Gender Differences in Responses to LLM-Generated versus Human-Written Donation Appeals

A Multi-factorial Analysis of Digital Charitable Giving

Bachelor End Project

Salah-din Mrait
Student Number: 1744976

Data Science Joint Degree
Tilburg University – Eindhoven University of Technology

Supervisor:

John Caffier

January 2026

Abstract

This study questions whether men and women respond differently to LLM-generated versus human-written donation appeals for cancer charities. Using experimental data from 725 U.S. participants who each evaluated six charitable appeals (4,350 total observations), I test whether the well-documented gender gap in charitable giving remains equally for both human and LLM-generated content. I measure three outcomes: engagement ratings, persuasiveness scores, and actual donation behavior. Mixed-effects regression models account for the nested data structure and test for gender $\times$ content source interactions. This research helps explain how a person's demographic traits change their response to LLM-generated content in social-good settings.

Keywords: charitable giving, donation behavior, gender differences, LLM-generated content, LLM, persuasion, prosocial behavior

Preface

This Bachelor thesis represents the culmination of my studies for the Joint Bachelor of Science in Data Science at Tilburg University and Eindhoven University of Technology.

I would like to thank my supervisor, John Caffier, for his help and advice throughout this project. I would also like to thank my co-students and group members, Joery Fabian Clements and Mohamad al Hallak, for their valuable collaboration and ideas during our time working together.

Salah-din Mrait
Tilburg / Eindhoven, January 2026

Contents

Chapter 1

Introduction

Charitable organizations face growing competition for donor attention and must continually improve how they ask for support. Digital fundraising has lowered the cost of reaching potential donors, yet success still depends on whether an appeal can trigger a sincere emotional response. A large body of empirical research shows that gender is one of the most consistent predictors of prosocial behavior. Across cultures and research settings, women donate more frequently and give higher amounts than men (Bekkers and Wiepking, 2011; Mesch et al., 2011). This difference is not simply about generosity. It shows deeper differences in socialization, moral identity, and how emotional messages are processed (Winterich et al., 2009; Diekman and Clark, 2015).

The rise of Large Language Models has added a new element to charitable communication. Organizations can now produce emotionally framed donation appeals without direct human involvement. These systems can imitate empathic language, yet they do not possess intention or emotional experience. This distinction matters in donation contexts, where perceived sincerity plays a key role. Research on algorithm aversion shows that people judge algorithmic output more harshly in domains that rely on judgment, intuition, or emotional meaning (Dietvorst et al., 2015; Castelo et al., 2019).

Women tend to respond more strongly to charitable appeals and place greater value on relational cues and emotional authenticity than men (Croson and Gneezy, 2009). If LLM-generated messages are perceived as not having a genuine moral agency, the group that contributes most to charitable giving may respond less positively to such appeals. This study questions whether revealing LLM authorship alters donation behavior differently for women and men.

1.0.1 Theoretical background and mechanisms

Understanding these results requires a closer look at the psychological differences in how men and women handle being persuaded.

Empathic concern and emotional processing Gender differences in charitable giving are often linked to differences in empathic concern. Women consistently show higher levels of empathy (Davis, 1983). Fundraising messages often use personal stories that invite readers to be emotionally engaged. When emotional engagement drives persuasion, perceived sincerity becomes important. Moral judgment is dependent on the perception of an intentional mind behind an action (Gray et al., 2012). If LLM are perceived as not having such a mind, emotional engagement may weaken.

Trust and algorithm aversion Algorithm aversion describes the tendency to lose confidence in algorithmic judgments after observing errors (Dietvorst et al., 2015). Later work shows that this reaction depends on the type of task. Algorithms tend to receive more trust in technical or rule-based settings than in tasks involving values, emotions, or moral persuasion. People often assume machines dont have any lived experience, which lowers trust in subjective areas (Castelo et al., 2019). Since women place greater weight on trust and relational signals in communication (Croson and Gneezy, 2009), they may react more negatively to emotional appeals once LLM authorship is revealed.

Moral reasoning: care vs. justice Differences in moral reasoning give another explanation for gendered responses. Women are more likely to adopt a care-oriented moral perspective focused on relationships and reducing suffering, whereas men lean more toward rule-based reasoning centered on rights and fairness (Gilligan, 1982). Moral Foundations Theory supports this view, showing that women score higher on the Harm/Care foundation (Graham et al., 2011). Emotional appeals written by LLM may succeed only if the sense of a caring relationship between donor and recipient remains intact. If artificial authorship weakens that perception, effectiveness may reduce for women in specific.

LLM systems can create emotionally engaging language, but they dont have moral agency. This creates tension between how effective it is linguistically and the perceived sincerity. This study tests how this tension affects women and men differently.

1.1 Research questions and hypotheses

This study addresses the following main research question:

> **Do women and men respond differently to LLM-generated versus human-written emotional donation appeals for cancer charities?**

Based on the theoretical interaction between gendered moral reasoning and algorithm aversion, I investigate three specific sub-questions:

1. **RQ1 - Perceived authenticity:** Is there a gender difference in perceived authenticity of LLM-generated emotional appeals compared to human-written appeals?

   H1: Women will report significantly lower persuasiveness scores (measured on a 1–7 Likert scale) for LLM-generated appeals than for human-written appeals, while men will show no such difference.

2. **RQ2 - Donation amounts:** Do women donate higher amounts than men regardless of content source, or does the gender gap in donation generosity differ between human-written and LLM-generated appeals?

   H2a: Women will allocate higher monetary amounts (measured in USD 0.00–0.10) to charities than men, regardless of whether the appeal is human-written or LLM-generated.

   H2b: The difference in mean donation amounts (USD) between men and women will be significantly larger for human-written appeals than for LLM-generated appeals.

