforward prompting—though any of the above approaches could also be used—then refine them by rewriting headlines to incorporate hypothesized features and ranking the results using a machine learning model trained to predict how linguistic changes affect engagement. The approach works with hypotheses from any source: those generated using computational methods like the ones above, extracted from literature, or developed through traditional creativity and insight.

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 input once it is programmed. Nevertheless, the output is a set of marketing hypotheses readily interpretable by humans.

Our contributions are threefold: First, we introduce a method to convert unstructured text into marketing insights (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 becomes digitized—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, providing a systematic approach for AI-assisted discovery of actionable hypotheses (De Freitas, Nave, and Puntoni 2025; Berger, van Osselaer, and Janiszewski 2025). More broadly, this method contributes to the research on data-driven discovery and hypothesis generation (McGuire 1997; Ludwig and Mullainathan 2024; Banker et al. 2024; Aka, Bhatia, and McCoy 2023; Adolphs et al. 2016; Zhou et al. 2024; Movva et al. 2025).

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 2024; Lee, Hosanagar, and Nair 2018; Berger, Moe, and Schweidel 2023; Berger, Kim, and Meyer 2021). Using our framework, we uncover new substantive 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.

Third, 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 a new method for

how to aggregate insights from thousands of A/B tests in the form of specific hypotheses that others can carefully test.

As a guide to the rest of the paper, we first describe the application and data used in this paper (Section 2). We then conduct a preliminary analysis to quantify the 'predictive gap' between existing knowledge and additional, semantic signal contained in the text, motivating our data-driven approach (Section 3). Afterwards, we detail our three-step framework—Hypothesize, Intervene, and Predict—for discovering and refining these hypotheses (Section 4). We then test a prioritized subset of the generated hypotheses on a hold-out set of experiments (Section 5). We conclude by discussing our findings, the framework's broader implications, and directions for future research (Section 6). Data and materials related to the studies conducted with human participants are summarized in the Appendix and available on the Open Science Framework (OSF).[3] Online experiments with human participants were reviewed by the Institutional Review Board at a U.S.-based university (IRB22-1611).

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;

[3] https://osf.io/d5xvb/?view_only=d58d4e38d43948eb8d87d25a513300f0

Deolankar et al. 2024).[4]

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 in our data, 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 and when this happens in a randomized-controlled trial, it suggests 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.[5] 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

We applied several preprocessing steps to prepare the data for analysis. First, we cleaned the raw text following standard procedures (e.g., removing non-visible characters and replacing non-ASCII characters with ASCII equivalents; see also Table 3 in Berger et al. 2020). For cases where multiple treatment arms in a trial had identical headlines (e.g., where only the image varied), we collapsed the rows, summing clicks and impressions.

[4]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)

[5]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, Section **??**.

Data Partitioning. We partitioned the data into four splits: training (40% of trials), intervening (10%), testing (10%), and a *lock-box* hold-out set (40%; to be used as the test set once the paper has undergone the peer review process).[6]. Because headlines were sometimes reused across trials, we first needed to group *trials* with overlapping headlines into "components."[7] We then partitioned using these components to minimize data leakage (Kapoor and Narayanan 2023; Egami et al. 2022; Ludwig, Mullainathan, and Rambachan 2025). More details into these partitions are provided in the Appendix, Section 1.

Outcome Definition. We measured engagement using the click-through-rate (CTR), 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.[8] For any headline H_a, we define this as the smoothed CTR estimate:

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

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

Given the experimental setup of the data, where multiple headlines were tested for the same story in randomized trials, we structure our analysis at the pair level. Altogether, our dataset contains over 282,000 unique pairs of headlines from over 17,000 trials across all splits (more details available in the Appendix). This pairwise approach effectively controls for topic-level confounds by focusing on the differences between headlines written for the same content.

[6] By including the lock-box set, the current manuscript effectively serves as a registered report (Nosek and Lakens 2014; Chambers and Tzavella 2021; Urminsky and Dietvorst 2024). A similar lock-box practice was implemented in Ludwig and Mullainathan (2024).

[7] 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, Trials 1-3 would all be assigned the same component.

[8] Other reasonable approaches 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.

Our primary outcome of interest is, therefore, the difference in CTRs between two headlines, H_a and H_b, from the same trial:

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

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

To analyze the headline text, we constructed three distinct sets of features. The first, a high-dimensional semantic representation of the text, serves as the primary input for our machine learning algorithm. The other two—a set of existing psycholinguistic features and a measure of human intuition—provide a crucial baseline of existing knowledge. This baseline allows us to later evaluate the novelty of the hypotheses our framework generates by comparing its discoveries to what is already known in the literature or intuitively understood by human judges.

Semantic Representation. We converted the raw headline text to its high-dimensional semantic representation using a pre-trained MPNet model (Song et al. 2020), which converts text into a vector of length 768 (for additional details, see Appendix).[9] We used these embeddings as inputs for our machine learning algorithm and to measure textual similarity and diversity.

Existing Psycholinguistic Features. To establish a baseline of known psychological effects, we replicated the feature set from Banerjee and Urminsky (2024) ("BU"), who mapped 51 psychological constructs to headlines using LIWC (Tausczik and Pennebaker 2010), TextAnalyzer (Berger, Sherman, and Ungar 2020), and curated word lists. This feature set includes constructs that Banerjee and Urminsky (2024) have shown affect click-through rates

[9]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

in this dataset, such as *reading ease*, *numeric reference*, and *visual language*.[10]

Human Labels. Despite the many constructs captured in BU's features, it is possible that some of what is known is not reflected in this set. Rather than enumerating additional constructs or creating new dictionaries (Humphreys and Wang 2018)—processes that are expensive in both time and money—we took another approach to capture hypotheses humans might intuit. We showed participants a pair of messages written for the same story and had them choose which one they thought performed better in A/B an test (for a similar approach, see Ludwig and Mullainathan 2024). By having participants choose which headline they think will perform better, we capture what they believe without them needing to put it into words (Malt et al. 1999; Batista et al. 2024). To capture these guesses, we conducted a study where 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. Each participant completed 10 "training" rounds, with feedback, and 30 "test" rounds, where they were paid based on accuracy (additional details available in the Appendix; data and materials are available on OSF). The resulting judgments, aggregated from these guesses, capture what is intuitively "known" about headline effectiveness and, alongside BU's psychological features, serve as useful benchmarks for our discovery process.

IS THERE ANY SIGNAL IN THE TEXT LEFT TO DISCOVER?

Textual cues have been shown to motivate engagement. In fact, using the Upworthy Research Archive, Banerjee and Urminsky (2024) test the effects of more than 50 psychological constructs. It is, therefore, reasonable to ask: is there anything left to discover? And, if so, how much?

To answer these questions, we first assess how well we can explain the outcome using

[10]Note that while we use the same features, our data splits and modeling specifications differ. Therefore, results may appear inconsistent with BU's work.

"what we already know," captured in features such as the ones used by Banerjee and Urminsky (2024), and then compare that to how much of the outcome is explained by a black-box algorithm, trained on sentence embeddings from past messages and outcomes.[11]

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 (G), or both combined. We report all our results out of sample to accurately compare the performance of the various models (see Table 1).[12]

We formed the BU predictor as follows. For each of the 51 psychological constructs used in BU's analyses, we take the difference in construct values between the headlines in each pair. The result is 51 features defined as the difference in a psychological construct (such as reading ease, numeric reference, or visual language). We then estimate an OLS regression of the form

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

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

[11] The intuition for this exercise is the following: the black-box algorithm approximates how much of the variation in the outcome is *explainable* given information contained in the text. If the black-box algorithm outperforms the known feature model, it is reasonable to suspect that it is picking up some information the known feature model is not. Furthermore, by combining the two models and measuring its performance, we could evaluate whether the information captured in the black-box algorithm is redundant or complementary to the known features model.

[12] 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.

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, 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 1).

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.

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 the section on Additional Features above for a description 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%]).

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 1, 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.[13] 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.

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

[13] 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, $F(2, 1689) = 6.70, p = .001$.

Table 1: 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 (linear)	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 (linear) + Human guess	0.049	0.136	0.568	0.629	0.608	0.692
ML only	—	0.130	—	0.639	—	0.687

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 captured through the ML algorithm is uninterpretable to human researchers.[14]

The current 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 prioritized to be tested. Throughout the process it should be apparent where the algorithm played a role (and where the human did). Our process consists of three steps: hypothesizing, intervening, and predicting (see Figure 1). For each step, we explain the procedure and the output (additional checks are reported in the Appendix).

