# Words that Work: Using Language to Generate Hypotheses

Rafael M. Batista James Ross

Presenter: Lixuan Feng

March 28th, 2025

# Motivation

- The current approach to exploring the space of possible insights in a given application
  - Depend on specific settings, such as context, platform, and population
  - Relies on both human ingenuity and trial and error to come up with
  - Hard to make compelling predictions in a particular real-world setting
- Developing the data-driven process that generated hypotheses
  - Language encompasses many dimensions, Machine learning (ML) help uncover patterns that humans may miss
  - Generating novel and interpretable hypotheses from text using a combination of LLMs, ML, and psychology experiments

# Research Queation

- How to developing a data-driven process that generated hypotheses
- Steps
  - Generating Hypotheses
  - Ranking Hypotheses
  - Filtering Hypotheses

# Contribution

1. Research on data-driven discovery and hypothesis generation
   - hypos are interpretable, novel, testable, and generalizable
   - maintain a transparent role for both human researchers and algorithmic processes
   - offers a practical tool to researchers, organizations, and policymakers
2. Generate new insights into what drives consumer behavior
   - Prior: exploring how researchers can use text to study consumer behavior
   - Extend: introduce a framework to convert unstructured text into marketing insights
3. Literature on studying how language affects engagement
   - Extend: uncover new insights, some adding to existing theories and others inspiring new questions.
4. Literature on organizational learning
   - Prior: continuously run A/B tests to learn how various messages affect consumers' behavior
   - Extend: demonstrates how to aggregate insights from thousands of A/B tests in the form of specific hypotheses

# Data-Driven Hypothesis Discovery

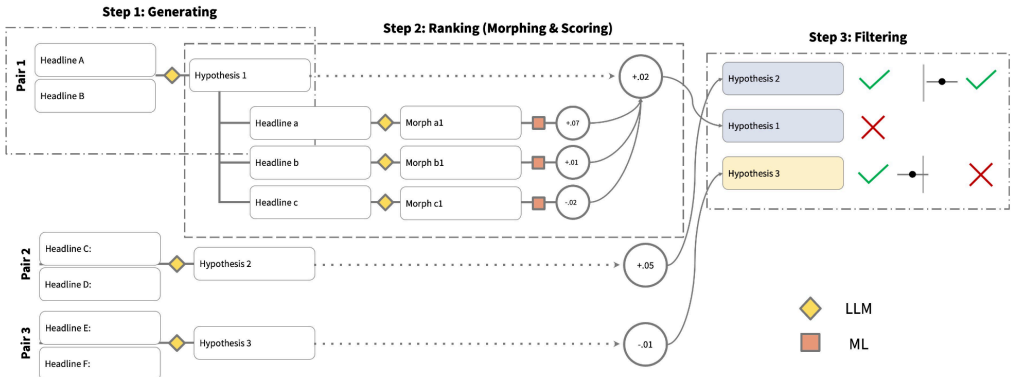


**Figure 1:** Overview of steps for generating and selecting hypotheses

# Step 1: Generate Hypotheses

- Provided OpenAI's GPT with 2,100 unique pairs of headlines written for the same story to produce 2,100 hypotheses.
- Example
  - Headline A: I Thought Long And Hard About Sharing This But I Decided I Had To Because These Dogs Need Our Help
  - Headline B: Don't Click This If You're Looking For Something That'll Make You Feel Better About The Human Race
  - GPT responded with, "Hypothesis: Incorporating reverse psychology leads to more engagement with a message."
- Assess quality: 79 human participants rate a subset of hypotheses
  - Most of the hypotheses were perceived to be clear, usable, and generalizable to new contexts

# Step 2: Rank the Hypotheses

- Rank-order hypotheses that insights most likely to affect the outcome (CTR)
- Part1: Morphing
- Applying each hypothesis to several different headlines, using GPT, producing a total of 250,000 morph-original pairs.
- Example
  - Hypothesis: Incorporating reverse psychology leads to more engagement with a message.
  - Headline A: Folks Who Work In Tipped Jobs Would Like You To Spend A Minute Looking At Something(Upworthy)
  - GPT produces "morphed"headline: You Probably Shouldn't Read This if You Think Tipping Is Optional

# Step 2: Rank the Hypotheses

- Part2: Scoring
  - how each morph might perform against the original headline
  - Using the ML algorithm to predict what the CTR would be for the morphed headline relative to the original headline
- Scoring produces an average PTE(predicted treatment effects) for each hypothesis
  - Primarily interested in the average effect per hypothesis

# Step 3: Filter the Hypotheses

- Clustered Selection
  - group similar hypotheses together using a sequential selection strategy
  - starting at the top, and select unique hypotheses
  - come across a hypothesis similar to one already selected, and exclude it from the list
  - reduced hypothesis set from 2,092 to 205
- Significance Testing
  - test whether PTEs were significantly greater than zero
  - Sixteen hypotheses had average PTEs that were positive and significantly greater than zero
- These steps constitute a data-driven framework for generating hypotheses before any confirmatory tests are conducted.

# Hypothesis Testing Using Hold - Out Set

- Regression set(six hypotheses)
  - participants were asked to "select the level which each trait is featured in this headline, from '1 (Low)'to '7 (High)'
  - $\Delta Ratings$: the relative intensity of the feature between two headlines

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

| | *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) |

# Application to online news headline

- Value of Click-Throughs
  - click-through rates (CTR), broad relevance for the consumption of news, but also for other domains such as advertising, influencer marketing

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

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

  - $\overline{CTR}$: the mean CTR calculated across all headlines
  - Impressions: Number of times the headline is displayed
- Upworthy Research Archive
  - dataset of 32,487 randomized trials (A/B tests) between 2013 and 2015

# Application to online news headline

- Absent in the Upworthy data are the explicit hypotheses each trial intended to test
- Existing Psycholinguistic Features
  - 51 psychological constructs used in BUs analyses
- Human Labels
  - capture any remaining information by collecting human guesses
  - Incentivized participants to choose from a pair of headlines —which one they believed had performed better in an A/B test
  - Evaluating the extent to which a headline exhibits features

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

  - $\Delta$Rating : the difference in construct values between the headlines in each pair

# Application to online news headline

- Predicting click-through rates
  - Adj R2: the proportion of the variation in CTR explained by the predictors.
  - Binary Accuracy: how well do humans pick the winning headline

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

- humans fail to see and that past research in marketing and psychology may not yet have discovered

# Ideas

- 多模态内容分析
  - 除了文本，还可以考虑多媒体内容（如图片、视频）对点击率的影响
  - 除了点击率衡量用户参与度，还可以用视频完播率，互动情况等数据
- 个性化语言策略
  - 结合用户画像数据，研究不同用户群体对不同语言策略的反应