3. **RQ3 - Persistence of gender gap:** Does the well-established gender gap in charitable giving remain equally for both human and LLM-generated appeals, or does content source moderate this relationship?

   H3: The likelihood of women selecting the "Like" category (ordinal rating: Dislike/Neutral/Like) will be significantly higher for human-written appeals compared to LLM-generated appeals.

1.2 Contribution and significance

This research connects work on gender differences in charitable giving (Bekkers and Wiepking, 2011) with research on human interaction with algorithmic systems. Prior studies on algorithm use focus on accuracy and advice-taking (Dietvorst et al., 2015; Logg et al., 2019). Few studies question how demographic factors shape responses to LLM in morally sensitive settings.

Nonprofit organizations increasingly rely on generative LLM for outreach. If such tools reduce trust or perceived sincerity among women, fundraising outcomes may suffer. By looking at gender and authorship, this research helps fundraisers decide whether to use LLM'S or a human writer.

1.3 Thesis structure

The remainder of this thesis is organized as follows. Chapter 2 describes the research methods, including the experimental design, variables, data preprocessing, and analytical approach. Chapter 3 presents the results in three phases: descriptive analysis, main effects models, and interaction analysis. Chapter 4 discusses the theoretical and practical implications of our findings.

Chapter 2

Methods

2.1 Data description and experimental design

The dataset comprises 4,350 observations from 725 unique U.S. participants, with each participant being tested with six donation appeals for cancer-related charities. The experimental design manipulated two primary factors: personalization type (counterfactual, generic, or personalized) and content source (human-written versus LLM-generated). Participants provided responses across three outcome measures: engagement ratings (Dislike/Neutral/Like), persuasiveness scores (averaged across three 7-point Likert items), and donation behavior (actual monetary allocations ranging from $0.00 to $0.10). The within-subjects design creates a data structure where each participant provided multiple responses, and each message was evaluated multiple times.

Participants saw social media posts that were presented as authentic fundraising appeals, but varied in authorship between a human writer and an LLM. The present analysis employs the message-level dataset (Data_LongFormat.csv), which contains 18 variables that represent both experimental manipulations and participant demographics, with each row corresponding to a single participant-message evaluation. This structure allows me to examine the within-subject response variation across experimental conditions while also comparing responses between male and female participants.

2.2 Variables

2.2.1 Outcome variables

Engagement is measured through a three-category ordinal rating (Dislike, Neutral, Like), that measures participants' immediate emotional reactions to donation appeals. This measure provides an idea into how the message is initially perceived before deeper cognitive processing.

Persuasiveness is measured using a composite score from three 7-point Likert scale items that look at different parts of perceived message effectiveness. Prior to analysis, internal consistency was verified using Cronbach's $\alpha = 0.89$, which showed reliability. The composite score is participants' evaluation of message quality and convincingness.

Donation behavior is operationalized as the monetary amount (in dollars) participants spent to each charity from their experimental budget. This measure gives an objective indicator of actual giving that completes self-reported subjective responses.

2.2.2 Independent variables

Gender serves as our main demographic variable and is coded as a binary variable (Male/Female) based on how participants identified themselves. I chose gender as the central variable because so much previous research shows clear, consistent differences in how men and women donate to charity.

Content source represents the experimental manipulation telling us whether appeals were written by humans or generated by LLMs. This variable allows for direct comparison of effectiveness between authorship types and is central to our research questions about whether established gender differences remain across both content sources.

2.3 Data preprocessing

Before analysis, several data quality checks were done. First, I verified that each row represents a unique participant-message combination by looking at the participant ID and post ID fields for duplicate entries. No duplicates were identified, confirming that the 4,350 observations are distinct evaluations. Second, I did a missing data assessment across all variables. The dataset contains no missing values, so there is no need for imputation procedures.

Categorical variables were recoded to follow a consistent naming format. Gender was coded with "Male" as the reference category to be able to directly test whether female participants show different response patterns. Content source was coded with "Human" as the reference category to test if LLM-generated content performs differently than human-written content. Numerical outcome variables were kept in their original continuous scales to preserve measurement precision.

A discrepancy was reveal between the initial project documentation (which referenced 750 participants) and the actual dataset. By looking at the amount of unique identifiers I noticed the correct number to be 725 participants. This has no effect on the validity of the analysis.

2.4 Analytical approach

All analyses were done using R (version 4.3.0) and RStudio. Key packages include **tidyverse** for data manipulation, **lme4** for mixed-effects models, **ordinal** for ordinal regression, **glmmTMB** for hurdle models, **ggplot2** for visualization, and **sjPlot** for model summaries.

Python (version 3.10) with **pandas**, **seaborn**, and **matplotlib** was used for supplementary visualizations.

2.4.1 Phase 1: Descriptive analysis

Initial descriptive statistics characterize the sample and looks at distributions of key variables. The distribution of male and female participants is reported and the appropriate sample sizes in both groups are verified. For outcome variables, I compute descriptive statistics (means, standard deviations, medians, ranges) overall and grouped by gender and content source.

I assess the distribution of persuasiveness scores using histograms and Q-Q plots to see whether they approximate normality, justifying the use of linear models. For donation amounts, I calculate the proportion of zero donations versus positive donations. I also look at correlations among the three outcome variables.

2.4.2 Phase 2: Main effects models

Separate models are estimated for each outcome variable. All models account for the nested data structure using mixed-effects specifications. The choice of model for each outcome variable is motivated by the distributional properties and measurement characteristics of the data.