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 (for link, see Introduction).

Step 1: Hypothesize

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"

[14]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.

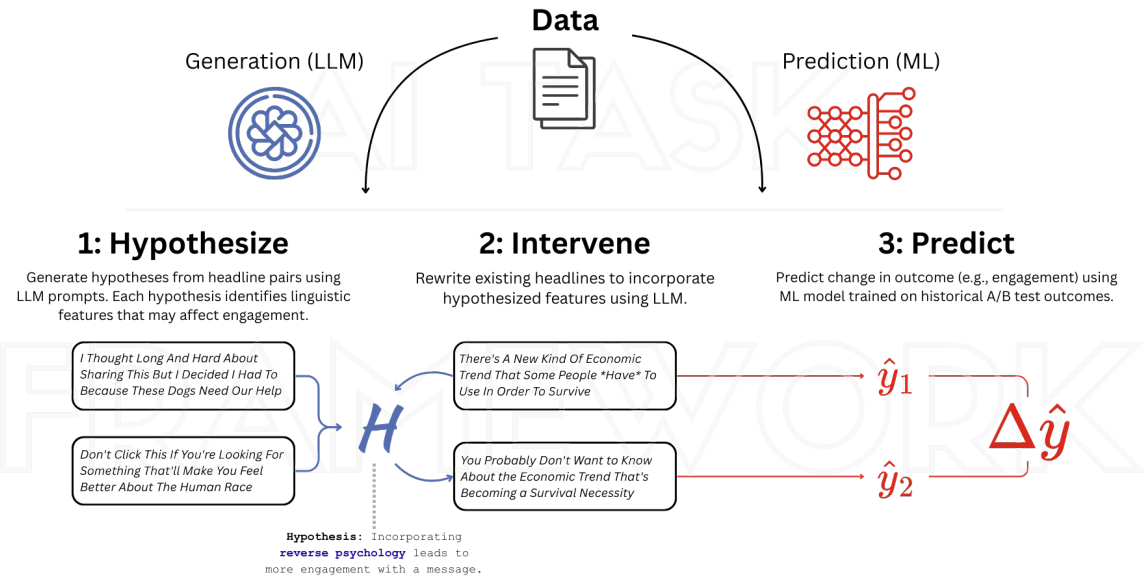

Figure 1: Overview of steps for generating and refining hypotheses

that could later be tested.[15]

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 (Adams et al. 2021), they are prone to confirmatory biases (Hartzmark, Hirshman, and Imas 2021; Bhatia 2014; Klayman and Ha 1987), and they may be constrained by beliefs that their creativity is finite (Lucas and Nordgren 2020). LLMs, instead, offer a way to do this task at scale while generating a diverse set of hypotheses.[16]

Procedure. We used OpenAI's GPT-4-Turbo to generate 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 headlines in each pair were always from the same trial.[17] We then

[15] 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).

[16] We find that LLM-generated hypotheses are at least as semantically diverse as those produced by humans (see Appendix, Section 5).

[17] 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.

restricted the sample to pairs with large predicted differences, specifically those from the top quartile of absolute predicted differences ($|\widehat{m}_{a,b}|$). From this subset, we randomly drew one pair per "component" (ensuring at most one pair per trial; see section on "Data Pre-Processing" above) in the *training* set to minimize leakage across steps.

Each pair was randomly assigned to one of five model temperatures and one of 288 prompt combinations (see Appendix, Section 2 for full details).[18] All prompts specified the hypothesis should be: (i) clear, (ii) generalizable, (iii) empirically plausible, (iv) unidimensional, and (v) usable.

Output. This step produced 2,100 hypotheses that appear to be clear and coherent (see examples in Table 2). To further assess the quality and gather a set of independent judgments, we recruited 79 participants on Prolific (prolific.com) [June 2024] who evaluated 106 hypotheses across several dimensions including clarity and generalizability. On every dimension, the average rating for most hypotheses was above the scale's midpoint. Importantly, raters believed every hypothesis could be applied to at least one other context beyond headlines, such as "product descriptions" or "billboard advertisements", suggesting broad applicability (see Appendix, Section 4).

Step 2: Intervene

The goal of the second step is to simulate *what if* a message were written according to a given hypothesis. Whereas a traditional experimental approach to evaluating thousands of hypotheses would require thousands of individual A/B tests, our approach simulates this process for every hypothesis at scale. That is, this step systematically creates a *counterfactual* version for a wide range of headlines, each rewritten to reflect one of the hypotheses generated in Step 1. This intervention allows us to ask: "What might this headline have looked like if it had incorporated this specific linguistic feature?" Intervening in this way allows us to then

[18] Prompts varied across roles (e.g., "editorial strategist"), hypothesis formats (e.g., "Hypothesis: X leads to more engagement"), and instructions. Although most prompts included specific headline pairs, some omitted them to serve as controls.

Table 2: 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 the OSF page provided at the end of the Introduction.

use the black box algorithm to assess how changing the text changes the predicted outcome.

Procedure. We again used OpenAI's GPT-4-Turbo with randomized model temperatures (.75; .9) to rewrite a random set of original Upworthy headlines (from the *Intervening* set) according to each of the 2,100 hypotheses. Each hypothesis was applied to approximately 70 different headlines, producing over 252,000 counterfactual headlines in total. Specifically, we provided the LLM with one hypothesis matched to one headline and a prompt that included three additional Upworthy headlines (for "few-shot learning", e.g., Min et al. 2022; Brown et al. 2020). The LLM was given instructions to rewrite the headline to incorporate the hypothesized feature while keeping the topic and meaning as similar as possible to the original (see Appendix, Section 2 for more prompt details). For instance, given the original headline, "There's A New Kind Of Economic Trend That Some People 'Have' To Use In Order To Survive," and the hypothesis, "Incorporating reverse psychology leads to more engagement," the LLM produced the rewritten, or counterfactual, headline: "You Probably Don't Want To Know About The Economic Trend That's Becoming A Survivor Necessity."

Output. The output contained 252,156 new headlines.[19] We then applied quality filters, removing headlines that exceeded 100 characters (reflecting Upworthy's apparent limits) and those generated from "control" prompts or from instructions to minimize the hypothesized feature.[20] Table 3 provides a set of examples.

In a set of follow-up exercises, we assessed whether the new, counterfactual headlines successfully incorporated the intended features while still resembling the originals. Using an

[19] In generating these counterfactual headlines, 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 counterfactual headlines, which resulted in a uniform sample of headlines for each hypothesis.

[20] Specifically, we removed 37,430 counterfactual headlines (17.85 per hypothesis) that were longer than 100 characters, 956 counterfactual headlines corresponding to hypotheses produced by "Control" prompts (i.e., that did not refer to an Upworthy headline), and 63,119 counterfactual headlines (30.1 per hypothesis) generated in response to prompts that instructed GPT to "dial down" the hypothesized feature. In post-hoc analysis, headlines corresponding to Control hypotheses and those meant to dial down features were predicted by the ML algorithm to be consistently worse than matched original headlines.

Table 3: Examples of hypotheses, original headlines and the associated rewrites

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 the OSF link provided at the end of the Introduction.

LLM to rate the presence of hypothesized features on a scale of 0 to 7, we found that when prompted to vary a specific feature, 53% of new headlines were rated as having *more* of the hypothesized feature compared to the original used to create it (40% of new headlines had the same value as the original headline, and only 7% had less of the hypothesized feature). As a comparison, when examining features we weren't trying to change, these features increased only 24% of the time, remained unchanged 57% of the time, and decreased 19% of the time. These results indicate the rewriting was working as intended, albeit imperfectly.[21] We also checked that the counterfactual headlines remained "in-distribution"—that is, similar enough to authentic Upworthy headlines in meaning and form for reliable ML predictions. By analyzing the distance in embedding space, we find that counterfactuals were semantically similar to their original headlines; indeed, they were more similar than two original headlines from the same trial were to each other. Furthermore, when human participants evaluated the AI-generated messages, they generally believed them to be written by humans and not AI-generated. While participants could detect some differences when directly comparing sets of headlines, both counterfactual and original headlines were broadly perceived as legitimate Upworthy content (see Appendix, Section 4 for more details).

