DATASCI 241 – Experiments and Causal Inference
Team Members: Kevin Coppa, Agostina Galluzzo, Armaan Hiranandani, Hyeong Geon Kim, Michael Mandujano

LLM vs. Human-Written Ad Copy for Survey Recruitment

Introduction

Recruiting participants for online psychology surveys is often time-intensive and costly. As large language models (LLMs) such as GPT-4o continue to improve in their ability to generate persuasive content, a natural question emerges: can LLMs outperform human-written ad copy in attracting and retaining survey respondents? Our research investigates this by asking:

Does GPT-4o-generated ad copy lead to higher engagement and recruitment for a short psychology survey compared to human-written copy?

Justification & Theoretical Motivation

In our research phase the team noticed that previous studies in marketing have proven that AI-generated content can in-fact outperform human-created materials in engagement metrics, like click-through rate (CTR). In 2019, JPMorgan Chase found that AI-generated advertising copy *significantly* improved CTRs over copy written by human agents. This pattern aligns with the theoretical argument that LLMs, trained on a massive collection of persuasive text, may have an advantage in capturing user attention. Our experiment seeks to test this proposition in the domain of participant recruitment for academic surveys.

Hypothese

Ads written by GPT-4 will result in a higher click-through rate (CTR) than those written by humans. Respondents recruited via GPT-4o-written ad copy will exhibit higher survey start and completion rates.

Experimental Design

Potential Outcomes Framework

Let $Y_1(1)$ denote the outcome (e.g., CTR or completion) for user i exposed to the GPT-4o ad, and $Y_1(0)$ denote the outcome under exposure to the human ad. Our estimand is the average treatment effect (ATE):

$\tau = E[Y_i(1) - Y_i(0)]$

Randomization

Reddit Ads Manager was used to randomly assign users to one of several ad treatments:

- Control A: Human-written ad (Neutral Tone)
- Control B: Human-written ad (Emotional Tone)
- Control C: Human-written ad (Professional Tone)
- Treatment A: GPT-4o (Neutral tone)
- Treatment B: GPT-4o (Emotional tone)
- Treatment C: GPT-4o (Professional tone)

Reddit's native ad delivery mechanism randomly delivers ads to users within selected subreddits (e.g., r/Psychology, r/SampleSize).

Treatment Description

Each ad was a text-based recruitment message for a "short 5-minute psychology survey." The control was written by our team. The GPT-4o treatments were generated by prompting the model with the same content but instructing it to rephrase in a neutral, emotional, or professional tone.

All GPT-4o ads linked to the same AI-reworded survey; the human ad linked to a human-written survey with semantically identical questions.

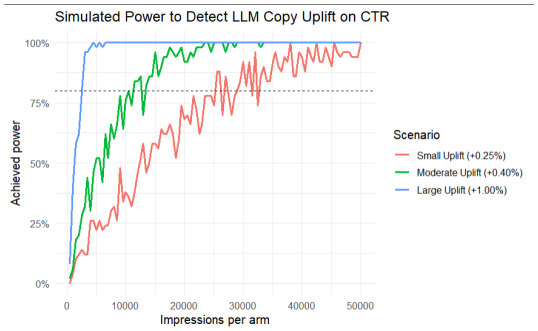
Participant Flow (CONSORT-style summary)

- **Eligible:** Reddit users browsing relevant subreddits
- **Assigned:** Automatically assigned to ad variant by Reddit
- **Clicked:** Users who clicked on the ad
- **Started Survey:** Users who began the Qualtrics survey
- **Completed Survey:** Users who completed all survey questions

Power Calculation

Using 50 simulation replicates per scenario, we estimated the statistical power to detect treatment effects based on pre-experiment assumptions. Assuming a baseline click-through rate (CTR) of 1.00%, we simulated three uplift scenarios: 25%, 40%, and 100%. We found that achieving 80% power would require approximately 25,500 impressions per arm for the smallest uplift, 10,500 for moderate uplift, and 2,500 for a large uplift. These results informed our minimum sample size requirements to reliably detect expected differences in CTR in our A/B test.

DATASCI 241 – Experiments and Causal Inference
Team Members: Kevin Coppa, Agostina Galluzzo, Armaan Hiranandani, Hyeong Geon Kim, Michael Mandujano



Minimum Sample Size Required per Arm for 80% Power

Scenario	Min Sample Size per Arm for 80% Power
Small Uplift (+25%)	25500
Moderate Uplift (+40%)	10500
Large Uplift (+100%)	2500

Data & Analysis Plan

CTR Analysis

The team used a few different methods when trying to evaluate whether ad copy written by ChatGPT had a measurable effect on user behavior compared to ads written by us. The main outcome we focused on was click-through rate (CTR), which we calculated by dividing the number of clicks by the number of times the ad was shown.