Engagement Model (Ordinal Logistic Mixed-Effects Regression): Because engagement ratings have a natural ordering (Dislike < Neutral < Like) but dont have equal intervals between categories, I use cumulative link mixed models. A standard linear model would incorrectly assume equal distances between categories and ignore the ordinal structure. The ordinal model preserves the ranking of the data while calculating the likelihood of a response moving from one category to the next. I include random intercepts for participants and posts to account for the nested structure, where observations within the same participant or post are more similar than observations across participants or posts. The model formula is:

$$\text{logit}(P(Y \leq j)) = \theta_j - (\beta_1 \text{Gender} + \beta_2 \text{Source} + u_{\text{participant}} + u_{\text{post}}) \tag{2.1}$$

where j indexes response categories, θ_j are threshold parameters, and u terms are random intercepts.

Persuasiveness Model (Linear Mixed-Effects Regression): After confirming approximate normality of the persuasiveness composite through diagnostic plots, I use linear mixed-effects models. This approach is appropriate because the composite score is continuous and approximately normally distributed. The nested structure requires random effects to avoid underestimating standard errors and inflating Type I error rates. The model includes random intercepts for participants (accounting for individual differences in rating tendencies) and posts (accounting for systematic differences in message quality). The model formula is:

$$Y_{ij} = \beta_0 + \beta_1 \text{Gender}_i + \beta_2 \text{Source}_{ij} + u_{0i} + u_{0j} + \varepsilon_{ij} \tag{2.2}$$

where Y_{ij} is the persuasiveness score, u_{0i} and u_{0j} are random intercepts, and ε_{ij} is the residual error.

Donation Model (Hurdle Mixed-Effects Model): Donation data exhibit substantial zero-inflation (35.3% non-donors) and right-skewed distributions among positive donations. A standard linear or Poisson model would perform poorly because it cannot sufficiently model both the high amount of zeros and the continuous positive values. A hurdle model separates these two processes: first, a binary logistic component predicts whether any donation happens (crossing the "hurdle" from zero to positive); second, a truncated Gaussian component predicts donation amount conditional on donating. This two-part structure mirrors the theoretical distinction between the decision to give and the decision of how much to give. The conditional component includes random intercepts for participants and posts. For the zero-inflation component, to get the model to converge, I simplified the structure by removing the random effects.

2.4.3 Phase 3: Interaction analysis

The main question I am asking is whether gender influences how people react to the content source. I test this by adding a Gender $\times$ Content Source interaction term to each model. A statistically significant interaction ($\alpha = .05$) suggesting that the gender gap differs depending on content source.

For significant interactions, I do simple slopes analysis to decompose the interaction pattern, estimating the effect of content source separately for men and women. I create interaction plots visualizing predicted values across all four conditions with confidence intervals.

Chapter 3

Results

3.1 Phase 1: Descriptive analysis

This section presents the descriptive statistics and visualizations for the three primary outcome variables: Engagement, Persuasiveness, and Donation Behavior.

3.1.1 Descriptive statistics

Engagement

Figure 3.1 displays the distribution of engagement ratings across gender and content source. Table 3.1 summarizes the frequencies.

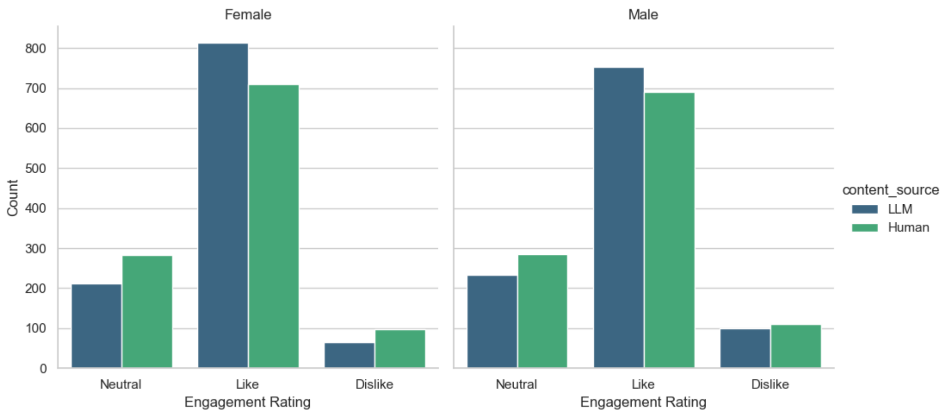


Figure 3.1: Distribution of Engagement Ratings by Gender and Content Source

Table 3.1: Engagement Ratings by Gender and Content Source

Gender	Source	Dislike	Neutral	Like
Female	Human	98	282	709
	LLM	65	211	813
Male	Human	110	286	690
	LLM	100	233	753
Overall		**373**	**1012**	**2965**

Persuasiveness

Figure 3.2 shows the distribution of persuasiveness scores. Table 3.2 presents the means and standard deviations.

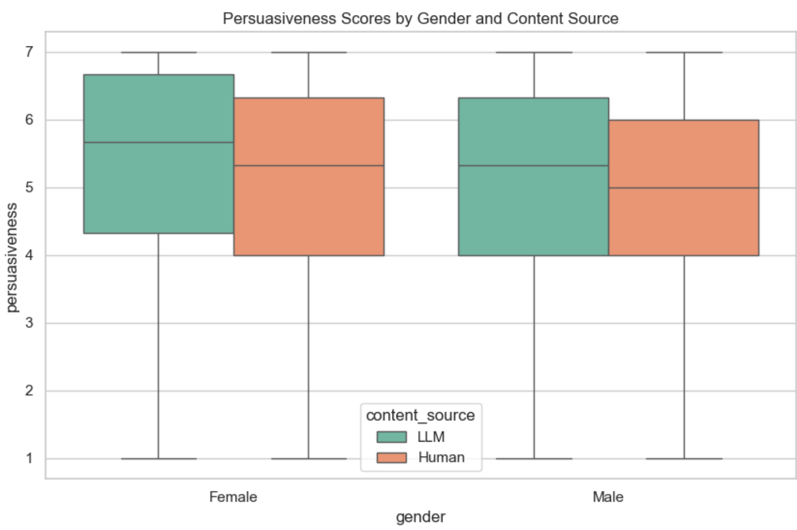


Figure 3.2: Persuasiveness Scores by Gender and Content Source

Table 3.2: Descriptive Statistics for Persuasiveness Scores (1-7 scale)