Step 3: Predict

The final step predicts how each hypothesis might affect engagement when implemented. Having created counterfactual headlines in Step 2, we now need to evaluate their likely performance. The ML algorithm provides a systematic way to predict treatment effects at scale, leveraging patterns learned from thousands of actual A/B tests. Importantly, this step does not substitute for experimentation but rather provides an educated guess to guide which hypotheses merit rigorous testing. This approach is efficient and scalable, enabling us to rank all hypotheses by their predicted impact before investing in field experiments.

[21] We suspect part of the reason some features were not incorporated is because they may have been harder to implement than others (e.g., 'reference to animals' versus 'question-asking'). We considered filtering out cases where the feature was not properly implemented but doing so would likely bias the process towards hypotheses with easily implementable features, potentially missing important insights about real-world applicability.

Procedure. We used the ML algorithm to predict $\widehat{\Delta\text{CTR}}$ for each counterfactual headline relative to its original ($\widehat{m}_{a,b}$), generating *predicted* treatment effects (PTEs) for every counterfactual-original pair.[22] Since we are primarily interested in each hypothesis's average effect across diverse contexts, we aggregated PTEs at the hypothesis level. This produces an average PTE for each hypothesis, capturing the effect it is predicted to have when applied to a random set of headlines covering different topics. For this step, we excluded eight hypotheses that were generated in Step 1 without referencing an actual Upworthy headline (what we referred to above as the "Control" prompts).

Output. This process generated average PTEs for all 2,092 hypotheses, creating a ranked list from most positive PTEs to most negative PTEs. The average PTE across all hypotheses was $-.00065$ (SD = .00086), which translates to approximately a 4% decrease in CTR relative to baseline. This suggests that, on average, a hypothesized feature applied to a randomly selected headline was likely to decrease engagement. This negative average may reflect several factors: the challenge of applying linguistic features irrespective of the original message's content, the difficulty of improving upon already well-crafted Upworthy headlines, and potential limitations in how effectively the LLM implements certain hypotheses. Nevertheless, there was substantial *variation* in predicted effects indicating meaningful differences between hypotheses; some hypotheses were predicted to substantially increase engagement while others were predicted to decrease it.[23]

Since we applied hypotheses uniformly across several content types regardless of fit, these predictions likely underestimate their potential effects. What we gain from this approach, however, is a more generalizable prediction. These predictions provide an empirical basis for prioritizing hypotheses for testing.

[22] This simulates the result of an A/B test conducted alongside the original Upworthy experiments—we have no actual CTR measurements for GPT-generated headlines, only ML predictions ($widehatm_{a,b}$)

[23] The standard deviation of $\widehat{\Delta\text{CTR}}$ at the 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 counterfactuals with nearly as much variation as we see in predictions from the original dataset.

Additional Filtering

Although this systematic ranking provides valuable guidance, two practical challenges remain before selecting hypotheses for testing: identifying redundant hypotheses (i.e., two or more hypotheses representing similar concepts) and accounting for noise in the ranking process. The following approach addresses these challenges through computational filtering.[24]

Addressing Redundancy Through Clustering. While our generative process produced over 2,000 hypotheses, many were conceptually similar. For instance, our process generated the hypotheses "posing a direct question to the reader" and "directly addressing the audience and asking a question," which appear to represent the same core idea. Simply testing hypotheses from the top of the unfiltered, ranked list would therefore result in testing multiple versions of the same concept.

To ensure conceptual diversity while still prioritizing the most promising hypotheses from our ranking, we implemented a sequential selection strategy. We began with the full list of hypotheses, rank-ordered by their average PTE. Starting with the top-ranked hypothesis, which we selected first, we iterated down the list. We evaluated each subsequent hypothesis for its functional similarity to those already selected. If a candidate hypothesis was deemed redundant with any previously selected one, we excluded it. If it was sufficiently distinct, it was added to our refined set, and the process continued.

Hypotheses were grouped based on the sort of change they induced when applied to an existing message. That is, rather than grouping hypotheses based on their description, we grouped them based on what they *did* to a message. To do this, we computed transformation vectors for each hypothesis by averaging the embedding differences between the original headlines and the counterfactuals generated during the Intervene step. These vectors capture

[24]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).

not what hypotheses say (e.g. "posing a direct question to the reader"), but rather the average semantic change between a message with and without the hypothesized feature (e.g., between a message that does and does not pose a direct question to a reader). We then used cosine similarity to determine if a candidate hypothesis was too similar to any in our selected set.[25] By selecting the highest-ranked representative for each cluster, this approach reduces redundancy without sacrificing the most promising candidates identified in the Previous step.

Handling Ranking Noise Through Significance Testing. While average PTEs provide systematic rankings, they may contain noise from the counterfactual generation and prediction processes. To identify hypotheses with robust predicted effects, we conducted statistical tests on the distribution of PTEs underlying each hypothesis's average. For each hypothesis, we performed one-sample t-tests to determine whether the set of PTEs was significantly different from zero, applying false discovery rate correction to account for multiple comparisons (Benjamini and Hochberg 1995). This filtering prioritizes hypotheses where the predicted effects appear consistent across multiple implementations, suggesting the feature's impact is not confined to a specific headline's topic or style.

Output. These two filtering stages systematically refined the set of 2,092 hypotheses. The clustering procedure first reduced the list to 205 conceptually distinct hypotheses (for the full set of hypotheses including their clusters, see the "hypothesis" dataset on OSF). From this set, significance testing identified all hypotheses with robust PTEs. After correcting for multiple comparisons, this yielded 16 hypotheses with significant positive effects (displayed in Table 4) and 114 with significant negative effects. From this final, refined pool of candidates, we selected a subset to test out-of-sample which we describe in the following section.

[25] Whether or not two hypotheses are similar is determined computationally according to a distance parameter, ε, which we set at .03. 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.

Table 4: 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 + Curiosity	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.*

Discussion

This process has produced a set of interpretable hypotheses, each with an average Predicted Treatment Effect (PTE) that estimates its potential impact on engagement across a diverse range of messages. The strength of this approach lies in its ability to systematically transform unstructured text from thousands of marketing experiments into a prioritized list of testable ideas. It achieves this by leveraging a machine learning model, trained on historical outcomes, to predict the effect of thousands of simulated interventions at scale. Importantly, these steps—Hypothesize, Intervene, and Predict—constitute a data-driven framework for generating hypotheses before any confirmatory tests are conducted, offering a structured and scalable engine for discovery.

The framework described above relies on several strategic decisions and assumptions. In the Hypothesize step, for instance, we presented headlines to the LLM in pairs rather than as a single large batch. This choice was informed by our own evidence that a pairwise approach generates a more semantically diverse set of hypotheses than presenting headlines in aggregate (see Appendix, Section 5). We also opted to sample each pair only once, though future work could sample the set of messages with replacement to generate several hypotheses from a single pair of messages. Our choice of OpenAI's GPT-4-Turbo reflects the state of the art at the time of our study, but the framework is LLM-agnostic. To mitigate sensitivity to prompting—a known challenge with LLMs (Ludwig, Mullainathan, and Rambachan 2025)—we systematically varied the prompt structure and randomly assigned pairs to one of 288 unique prompt combinations.

During the Intervene step, the rewriting process necessarily balanced two competing goals: meaningfully incorporating the hypothesized feature while ensuring the new headline remained similar to the original data. The tension between these goals is reflected in our results. While a modest majority (53%) of rewrites successfully incorporated the target feature, a significant portion (40%) did not, and human raters could detect subtle differences between the original and AI-generated versions. Despite these imperfections, the rewrites

remained semantically faithful to their originals and were validated as authentic and of equivalent quality in a series of pre-registered equivalence tests (see Appendix, Section 4).

Finally, the Predict step relies on the critical assumption that the counterfactual headlines are sufficiently similar to the original data for the ML model's predictions to be valid. We conducted several pre-registered experiments and computational checks to support this assumption but our approach has no guarantees that this will always hold. Furthermore, our Siamese network architecture was chosen specifically to model the change in CTR, a structure that naturally aligns with the A/B test data from which the model learns but other modeling approaches for different datasets could yield a different set of results.

One might also be concerned that the ML model in our Predict step functions as an imperfect verifier, which could risk identifying 'false positives'—hypotheses that our ML model favors but that would fail in real-world tests (Stroebl, Kapoor, and Narayanan 2024). However, our framework has two key structural safeguards that mitigate this concern. First, the generator (the LLM) is decoupled from the verifier (the ML model); it has no access to the model's predictions and thus cannot learn to exploit its weaknesses. Second, by averaging predictions across dozens of interventions for each hypothesis, we prioritize generalizable patterns over single, anomalous predictions. These design choices reinforce that the framework is a tool for prioritization, not a substitute for empirical validation.