In our models, we included the type of ad someone saw (AI vs. Human), the tone of the ad (Neutral, Emotional, or Professional), and the interaction between these two factors. These variables came directly from how we designed the experiment.

The analysis happened in two main parts. First, we ran basic proportion tests comparing CTRs between AI and Human ads within each tone category. The biggest difference showed up in the Neutral tone, where the AI ad got a 2.01% CTR compared to 1.50% for the human version. This result was statistically significant ($p = 0.000098$), with a confidence interval ranging from 0.22% to 0.81%. On the other hand, in both the Emotional and Professional tones, the CTR differences between AI and Human ads weren't significant ($p = 0.178$ and 0.766, respectively).

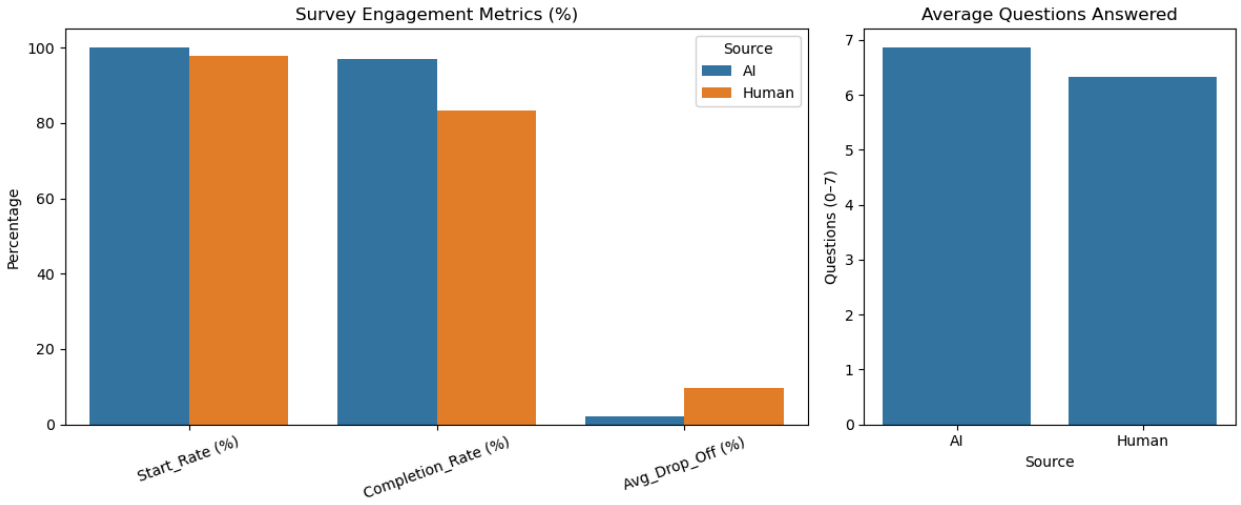
Second, we ran a logistic regression with CTR as the outcome variable. This model included whether the ad was written by AI or a human, the tone of the ad, and their interactions. We used Human Neutral as the baseline group. The results showed that the Neutral AI ad clearly performed better than the baseline ($p < 0.001$), but that the Emotional and Professional AI ads actually did worse than their human counterparts. These interaction effects were significant ($p < 0.001$ for Emotional and $p = 0.003$ for Professional), which shows that it's not just the AI label that matters but also it's how that AI copy is written. Overall, this part of the analysis helped reveal patterns that weren't obvious just by looking at the raw CTRs.