Gender	Source	Mean	SD
Female	Human	4.93	1.72
	LLM	5.18	1.64
Male	Human	4.87	1.66
	LLM	5.08	1.61
Overall		**5.02**	**1.66**

Donation behavior

Figure 3.3 displays the mean donation amounts. Table 3.3 summarizes the descriptive statistics.

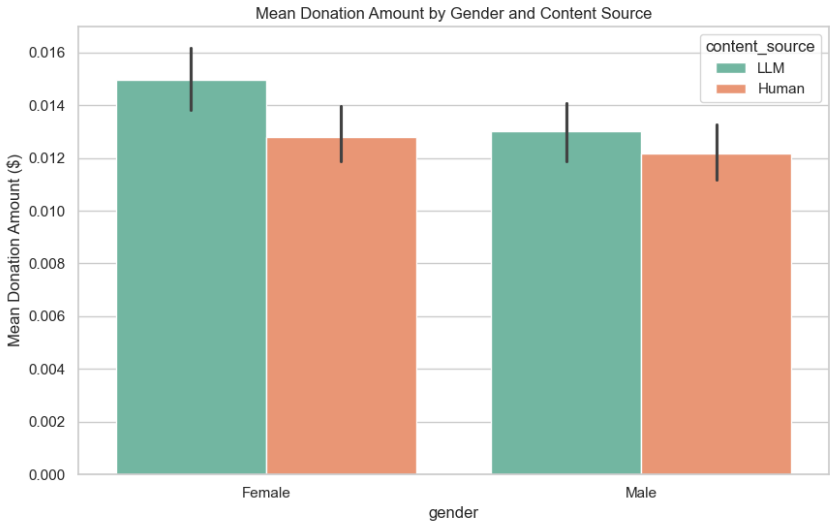


Figure 3.3: Mean Donation amount by Gender and Content Source

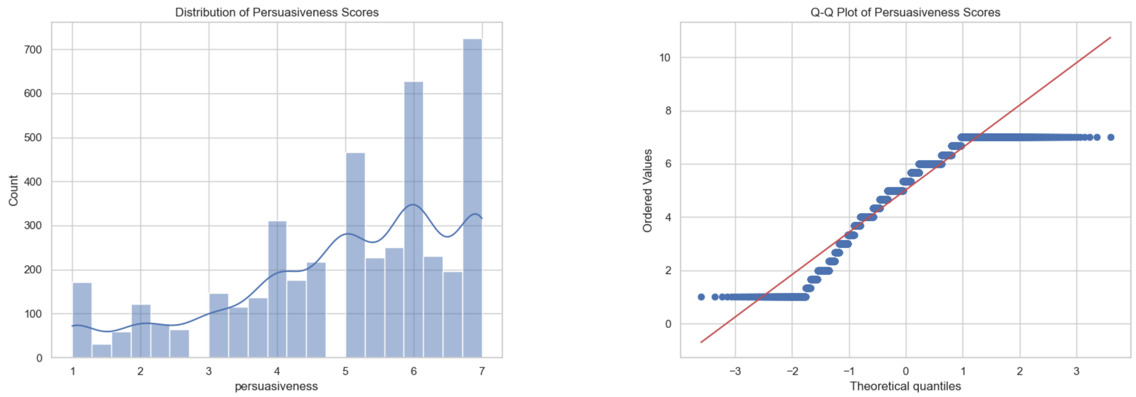
Table 3.3: Descriptive statistics for donation amount (in dollars)

Gender	Source	Mean	SD
Female	Human	0.0128	0.0181
	LLM	0.0150	0.0198
Male	Human	0.0122	0.0178
	LLM	0.0130	0.0181
Overall		**0.0132**	**0.0185**

3.1.2 Model assumptions

Normality of persuasiveness scores

To check whether linear mixed models are appropriate for analyzing persuasiveness scores, I looked at their distribution. Figure 3.4 presents the histogram and Q-Q plot.



(a) Histogram

(b) Q-Q Plot

Figure 3.4: Distribution of persuasiveness scores

The distribution approximates normality reasonably well. In the Q-Q plot, points align with the theoretical line even with some deviation at the tails. This justifies the use of Linear Mixed Models for the analysis of persuasiveness.

Zero-Inflation in donation data

For donation amounts, I assessed the prevalence of non-donors. Figure 3.5 shows the distribution, highlighting the large spike at zero.