The filtering stage introduces additional methodological choices. Our clustering approach groups hypotheses by their functional transformations rather than textual similarity, based on the assumption that hypotheses creating similar semantic changes may operate through related mechanisms. The choice of distance threshold ($\varepsilon = .03$) balances preserving diverse concepts while removing redundant ones, though other thresholds could group hypotheses differently. The significance tests apply conservative corrections for multiple comparisons, prioritizing effects that appear robust over including more hypotheses. However, this approach has important limitations. With 114 hypotheses showing significant negative effects, the significance test may conflate genuinely harmful features with those that are simply

poorly matched to the diverse content types they were applied to. For instance, a hypothesis about humor might show consistent negative PTEs not because humor inherently decreases engagement, but because applying humorous language to serious news stories reliably fails. This highlights a challenge for any method that aggregates effects across varied content: a consistently negative signal may indicate a feature to avoid, or it may point to an important, yet-unspecified, interaction.

Ultimately, the framework and subsequent filtering procedures are designed to systematically convert a large corpus of text into a small, curated set of testable insights. This entire process is not an end in itself, but a means to generate promising, data-driven questions for empirical study. Having produced a refined pool of hypotheses ordered by their predicted effects, we handpicked a subset for rigorous out-of-sample testing (see Table 4). The following section describes how we tested these hypotheses on a hold-out set of experimental data.

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 (Egami et al. 2022). 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 for 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 *test* 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 counterfactual headlines. This set was used previously only to gather human labels (see subsection above where we describe Additional Features). 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 test 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*.[26]

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

[26]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.

from the 117 original headlines used for generating counterfactual headlines and the 404 counterfactual headlines corresponding to each of the six selected hypotheses. These ratings were meant to check whether the counterfactual generating procedure indeed modified 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, Subsection 4.2.5).

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 test set, 4,212 for original headlines used to generate counterfactual messages, and 14,988 for counterfactual headlines themselves). Each headline was rated on each feature a median of six times (IQR: 4, 8).

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 5 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 (see Figure 2). A fifth had a marginal effect ($p = .094$). Of these five, all

Table 5: 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.

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.[27]

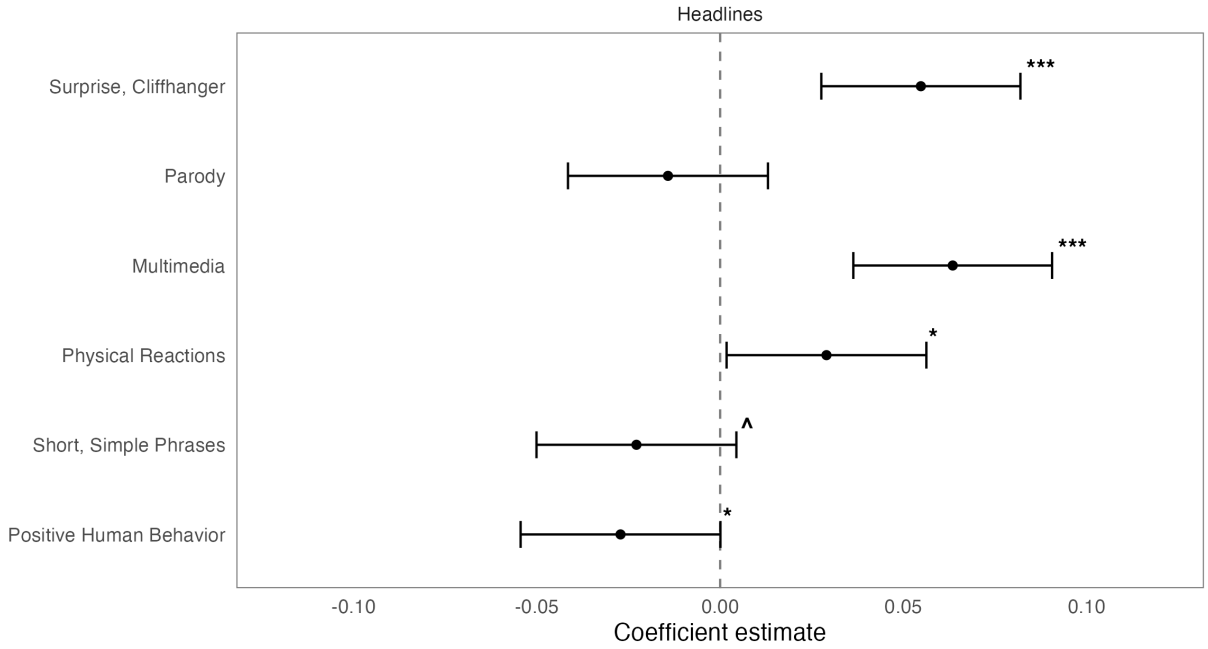

Figure 2: Coefficient estimates for hypotheses where the outcome is ΔCTR for Headlines. Values shown are pulled from regressions, where a one unit change in variable corresponds to a $\widehat{\beta}$-unit change in the standard deviation of CTR.

Through this process, we have uncovered several hypotheses. 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 (2024), which was estimated on the training partition (i.e., $\widehat{BU}$). The results of these regressions are available in Table 6. 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 multivariate model, where three of the six features are significant predictors (ps $< .05$).

[27] 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.

Table 6: 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.

In another specification, we compare a baseline model that regresses ΔCTR on all 51 features from Banerjee and Urminsky (2024) (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 7.

Table 7: 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.

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.[28] 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 of the six features—*surprise, cliffhanger* and *multimedia reference*—appear to capture information that is sufficiently different from the 51 psychological constructs derived in Banerjee and Urminsky (2024). 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.

GENERAL DISCUSSION

This paper introduces a framework—Hypothesize, Intervene, and Predict—that researchers and marketers can use to discover and prioritize testable insights from unstructured text data. Our approach leverages a large-language model (LLM) to first *Hypothesize* linguistic features from observed text patterns. An LLM then acts as a tool to *Intervene* by rewriting messages to incorporate these hypothesized features. Finally, a machine learning (ML) model, trained on historical behavioral data, *Predicts* the change in engagement resulting

[28] 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).

from these interventions. Using a dataset with over 60,000 headlines and 32,000 A/B tests, we generated and refined hundreds of hypotheses about what features of language affect engagement, then validated a subset out-of-sample on nearly 2,000 real-world experiments. We found causal evidence supporting five out of six hypotheses. For instance, describing physical reactions significantly increases engagement, while focusing on positive aspects of human behavior decreases it. The framework's innovative aspect lies in the steps for refining the set of hypotheses—using LLMs to create counterfactuals and ML to predict treatment effects—which allows researchers to learn from past campaigns and prioritize hypotheses *before* investing in new experiments.

Contributions

Toolkit for Discovery. This framework has several implications for how researchers approach discovery. It provides a systematic workflow for converting large text corpora into testable hypotheses. Researchers working with text messages, social media posts, customer reviews, or online ads now have a replicable process: generate hypotheses using LLMs, create counterfactuals that operationalize those hypotheses, and use ML to predict which hypotheses warrant empirical testing. Automated content analysis such the approach proposed in this paper, allows researchers to "cast a wider net" during discovery (Berger, van Osselaer, and Janiszewski 2025). Our framework goes further by addressing a critical next step: once researchers have cast that wider net, how do they prioritize which ideas to investigate? The Intervene and Predict refinement steps offers a principled solution, systematically evaluating hypotheses against patterns in historical data before committing resources to new experiments. The framework is modular by design, meaning researchers can generate hypotheses using one dataset and refine them using another. Importantly, this approach nudges researchers to be deliberate and transparent about their decisions. To use this framework, researchers must specify their modeling approach, prompts, and clustering thresholds. Making these choices explicit not only strengthens the scientific rigor of the discovery process

but also makes it reproducible, allowing other researchers to adapt the framework while documenting their own methodological decisions.