Survey Completion & Drop-Off Analysis

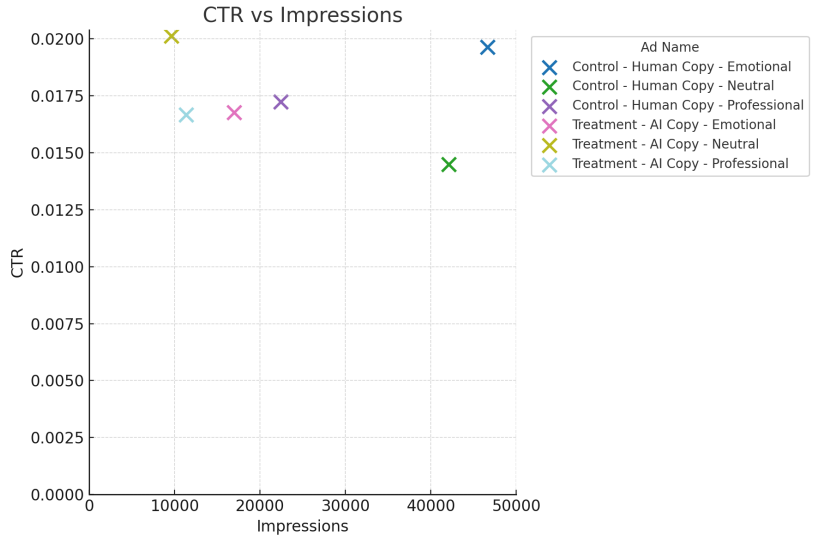
While click-through rate (CTR) reflects ad effectiveness at attracting potential participants, we also examined post-click engagement, focusing on (1) completion rate, which we defined as the proportion of survey starters who finished the survey, and (2) drop-off fraction, the portion of the survey left incomplete, with 0 indicating full completion and 1 indicating no questions answered. These were calculated based on participant's answers, or lack-there-of, for questions 1-7 of the survey.

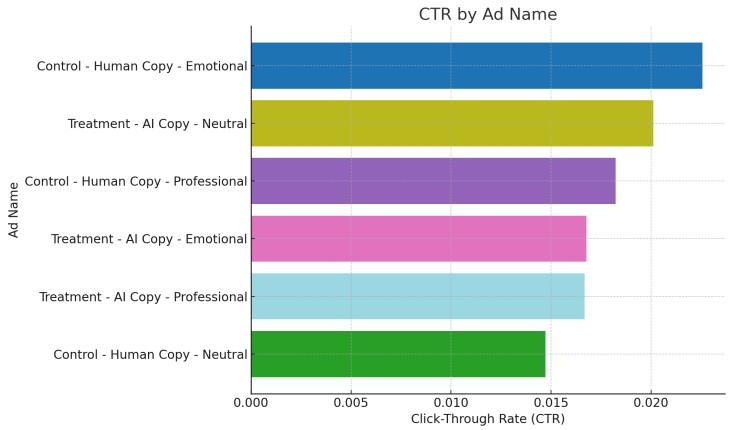
For this purpose, due to limited data, we aggregate our survey results into Human ad-copy and AI ad-copy, no longer accounting for tone used. Across all tones combined, completion rates were significantly higher for participants recruited via GPT-4o ad-copy (97.3%) compared to Human ad-copy (83.4%) (chi-square $p = 0.0012$).

Similarly, drop-off fractions were lower in the AI group (2.04%) compared to the Human group (9.73%). Due to skewness of our data we used a Mann–Whitney U test confirming the difference ($p = 0.0006$). This pattern suggests that GPT-4o ads did not just draw more clicks in certain tones but they also tended to recruit participants who were more likely to complete the survey once started.



DATASCI 241 – Experiments and Causal Inference
Team Members: Kevin Coppa, Agostina Galluzzo, Armaan Hiranandani, Hyeong Geon Kim, Michael Mandujano





Limitations

According to our power analysis, the number of impressions for AI-generated ads fell below the minimum required to confidently detect at least a 25% uplift in click-through rate.

Reddit's ad delivery algorithm dynamically optimizes for performance at the campaign level, allocating more impressions to variants that appear to perform better or cost less to serve. As a result, equal targeting and budget settings did not guarantee nor result in equal impressions across ad arms, which may have introduced imbalance in exposure unrelated to our experimental design.

While ad performance data (CTR, impressions, reach) was segmented by sentiment and tone, the survey dataset did not record which sentiment version each respondent saw, preventing tone-level analysis of engagement and completion. Demographic data was collected only at the

end of the survey, meaning analyses involving demographics included only participants who completed the survey, potentially biasing results toward more engaged respondents. Finally, because AI- and Human-generated ads linked to slightly different surveys, observed differences in post-click engagement (such as completion rates and drop-off patterns) may reflect differences in survey wording or flow as well as ad copy effectiveness.

Conclusion

Our experiment examines whether GPT-4o-generated ad copy can improve participant recruitment in online research. Results show that in the Neutral tone, GPT-4 ad copy significantly outperformed human-written ads in CTR. However, in Emotional and Professional tones, human copy performed better, with statistically significant interactions revealed through logistic regression.

These findings suggest that while LLMs are powerful tools, their performance is highly sensitive to context and tone. Future experiments should isolate survey content from ad copy more carefully to better attribute effects.