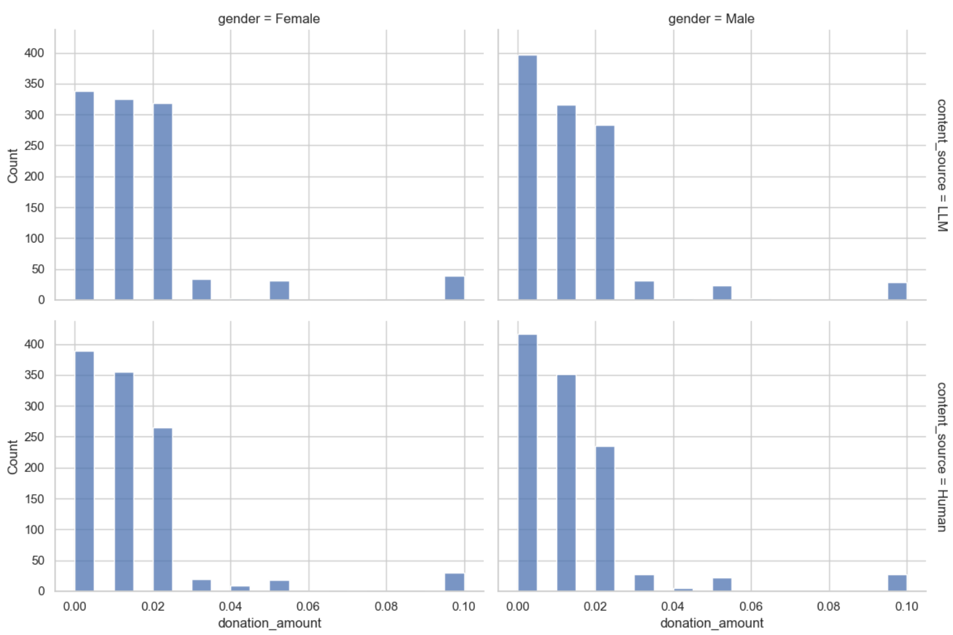


Figure 3.5: Distribution of donation amounts by gender and content source

Table 3.4 reports the proportion of zero donations.

Table 3.4: Proportion of zero donations

Gender	Source	Zero donations (%)
Female	Human	35.7%
	LLM	31.0%
Male	Human	38.0%
	LLM	36.3%
Overall		**35.3%**

With 35.3% zero values overall, this zero-inflation justifies using a Hurdle Model, which separately models the decision to donate (binary) and the amount donated (truncated continuous).

3.1.3 Correlations

Finally, I inspected correlations among the three outcome variables. Table 3.5 presents the Pearson correlation coefficients.

Table 3.5: Correlations among outcome variables

	Engagement	Persuasiveness	Donation amount
Engagement	1.00		
Persuasiveness	0.49	1.00	
Donation amount	0.15	0.19	1.00

The results show positive but moderate correlations. Engagement and Persuasiveness share the strongest relationship ($r = 0.49$), while Donation amount is more weakly correlated with both Engagement ($r = 0.15$) and Persuasiveness ($r = 0.19$). This suggests that the outcomes are related but measure separate constructs, justifying analyzing all three separately.

3.2 Phase 2: Main effects models

This section presents the results of the mixed-effects models estimating the main effects of gender and content source.

3.2.1 Engagement model

I fitted a Cumulative Link Mixed model (Ordinal logistic mixed-effects regression) to analyze engagement ratings. The model included fixed effects for gender and content source, with random intercepts for participants and posts.

Table 3.6 presents the estimated odds ratios (OR), 95% confidence intervals (CI), and p-values.

Table 3.6: Engagement model results (Ordinal logistic mixed-effects)

Predictor	OR	95% CI	p-value
Gender (Female)	1.37	[0.92, 2.04]	.125
Content Source (LLM)	1.58	[1.29, 1.93]	$< .001$

Reference categories: Male and Human.

The results show that content source had a significant effect on engagement. LLM-generated appeals had 58% higher odds of receiving a more favorable engagement rating compared to human-written appeals ($OR = 1.58$, $p < .001$).

Gender did not have a statistically significant main effect on engagement ($OR = 1.37$, $p = .125$). The confidence interval [0.92, 2.04] includes 1, suggesting women were not significantly more likely than men to give higher engagement ratings overall.

3.2.2 Persuasiveness model

I fitted a Linear Mixed-Effects Model to analyze persuasiveness scores. The model included fixed effects for gender and content source, with random intercepts for participants and posts.

Table 3.7 presents the unstandardized coefficients (b), 95% confidence intervals, and p-values.

Table 3.7: Persuasiveness model results (linear mixed-effects)

Predictor	b	95% CI	p-value
Gender (Female)	0.06	[-0.14, 0.26]	.544
Content Source (LLM)	0.20	[0.12, 0.28]	$< .001$

Reference categories: Male and Human.

Consistent with engagement results, content source had a significant positive effect on persuasiveness. LLM-generated appeals were rated as significantly more persuasive than human-written appeals, with an increase of 0.20 points on the 7-point scale ($b = 0.20$, $p < .001$).

Gender did not significantly predict persuasiveness scores ($b = 0.06$, $p = .544$), suggesting that men and women found the appeals equally persuasive on average.

3.2.3 Donation model

Given the zero-inflated nature of donation data, I employed a Hurdle mixed-effects Model to separately analyze the decision to donate and the amount donated among donors.

Conditional model (Amount donated)

Table 3.8 presents results for the conditional component, predicting donation amount for those who donated.

Table 3.8: Donation Model: conditional component (among donors)

Predictor	b	95% CI	p-value
Gender (Female)	0.0013	[0.0001, 0.0024]	.026
Content Source (LLM)	0.0018	[0.0004, 0.0032]	.011

Reference categories: Male and Human.

Among participants who chose to donate, women donated significantly higher amounts than men ($b = 0.0013$, $p = .026$). Also, LLM-generated appeals got significantly higher donation amounts compared to human-written appeals ($b = 0.0018$, $p = .011$).

Zero-Inflation Model (decision to donate)

The zero-inflation component, which predicts the likelihood of *not* donating, did not identify any statistically significant predictors. Neither gender ($p = .412$) nor content source ($p = .287$) significantly influenced the binary decision of whether to donate. This suggests these factors influenced the *magnitude* of generosity among donors but not the initial threshold decision to give.