Insights into Consumer Language. Beyond the methodological contributions, this work generates substantive insights about how language affects engagement. We identified dozens of hypotheses predicted to have significant effects on click-through rates, and when tested out-of-sample, many showed the predicted patterns. For instance, describing physical reactions significantly increases engagement, while focusing on positive aspects of human behavior decreases it. Yet these findings represent only the beginning of what is needed to build comprehensive understanding. As readers of this journal well know, identifying an effect is valuable, but substantial work remains to understand its mechanisms, establish its generalizability and boundary conditions, and explore potential downstream consequences. Why do physical reactions increase engagement? Under what conditions does this effect hold or reverse? How does it influence not just clicks but sustained attention or purchase behavior? Answering these questions are still the work of a human researcher. By systematically generating then screening hundreds of potential hypotheses and prioritizing those with the strongest predicted effects, we reduce the search costs of discovery. This allows researchers to focus their efforts on fewer, more promising candidates for deeper investigation.

Helping Organizations Learn from Past Experiments. Organizations today continuously run thousands of A/B tests to optimize their messaging, yet many fail to aggregate insights contained across experiments. Individual tests might reveal that one headline outperforms another, but without a systematic approach to aggregating findings, organizations learn *what* works in isolated instances, often, without any insight into *why*. Our framework offers a practical solution to this challenge. By applying the Hypothesize-Intervene-Predict process to historical A/B test data, organizations can search for patterns across multiple experiments. As in the case of Upworthy, organizations often change several elements of a message at once. As academic researchers will know, this poses a challenge for *testing*

a specific hypothesis because changing many things at once introduces several confounds. This bug, however, turns out to be a feature when it comes to discovery. More variation across messages creates a richer dataset for pattern detection. While any single test remains ambiguous about which change mattered, analyzing patterns across hundreds of such tests reveals which features consistently predict engagement. The framework thus transforms accumulated A/B test data from isolated findings into generalizable principles of message effectiveness. Furthermore, the framework's flexibility enhances its practical value. Marketing teams can generate hypotheses from competitor messaging or public datasets, then refine them using internal ML models trained on proprietary data containing text and outcomes. This expands the space of possible discoveries beyond what exists in a single organization's data. Companies with multiple brands, products, or customer segments can generate hypotheses once and refine them separately for each context, building a library of insights about when and where different linguistic features may be driving engagement. In this way, the framework transforms the common practice of A/B testing from an optimization tool into an engine for strategic learning.

Limitations and Future Directions

Our framework's primary strength—and its main limitation—is that it is inherently data-driven. Unlike theory-driven approaches that start with a model of the world, our method begins with an observation in the data. The benefit is that we know the effects we prioritize for testing exist; subsequent research must then focus on understanding the effects' generalizability, causes, and consequences. The downside, however, is that the scope of discovery is necessarily constrained by the patterns contained within the initial dataset. This core characteristic has several important implications for the types of hypotheses we generate and how they should be interpreted.

First, the hypotheses our framework produces tend to be more substantive, or applied, than theoretical, or basic (Lynch et al. 2012; Blanchard et al. 2022). In designing our process,

we deliberately aimed to generate "empirically plausible" hypotheses that could be observed in the data without requiring extensive background knowledge (Ludwig and Mullainathan 2024). This choice carries a trade-off: the resulting hypotheses are simpler to communicate and easier to test with available data, but they do not come with any accompanying theory. As De Freitas, Nave, and Puntoni (2025) observe, current generative AI models are better suited for producing incremental "small ideas" rather than groundbreaking "big ideas" (see also Lee and Chung 2024). Our framework reflects this reality, serving as a powerful engine for discovering and refining these valuable, though often less radical, insights. The framework is modular, however, and could be adapted to generate more theoretically-grounded hypotheses. For instance, one could fine-tune an LLM on specific academic literature (Banker et al. 2024) or iteratively incorporate data alongside existing literature into the prompting process (Liu et al. 2025). Similarly, while we constrained our hypotheses to be simple and mostly unidimensional to maximize interpretability, the framework could be modified to generate more complex, interaction-based hypotheses. Early prompts we tried, for instance, produced dynamic hypotheses suggesting to "begin with a positive emotional state and then transition to a negative one, depicting a journey of emotional upheaval." Whether such complexity improves or hinders practical use remains an open question worth exploring.

Second, without a guiding theory, it can be difficult to contextualize data-driven findings or predict when they might generalize. For example, our results show that referencing multimedia significantly increases engagement in Upworthy headlines but decreases it in social media posts (see Appendix, Section 7). A purely data-driven approach can identify this reversal, but it cannot explain it. Is the difference due to platform norms, audience expectations, or the type of content? These questions underscore a crucial and broader point: data-driven discovery does not eliminate the need for theory. If anything, it makes theory more critical. Without a theoretical lens to connect different insights, our findings risk "talking past" one another.

Science requires both data-driven and theory-driven approaches (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). Our framework is firmly a tool for the former. By reducing the costs of discovery, it does not replace the creative and theoretical work of human researchers; rather, it augments it. The transparency of our outputs leaves room for human researchers to examine the generated hypotheses with an eye for theoretically interesting patterns. Building theory from these patterns remains fundamentally human work.

REFERENCES

Abbott, Andrew Delano (2004), *Methods of discovery: heuristics for the social sciences* Contemporary societies, New York: W.W. Norton & Co.

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) 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.

Aubin Le Quéré, Marianne and J. Nathan Matias (2025), "When curiosity gaps backfire: effects of headline concreteness on information selection decisions," *Scientific Reports*, 15 (1), 994 https://www.nature.com/articles/s41598-024-81575-9, publisher: Nature Publishing Group.

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 (2024), "The Language That Drives Engagement: A Systematic Large-scale Analysis of Headline Experiments," *Marketing Science* https://pubsonline.informs.org/doi/full/10.1287/mksc.2021.0018, publisher: INFORMS.

Banker, Sachin, Promothesh Chatterjee, Himanshu Mishra, and Arul Mishra (2024), "Machine-assisted social psychology hypothesis generation," *American Psychologist*, 79 (6), 789–797.

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/.

Berger, Jonah, Stijn M. J. van Osselaer, and Chris Janiszewski (2025), "Casting a Wider Net: Using Automated Content Analysis to Discover New Ideas," *Current Directions in Psychological Science*, page 09637214251315716 https://doi.org/10.1177/09637214251315716, publisher: SAGE Publications Inc.

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.

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.

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.

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.

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) https://www.pnas.org/doi/abs/10.1073/pnas.2021636118.

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.

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 Freitas, Julian, Gideon Nave, and Stefano Puntoni (2025), "Ideation with Generative AI—in Consumer Research and Beyond," *Journal of Consumer Research*, 52 (1), 18–31 https://doi.org/10.1093/jcr/ucaf012.

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.

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.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart (2022), "How to make causal inferences using texts," *Science Advances*, 8 (42) https://www.science.org/doi/full/10.1126/sciadv.abg2652.

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.

Glaeser, Edward L. "Researcher Incentives and Empirical Methods," (2006) https://www.nber.org/papers/t0329.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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, Byung Cheol and Jaeyeon (Jae) Chung (2024), "An empirical investigation of the impact of ChatGPT on creativity," *Nature Human Behaviour*, 8 (10), 1906–1914 https://www.nature.com/articles/s41562-024-01953-1, publisher: Nature Publishing Group.

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.

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.

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.

Liu, Haokun, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan "Literature Meets Data: A Synergistic Approach to Hypothesis Generation," (2025) http://arxiv.org/abs/2410.17309.

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.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan "Large Language Models: An Applied Econometric Framework," (2025) https://www.nber.org/papers/w33344.

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

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.

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.

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.

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.

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.

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.

Movva, Rajiv, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson "Sparse Autoencoders for Hypothesis Generation," (2025) http://arxiv.org/abs/2502.04382.

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.

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.

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.

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.

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.

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.

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.

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.

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) https://www.science.org/doi/10.1126/sciadv.adn2555.

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.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu "MPNet: Masked and Permuted Pre-training for Language Understanding," (2020) http://arxiv.org/abs/2004.09297.

Stroebl, Benedikt, Sayash Kapoor, and Arvind Narayanan "Inference Scaling fLaws: The Limits of LLM Resampling with Imperfect Verifiers," (2024) http://arxiv.org/abs/2411.17501, arXiv:2411.17501 [cs].

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.

Toubia, Olivier and Oded Netzer (2017), "Idea generation, creativity, and prototypicality," *Marketing science*, 36 (1), 1–20.

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.

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.

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.

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.

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

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 Large Language Models to Generate and Refine Hypotheses from Text

October 9, 2025

CONTENTS

WEB APPENDIX 1: PREPARING THE DATA

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 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.

