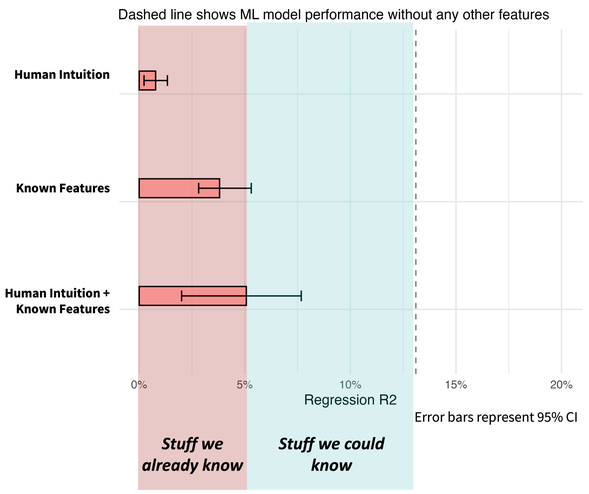
Words that work: Using large language models to generate and refine hypotheses

Rafael M. Batista & James Ross

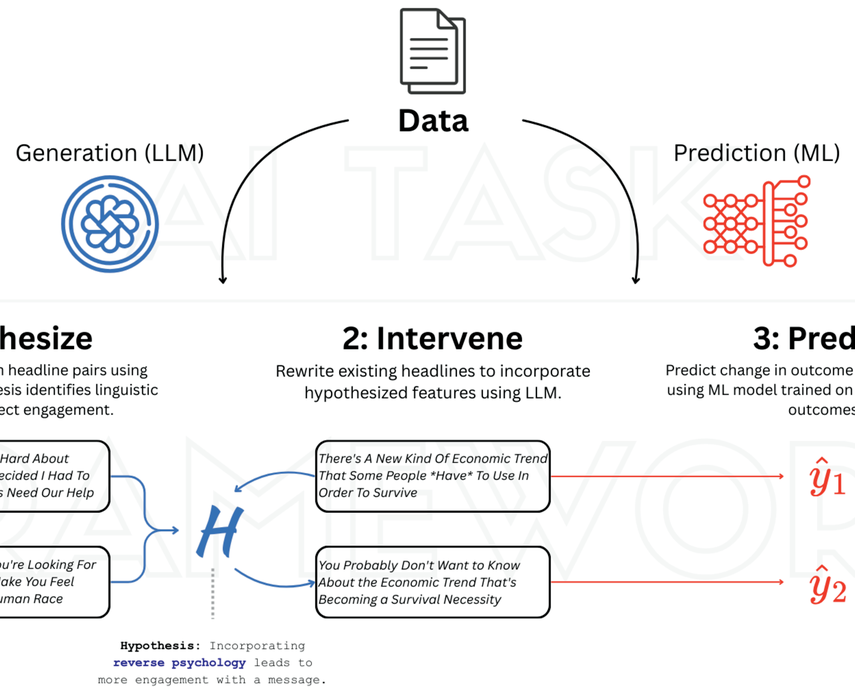
Summary
The ability to generate novel, testable hypotheses is fundamental to scientific discovery, yet this process remains largely dependent on human intuition and serendipity. We present a systematic framework for generating interpretable hypotheses from unstructured text data. Using a dataset of news headlines as a proof-of-concept, we uncover known and previously unidentified psychological insights into what drives online engagement. When tested out-of-sample, several hypotheses reveal significant effects

Is there anything to discover?
Yes! Preliminary analysis compares **model of known features** (based on review by Banerjee & Urminsky, 2024) to **black-box ML algorithm** (trained only on sentence embeddings) and finds the latter significantly outperforms the former in explaining variance of outcomes.



Related Works:
- Ludwig and Mullainathan (2024)
- Movva et al. (2025)
- Zhou et al. (2024)
- Banerjee and Urminsky (2024)
- Matias, Munger, Le Quere, and Ebersole (2021)

We introduce a framework that combines LLMs, ML, and human evaluation to generate and refine a set of interpretable hypotheses from unstructured text data.



Pre-Requisites
- **Data**: A dataset consisting of pairs of text, along with a measurable outcome of interest for each pair
- **ML algorithm**: An ML algorithm designed to predict the outcome of interest from each pair of text inputs
- **LLM**: A generative language model, such as GPT

Application: What Drives Engagement with News Headlines?

We apply this pipeline to the Upworthy Research Archive (Matias et al., 2021): a dataset of 32,487 randomized field experiments, measuring click-through rates ("CTR") for variations of news headlines written for the same story.

1

We used GPT-4 to generate 2,100 hypotheses from 2,100 unique pairs of headlines

2

We rewrote ~70 headlines according to a given hypothesis, producing ~252k new headlines, each matched to an original Upworthy headline.

3

We used the ML model, trained on embeddings of original headlines, to *predict* the treatment effect for each pair and averaged predicted effects by hypothesis.

To refine the set, we ranked hypotheses by predicted treatment effects, clustered hypotheses that produced similar semantic changes when applied to a message, and removed any remaining hypotheses with weak predicted effects.

Hypotheses Generated

1. *Framing a message with an **element of surprise followed by a cliffhanger** makes people more likely to engage with a message.*
2. *Incorporating the **concept of parody** makes people more likely to engage with a message.*
3. *Incoporating **multimedia evidence** in a headline results in more engagement with a message.*
4. *Describing **physical reactions** makes a message more engaging.*
5. ***Shortening and simplifying phrases** affects engagement with a message [negatively].*
6. *Focusing on **positive aspects of human behavior** affects engagement with a message [negatively].*

Testing Hypotheses Out-of-Sample
To validate our framework, we collected headlines from 1,693 experiments from a holdout set, and recruited 800 participants to rate each headline on each of the six hypothesized features. We used these labels in a regression of the form:

$$\Delta CTR = \beta_0 + \beta_r \cdot \Delta Rating + \varepsilon,$$

for each pair of headlines from the same trial. These tests were pre-registered on AsPredicted.org.

Four of the six hypotheses showed significant effects in the predicted direction. Notably, several of these effects persisted even after controlling for previously known linguistic features.

	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