3.3 Phase 3: Interaction analysis

To test whether the effect of content source depends on gender, I added a Gender $\times$ Content Source interaction term to each model.

Table 3.9 summarizes the interaction test results.

Table 3.9: Summary of interaction effects (Gender $\times$ Content Source)

Outcome	Interaction Coef.	p-value	Significant?
Engagement	OR = 1.53	.011	Yes
Persuasiveness	$b = 0.06$	.428	No
Donation amount	$b = 0.0002$	.735	No

3.3.1 Engagement interaction

The interaction between gender and content source was statistically significant for engagement ($OR = 1.53$, $p = .011$).

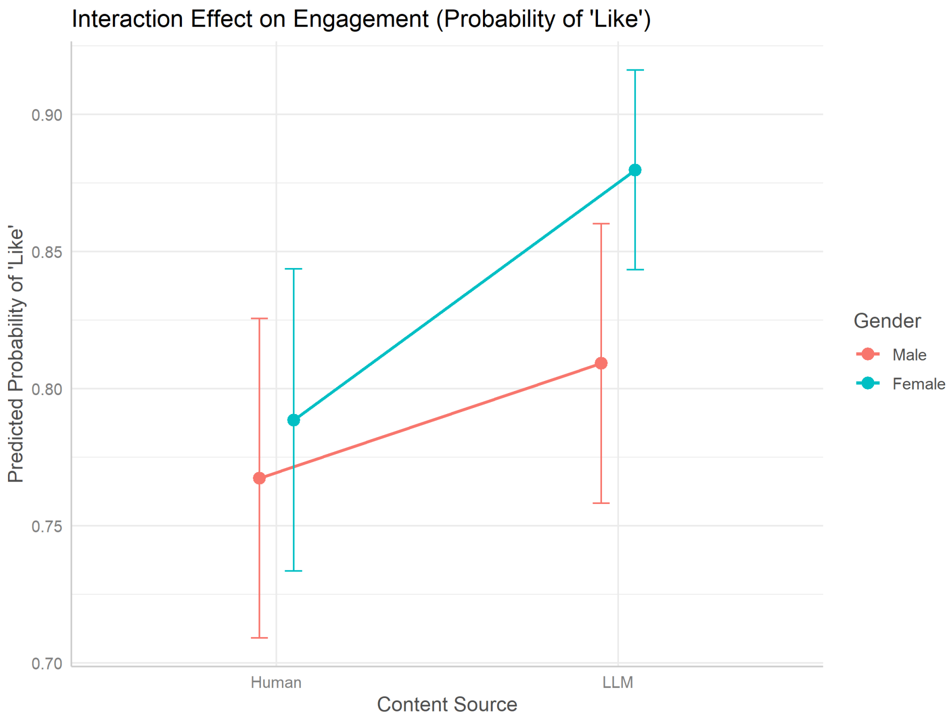


Figure 3.6: Interaction effect on engagement (Probability of 'Like')

Figure 3.6 shows that the preference for LLM-generated content was stronger among women than men. The odds ratio of 1.53 suggests that the positive effect of LLM content (vs. human) on engagement is significantly larger for female participants. In this graph, the Y-axis shows the chance that a participant will choose the 'Like' rating, and the X-axis shows the content source (Human or LLM). The two lines represent men and women. Both lines go up, meaning everyone preferred the LLM content. But, the line for women is steeper than the line for men. This steeper slope shows the significant interaction I found: the boost in engagement from LLM content is bigger for women than it is for men.

3.3.2 Persuasiveness interaction

For persuasiveness, the interaction term was not statistically significant ($b = 0.06$, $p = .428$).

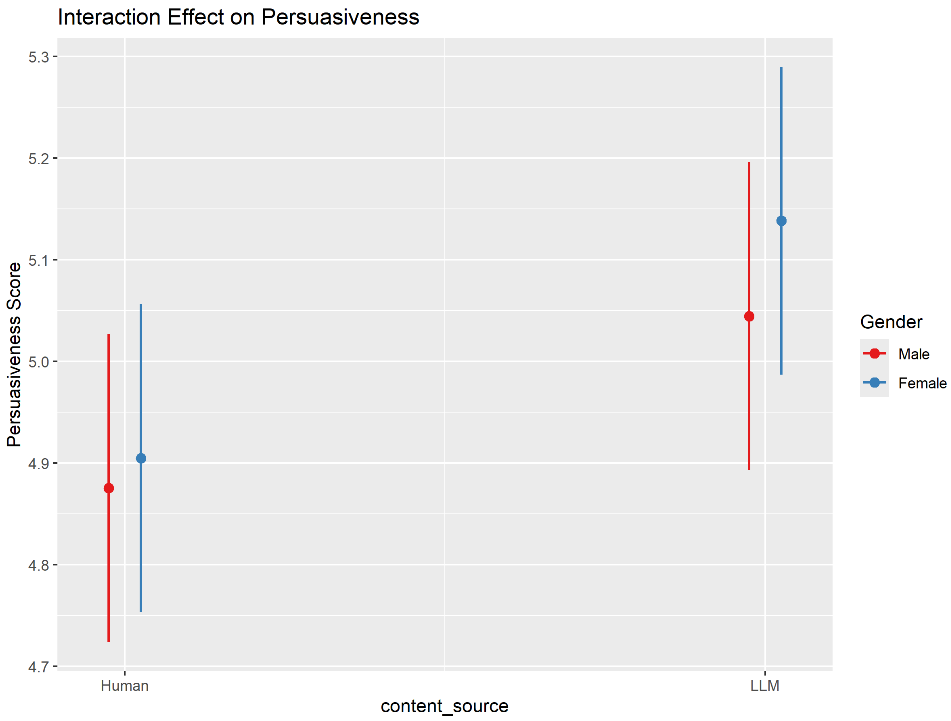


Figure 3.7: Interaction effect on persuasiveness

Figure 3.7 illustrates the interaction pattern, where the Y-axis shows the persuasiveness score (1-7 scale) and the X-axis shows the content source. Both lines go up, showing that LLM content was rated higher. even if a larger gap between male and female responses appears visible for LLM-generated content, the statistical test confirms this difference is not significant. I cannot conclude the gender gap differs by content source for persuasiveness.