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, also Egami et al. 2022; Ludwig, Mullainathan, and Rambachan 2025). Therefore, we resampled the complete set into new splits by "component", which we defined so that we could group *trials* with overlapping headlines.[1] This ensured that headlines repeated within and across trials were contained within the same split. Table 1 displays the headline, trial, and component counts by partition. 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 **??**. 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 *generating counterfactual* set (10% of trials; 3,366 trials). This set was used to produce counterfactual headlines, described in Section **??**.

- 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. 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.[2]

[1] 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 2023), Trials 1-3 would all be assigned the same component.

[2] 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).

Table 1: Counts for Headline Data

	Splits				
	Train	*Intervene*	*Test*	*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.

Outcome Definition

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.[3] For simplicity, we refer to Smoothed

[3]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.

CTR as CTR in the remainder of this paper. Table 1 provides a summary of the outcome measure at the headline-level, pair-level, and trial-level.

Analysis Structure

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.
- Generating Counterfactual set (10%): 29,600 unique pairs from 1,807 trials. However, because the rewriting of headlines is done at the headline level, the pairwise dataset is not used for the generating counterfactual 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.

1.4.1 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.[4] 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.

[4]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

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.

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

1.5.1 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 2024; 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 (2024, "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 feature set representing the 51 psychological constructs used in BUs analyses.[5] Notably, the set also includes features that Banerjee and Urminsky (2024) have shown affect click-through rates in this dataset, features such as *reading ease*, *numeric reference*, and the use of *visual language*.[6]

1.5.2 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.[7]

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

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

[6] 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.

[7] 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, ($t(300) = .93$, $p = .36$)

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 response beyond that.[8] 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.

WEB APPENDIX 2: PROMPTING MATERIALS

We use large language models to generate hypotheses, write counterfactual messages, 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=d58d4e38d43948eb8d87d25a513300f0`.

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. To ensure systematic variation and robust hypothesis generation, we designed a comprehensive prompting strategy that varied three key elements: the assigned role for the language model, the required structure for the hypothesis output, and the instructional context provided. We also imposed quality requirements to ensure each resulting hypothesis satisfied our goals of clarity, generalizability, empirical plausibility, unidimensionality, and usability. The basic prompt format is shown Section 2.2.1.

This systematic approach generated 288 unique prompt combinations by combining 4 different instruction types × 9 assigned roles × 8 hypothesis format templates. Each of the 2,100 headline pairs was randomly assigned to one of these prompt combinations and one of

[8] 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.

five model temperatures (.4, .6, .8, 1.0, 1.2) to minimize the chance our results were due to any specific prompt configuration.[9]

The three systematic variations worked as follows. First, all prompts began with assigning a *role*: "Assume you are {`role`}..." where {`role`} was replaced with one of nine assigned roles, such as "an editor of a top marketing journal such as the Journal of Consumer Research" or "a communication scientist researching the effects of linguistic framing on reader perception." Second, all prompts included a specific *hypothesis format* that constrained the response structure. 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 templates that started with "Hypothesis:" and ended with a reference to the feature's effect on engagement. To illustrate, one format reads: "Hypothesis: ____ leads to {`direction`} engagement with a message," where {`direction`} was filled with "more [less]" or "increases [decreases]" depending on whether $\widehat{m}_{a,b}$ was positive or negative. Third, prompts varied in the *instructions* provided to GPT. Three of the four instruction templates included the specific pair of headlines that were meant to generate a hypothesis, while one template included no headlines and served as a *Control* to assess whether hypotheses generated with access to our dataset differed from those generated without specific examples.[10] Other instruction variations included providing examples of ideal hypotheses or listing known constructs from Banerjee and Urminsky (2024) and asking GPT to "look for patterns not yet known." Below, we include examples 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.

[9] "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.

[10] Since we planned to exclude Control prompts from the rest of the pipeline, these instructions were undersampled before being matched to pairs.

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.[11]
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 assigned roles, one of the 8 hypothesis format templates, and one of the 4 instruction types (the three listed, plus the possibility of no variation), resulting in 288 total prompt combinations.

2.1.1 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}
```

[11] 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.

```
Please make sure that the hypothesis is:
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 Counterfactual Headlines

Our prompt for generating counterfactual headlines 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.

Each headline-hypothesis pair was randomly assigned to one of several model temperatures (.75 or .9) to encourage response variation. Unlike the prompts used to generate hypotheses, we did not vary the role; instead, all 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 to allow for "few-shot learning" of the headline distribution. In addition to this base prompt, 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 counterfactual should be as similar to the original headline in nearly every way except for the feature being changed.

2.2.1 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.

2.3.1 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 3: FORMALIZING THE STEPS OF HYPOTHESES GENERATION AND REFINEMENT FRAMEWORK

Step 1: Hypothesizing

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: Intervening

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.

Step 3: Predicting

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:

$$PTE(h) = \mathbb{E}\left[m(x, x') \mid (x, h, x') \in \mathcal{M}\right], \tag{3}$$

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.

Additional Filtering

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 4: QUALITY CHECKS FOR GPT TASKS

Quality of hypotheses

4.1.1 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.[12] 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.

[12]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).

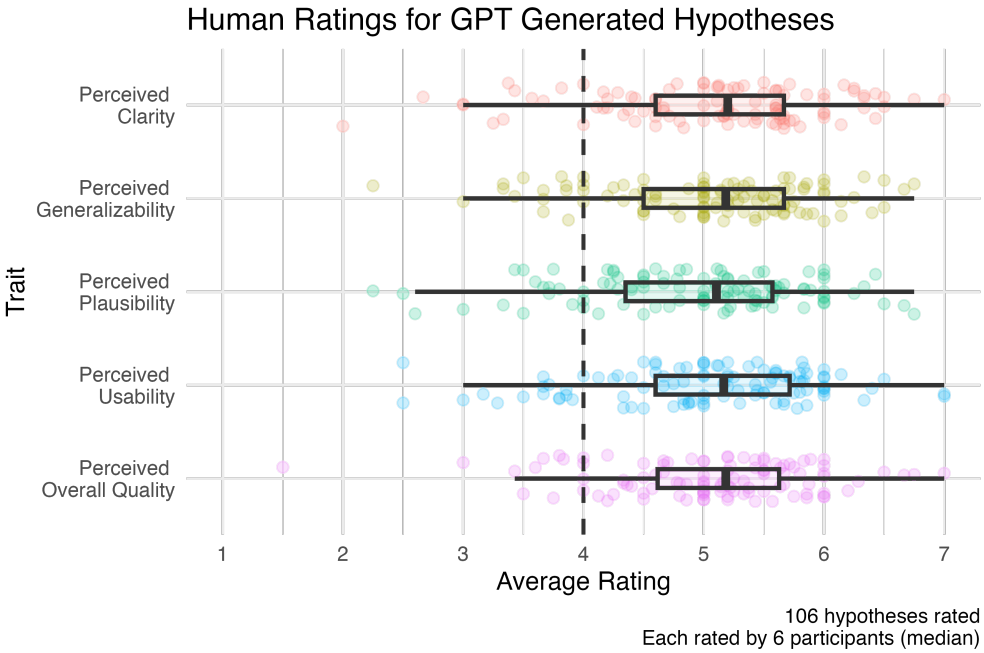

Figure 1: Human ratings for GPT-Generated Hypotheses.

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 3.

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 3: 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 counterfactual generation

Our first checks are aimed at confirming the quality of the GPT-generated counterfactual headlines.

4.2.1 Are counterfactuals 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.

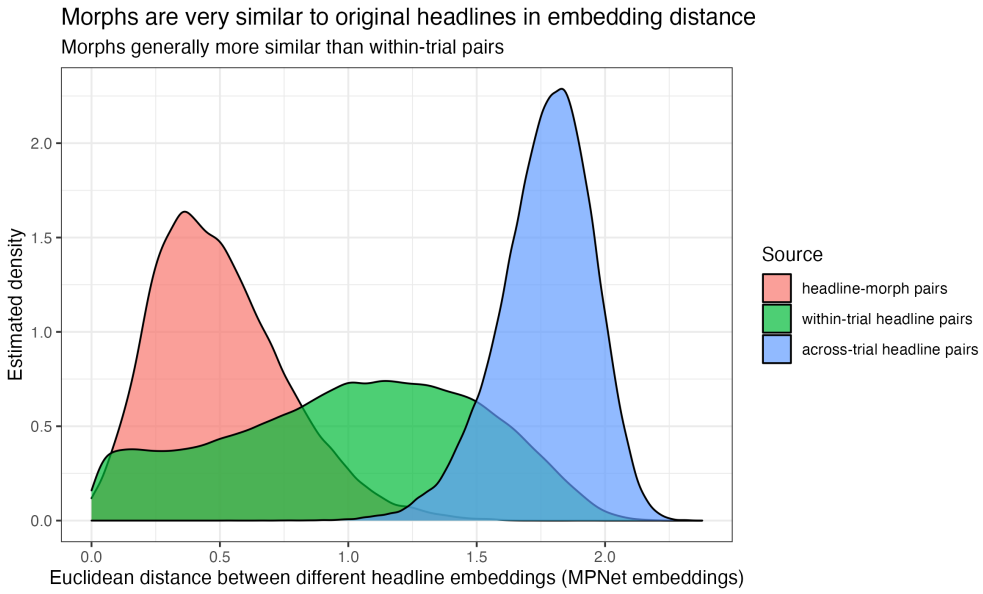

Figure 2: Average Euclidean distance between pairs.

4.2.2 Experiment 1: How do humans' attitudes towards GPT-generated 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 4). 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 4: 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.

4.2.3 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.

4.2.4 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 $t(994) = 3.37$, $p < .001$.

4.2.5 Do counterfactual headlines 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 5. 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 5, 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 5: 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 6. 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 6: 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 5: 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.[13] 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 7. 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.

[13] 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 7: Pairwise distances between embedding vectors for hypothesis, generated from various strategies.

WEB APPENDIX 6: 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 4.2.5, 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 4.3 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 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.

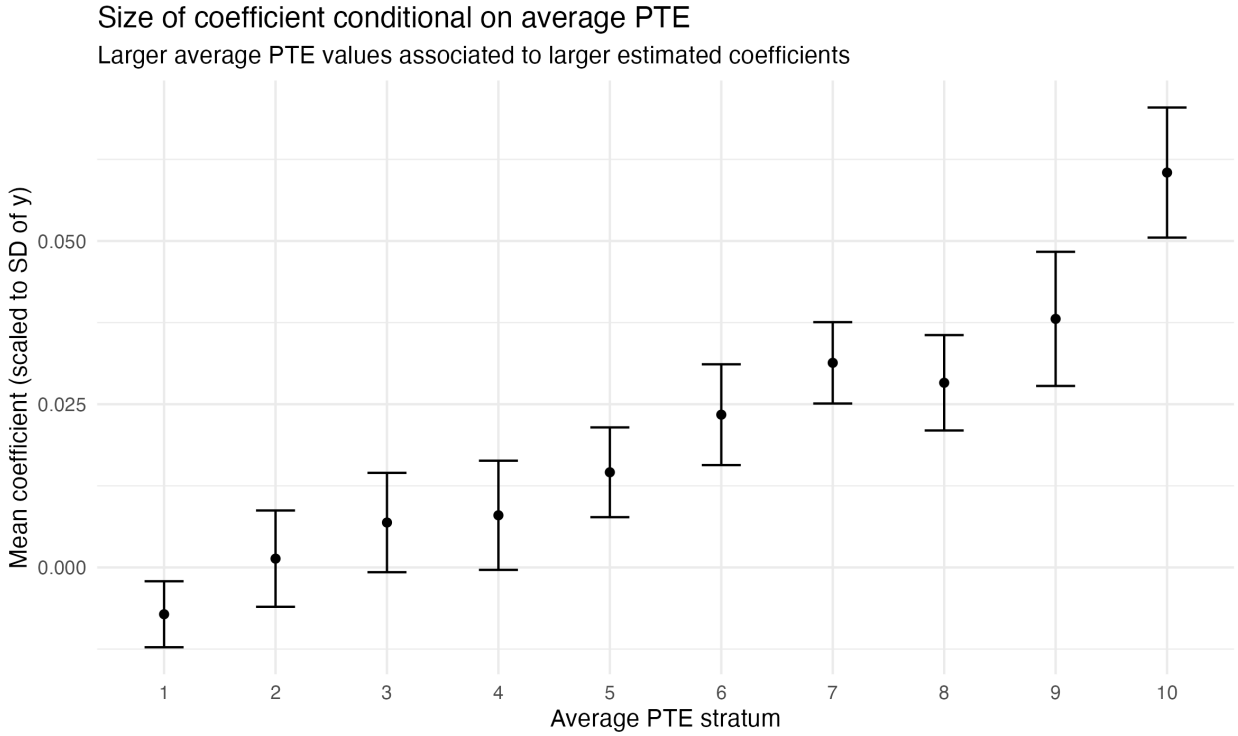

Figure 3: Coefficients tend to increase as a function of average PTE. Error bars show 2 standard errors of the coefficient values within each stratum.

WEB APPENDIX 7: GENERALIZING HYPOTHESES TO NEW CONTEXTS

Social Media Partner Data

7.1.1 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

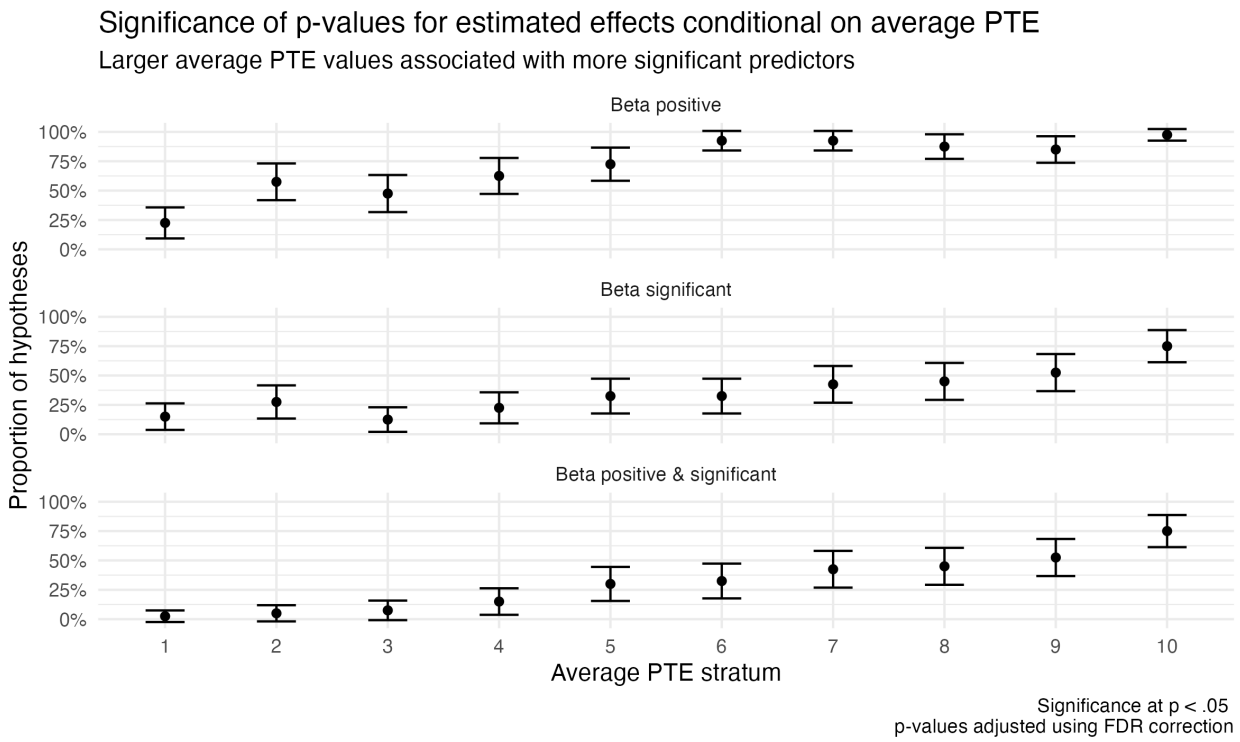

Figure 4: The significance of predictors increases as a function of average PTE. Error bars show 2 standard errors of the shown proportion within each stratum.

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 8.

Table 8: 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.

7.1.2 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*.

7.1.3 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 9. 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 9: 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.