3.3.3 Donation interaction

The interaction effect on donation amounts among donors was not statistically significant ($b = 0.0002$, $p = .735$).

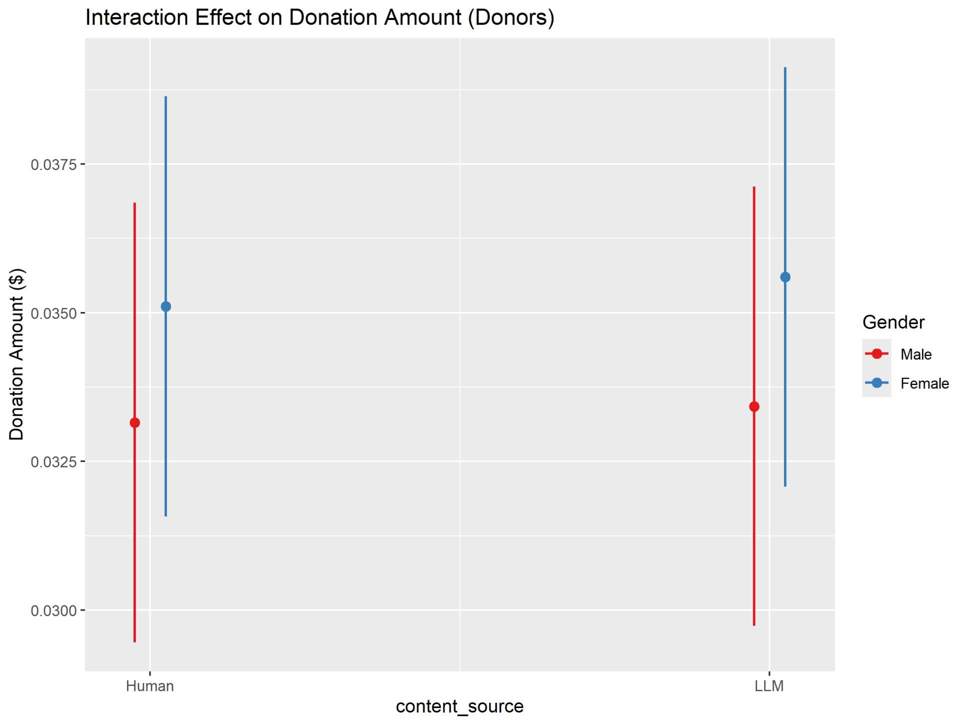


Figure 3.8: Interaction effect on donation Amount

This plot (Figure 3.8) shows the donation amounts for participants who chose to donate. The Y-axis is the dollar amount, and the X-axis is the content source. The female line is higher than the male line, which shows that women donated more. Both lines also go up for the LLM content, meaning LLM appeals led to higher donations. The lines are almost parallel, which confirms there is no significant interaction. This means the increase in donation amount from using LLM is about the same for both men and women.

so, a significant Gender $\times$ Content source interaction was found only for engagement, showing women's preference for LLM-generated appeals is stronger than men's. For persuasiveness and donation behavior, the effect of content source did not significantly differ by gender.

Chapter 4

Discussion

This study questioned whether men and women respond differently to LLM-generated versus human-written donation appeals for cancer charities. Using data from 725 participants evaluating 4,350 charitable appeals, I tested whether the well-documented gender gap in charitable giving remains across different content sources. The results show a clear pattern. LLM-generated appeals outperform human-written appeals on all three outcomes. Gender moderates this effect only for immediate emotional engagement, not for persuasiveness judgments or donation behavior.

4.1 Main findings and interpretation

4.1.1 The LLM advantage across outcomes

One of the clearest findings is that LLM-generated appeals performed better than human-written appeals on every outcome. Participants had higher engagement, judged the appeals as more persuasive, and donated slightly more money after reading LLM-generated text. These effects held up even after I factored in individual differences and the nested structure of the data.

Recent work helps explain this pattern. Modern language models have been shown to outperform human persuaders, even when those humans are financially incentivized (Schoenegger et al., 2025). This suggests that LLMs have reached a steady level of persuasive skill that actually outperforms the average human writer. Instead of being unusual or artificial, LLM-generated messages may show an altered version of persuasive techniques learned from large amounts of existing text.

This interpretation fits with the idea that LLMs reliably produce clear, fluent, and structured appeals. These qualities may be enough to improve engagement and persuasion in donation contexts, even without genuine emotional experience.

4.1.2 Gender and content source

Gender differences appeared only at the level of immediate engagement. Women reacted more positively to LLM-generated appeals at this early stage. This difference did not appear for persuasiveness judgments or donation amounts. Specifically, the analysis did not support H1, as the interaction between gender and content source regarding perceived persuasiveness was not statistically significant.

A useful explanation comes from research on fast and slow evaluation processes. First impressions of LLM-generated content are often formed through quick, automatic processing, instead of thought through evaluation. (Li et al., 2025). Slower and more reflective processing is used for judgments that require evaluation or action. Engagement ratings are likely driven by these quick reactions, which are sensitive to tone and emotional smoothness. While the theoretical background in Chapter 1 suggested that women might be more skeptical of artificial emotional text (Dietvorst et al., 2015), these results suggest that the high linguistic quality of LLMs effectively outweighs the aversion on the intuitive level.