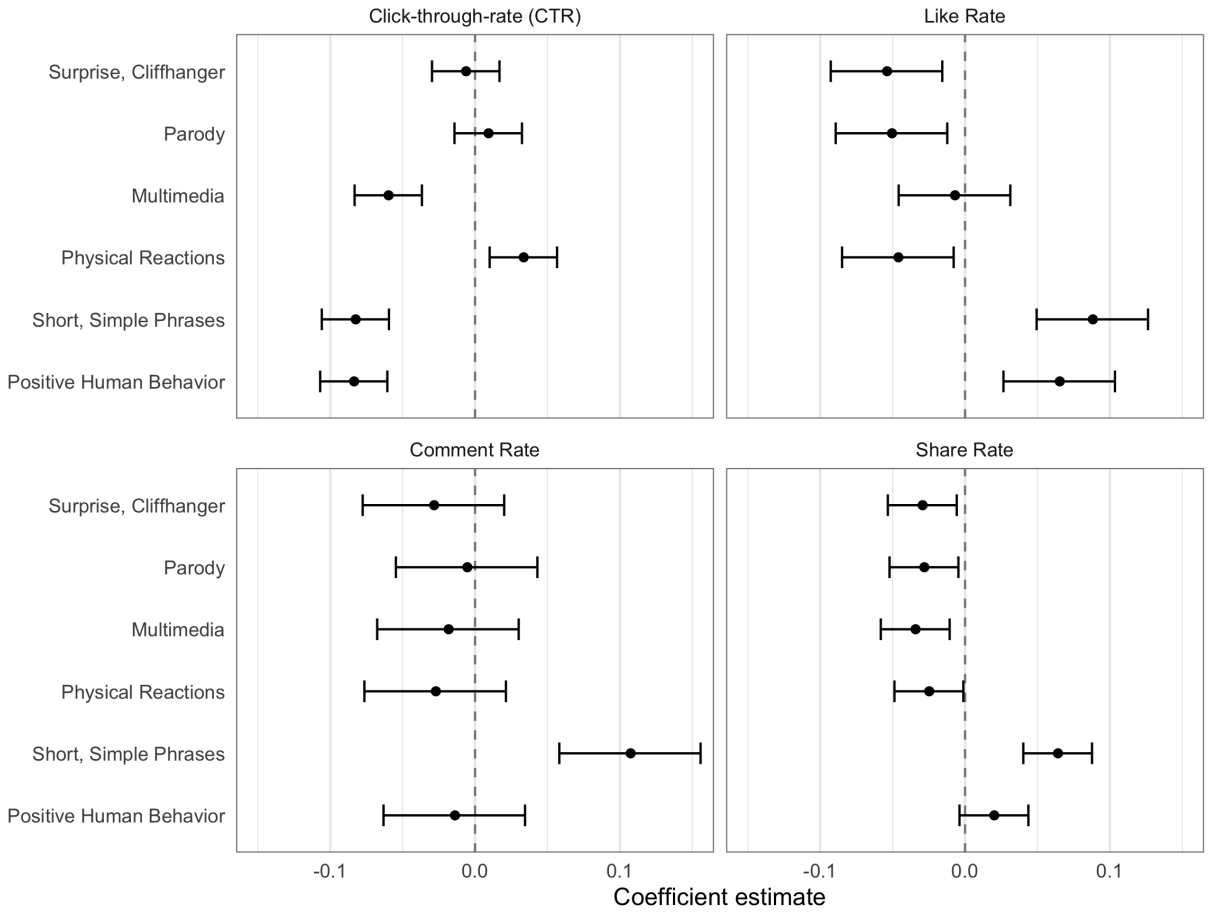

Figure 5: The coefficients for a range of outcomes and hypothesized features. Shown is the coefficient estimates with 95% confidence intervals for each outcome and feature combination.

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.

7.2.1 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 10.

Table 10: 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.

7.2.2 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*.[14] 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*.

7.2.3 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 11. 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 ($p < 0.05$).

[14] 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 11: 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.

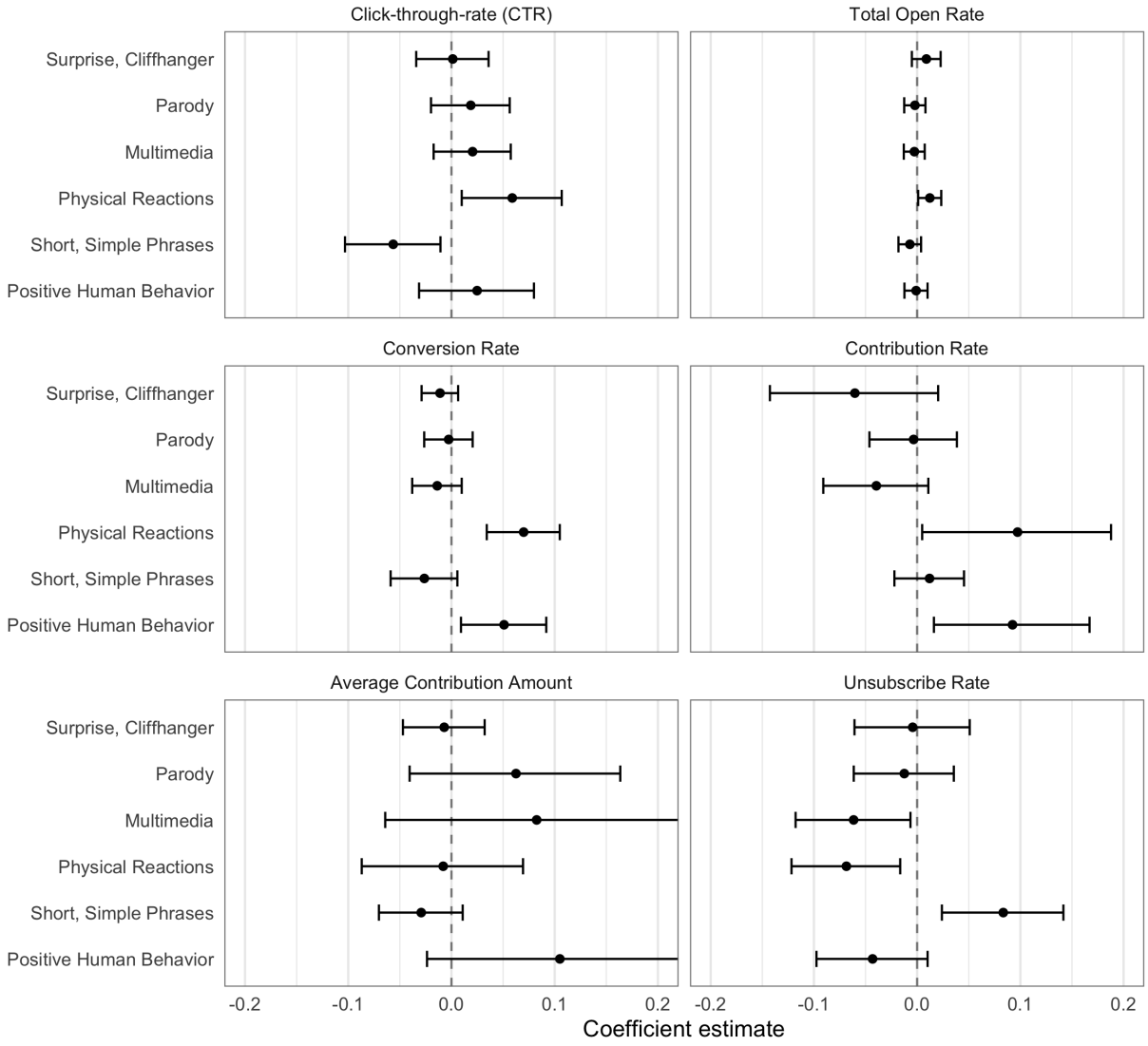

Figure 6: Most features have a non-significant relationship to most outcomes of interest. Shown is the coefficient estimates with 95% confidence intervals for each outcome and feature combination.

WEB APPENDIX 8: 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 12 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 9: ADDITIONAL FIGURES

Table 12: 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 13: 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

Banerjee, Akshina and Oleg Urminsky (2024), "The Language That Drives Engagement: A Systematic Large-scale Analysis of Headline Experiments," *Marketing Science* https://pubsonline.informs.org/doi/full/10.1287/mksc.2021.0018, publisher: INFORMS.

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.

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

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.

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/.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart (2022), "How to make causal inferences using texts," *Science Advances*, 8 (42) https://www.science.org/doi/full/10.1126/sciadv.abg2652.

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.

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.

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.

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.

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.

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.

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.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan "Large Language Models: An Applied Econometric Framework," (2025) https://www.nber.org/papers/w33344.

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.

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.

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.

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.

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.

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

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu "MPNet: Masked and Permuted Pre-training for Language Understanding," (2020) http://arxiv.org/abs/2004.09297.

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.

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.

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