Research on LLM empathy supports this explanation. LLM-generated empathic responses often receive higher ratings for comfort, validation, and emotional understanding than human responses (Wenger, 2024). At the same time, people remain hesitant to seek empathy from LLM when given a choice. This pattern suggests that LLMs can trigger strong emotional reactions without fully replacing the preference of human written text. The discrepancy between the significant engagement ratings and the non-significant results for persuasiveness reveals a "sincerity gap." Because participants were not told of the authorship, the psychological mechanisms of algorithm aversion were likely not activated. This suggests that such aversion is driven by the perceived identity of the source rather than the text itself.

4.1.3 Gender differences in donation amounts

A familiar pattern appeared in donation behavior. Among participants who chose to donate, women gave more money than men. This difference was small but meaningful given the limited donation budget used in the experiment. The same pattern appeared for both human-written and LLM-generated appeals. Because the gender gap in giving did not change based on the source of the message, H2b was not supported.

Gender did not predict whether participants donated at all. Men and women were equally likely to give something. The difference emerged only in the amount donated. This suggests that gender differences relate more to generosity than to willingness to participate. These results support the idea that Gilligan's Gilligan (1982) "Care" orientation is a strong predictor of behavior, no matter which mediator is involved. The fact that this pattern holds for LLM-generated appeals suggests that Women's higher donation rates seem to be driven more by the emotional distress described in the message than by who actually wrote it. Even when the appeal is generated by a machine, women's focus on the "Harm/Care" foundation (Graham et al.,

2011) remains the dominant driver of their prosocial behavior, as the machine-generated text successfully copies the emotional cues necessary to get a response.

4.2 Theoretical implications

These findings build on earlier research on gender differences in social behavior and apply it to settings mediated by LLMs. The psychological drivers that usually explain why women give more to charity still seem to be true, even when the appeal is written by an LLM.

The results also add to debates about trust in LLM-generated content. Evidence shows that LLM-generated responses can feel emotionally supportive and sensitive. At the same time, The perceived effectiveness of an empathic response is significantly reduced when the recipient knows the source is an LLM (Yin et al., 2024). Since participants in this study were not told about authorship, the results show responses to the text itself rather than how they look toward LLMs.

Lastly, our results point to a larger issue regarding emotional truth. LLMs are now capable of mimicking deep emotions despite having no actual conscience or moral foundation (Cuadra et al., 2024). In this study, The strong engagement I saw might just show that the emotional triggers hit the right buttons, not necessarily that people actually cared about the cause. This does not reduce the practical effectiveness of LLM-generated appeals, but it does suggest a need for care when using fake empathy in charitable contexts.

4.3 Practical implications for nonprofit organizations

For nonprofit organizations, the results suggest that LLMs can generate donation appeals that perform at least as good as human-written content. This may be especially useful for organizations with limited resources.

Disclosure policies must be cautious. Earlier research shows that labeling content as LLM-generated can reduce its impact. Organizations may face a trade-off between transparency and effectiveness.

The distinction between donating at all and donation amount suggests different strategies for different goals. High-quality appeals appear sufficient to prompt donations. Increasing donation amounts may require more approaches beyond appeal wording.

Chapter 5

Conclusion

This study looked at whether men and women respond differently to donation requests created by LLM. I collected data from 725 participants to test if the source of the message changes how people give. The results show that LLM-generated content outperforms human-written text. Participants engaged more with the LLM messages, rated them as more persuasive, and donated larger amounts. This was true for both male and female participants.

I found a gender difference in only one area. Women showed a stronger initial preference for LLM messages than men did. But this reaction did not change their judgments of persuasiveness or their final donation amounts. Women were more generous than men on average. But the difference in giving between genders was the same for human and LLM appeals.

The findings suggest that gender does not limit the value of LLM in fundraising. The data did not support the idea that women would dislike artificial emotional appeals. Nonprofits can use these tools to improve outcomes without pushing away female donors. Donors respond to the quality of the content, not to the author.

Bibliography

Bekkers, R. and Wiepking, P. (2011). A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly*, 40(5):924–973.

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474.

Cuadra, A. et al. (2024). The illusion of empathy? notes on displays of emotion in human-computer interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113–126.

Diekman, A. B. and Clark, E. K. (2015). Beyond the damsel in distress: Gender differences in prosocial behavior and implications for outreach to women. *Psychology of Women Quarterly*, 39(2):159–175.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Gilligan, C. (1982). *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press, Cambridge, MA.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385.

Gray, K., Young, L., and Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, 23(2):101–124.

Li, X. et al. (2025). How do users perceive ai? a dual-process perspective. *Computers in Human Behavior Reports*, 11:100166.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Mesch, D. J., Brown, M. S., Moore, Z. I., and Hayat, A. D. (2011). Gender differences in charitable giving. *International Journal of Nonprofit and Voluntary Sector Marketing*, 16(4):342–355.

Schoenegger, P. et al. (2025). Large language models are more persuasive than incentivized human persuaders. *arXiv preprint arXiv:2505.09662*.

Wenger, J. (2024). Participant preference for human versus ai empathy. Research talk, Empathy and Moral Psychology Lab. Available on YouTube.

Winterich, K. P., Mittal, V., and Ross, W. T. (2009). Donation behavior toward in-groups and out-groups: The role of gender and moral identity. *Journal of Consumer Research*, 36(2):199–214.

Yin, M., Jia, L., and Wakslak, C. (2024). Ai can help people feel heard, but an ai label diminishes this impact. *Proceedings of the National Academy of Sciences*, 121(10):e2313923121.