EXAMINING THE EFFECTS OF SOCIAL MEDIA BOTS ON ONLINE DISCUSSIONS: EVIDENCE FROM AN OBSERVATIONAL AND AN EXPERIMENTAL STUDY

By

Ruth Jin-Hee Heo

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Communication – Doctor of Philosophy

2025

ABSTRACT

This dissertation explores the impact of bots on online public discourse, specifically focusing on human users' language and attitudes in response to online interactions. The findings from Study 1 suggest that while bots generally exhibited lower levels of politicization, polarization, and neutrality, they displayed higher levels of anger, disgust, fear, and joy. Some of these features, specifically, politicization and disgust of bots can still influence humans over time. Furthermore, Studies 2 and 3 compared the effects of LLM-generated content versus non-LLM-generated content on individuals' attitudes. The results show that LLM-generated contents can subtly influence users, as individuals often struggle to distinguish between human-like and machine-generated content on social media. As bots become more sophisticated with technological advancements, they are increasingly capable of shaping human attitudes in ways that are nearly indistinguishable from human interactions. Given the pervasive use of such technologies on social media, understanding their relational impact on humans is becoming crucial.

TABLE OF CONTENTS

INTRODUCTION

Previous research has consistently shown that the presence of bots increased the likelihood of social media users' exposure to triggering and/or biased content (Bail et al., 2020; Badawy et al., 2018). Specifically, bots played a role in the dissemination of misinformation and the exacerbation of conflicts during significant events, such as the 2016 U.S. presidential election (Bail et al., 2020; Badawy et al., 2018; Keller & Klinger, 2019; Shao et al., 2017), the COVID-19 pandemic (Ferrara, 2020; Xu & Sasahara, 2021), and the Catalan referendum for independence (Stella et al., 2018). However, bots' direct influences on humans remain largely speculative (e.g., Aldayel & Magdy, 2022; Bail et al., 2020; Caldarelli et al., 2020; Ferrara, 2020).

In recent years, the emergence of large language models (LLMs) has transformed the landscape of social media platforms. LLMs enable users to deploy bots effectively and produce highly deceptive content in a rapid fashion (Zhang et al., 2024). For example, this includes the viral video that falsely depicted Volodymyr Zelensky, the president of Ukraine, announcing a surrender to Russia (Satariano & Mozur, 2023). LLMs not only disrupt the content produced in social media but also have contributed to the emergence of more nuanced and sophisticated bots on the platforms. With the rise of so-called LLM-powered bots, which exhibit increasingly human-like features, there is potential for these bots to exert social influence equivalent to that of humans. Nonetheless, a lack of evidence leads to questioning the extent of the impact of information provided by these bots, how this will influence humans, and what the consequences of this influence may be.

The purpose of the dissertation is to examine the potential influence of social media bots' presence on humans in the context of online discussions about genetically modified organisms

(GMOs). GMOs have emerged as a technological advancement aimed at enhancing the quality and quantity of agricultural crops (Rathod & Hedaoo, 2022; Sohi et al., 2013). Historically, GMOs have been a controversial topic, with divided opinions—some people emphasize their benefits, while others focus on the potential risks. This context provides a valuable opportunity to examine how bots may shape public discourse, especially in light of the ongoing controversies surrounding GMOs. This research comprises three studies, and Study 1 utilizes observational data to examine bots' role in social media discourse by analyzing how their content influences that of humans. Using time series prediction, it analyzes linguistic elements in tweets to determine the extent to which bots' linguistic features affect those of humans or the extent to which humans' linguistic features affect those of bots. Study 2 further employs a controlled experiment to investigate the causal influence of LLM-generated content as well as non-LLM-generated content on human attitudes. Specifically, this study adapts methods involving LLMs to create personas based on respondents' answers from a nationally representative poll (Hewitt et al., 2024). In Study 3, actual human subjects were recruited correspondingly to the same experimental setup of Study 2, which increases the generalizability of the results as well as reinforces the validity of Study 2's results. The present research not only assesses the overall impact of bots but also validates the growing influence of LLM-generated content in the online environment. Given the concerns over the drastic development of the new technology, the findings could contribute to contemplating the ways to mitigate unforeseeable consequences driven by them.

LITERATURE REVIEW

Social Media Bots and their Impacts on Humans

Bots are automated software programs designed to simulate human behavior. While they can facilitate interactions with users, bots are often deployed to disseminate information according to specific agendas. In some cases, bots serve benign purposes, potentially promoting free speech, political discourse, and even social activism (Ferrara et al., 2016; Gorwa & Guilbeault, 2020; Savage et al., 2015). However, recent research has focused on the malicious intent of implementing bots for manipulating public discourses. Bots have been identified as key contributors to the spread of misinformation, hate speech, and identity theft (Stocking & Sumida, 2018). Specific to social media platforms, bots often masquerade as human users, thereby polluting public conversations. Their malicious influence on social media, specifically Twitter (now X), was repeatedly reported as bots disseminated politically charged posts to amplify a political agenda (Bail et al., 2020; Badawy et al., 2018), violent content that contained negative and inflammatory narratives to intentionally induce conflicts among politically divisive groups (Stella et al., 2018), and misinformation that interrupted credible information sources (Broniatowski et al., 2018; Shao et al., 2017; Xu & Sasahara, 2022).

Despite predicaments caused by bots, such as repeated exposure to harmful, triggering content, it is yet unclear if their influence is actually strong enough to directly change one's attitude toward a topic (e.g., Aldayel & Magdy, 2022; Bail et al., 2020). For example, Bail et al. (2020) examined the role of the Russian Internet Research Agency (IRA) in the context of the 2016 U.S. presidential election; results indicated bots had a role in polluting the public discourse over the election, but their actual influence on other users' attitudes and behaviors has not been confirmed (Bail et al., 2020). Similarly, Aldayel and Magdy (2022) investigated bots' effect on

users' stances in the context of political and social events such as climate change, feminism, Brexit, and other movements. Their findings suggested that the relatively rare occurrences of social media bots and lack of evidence for a direct relationship between bots and users' stances led to an inconclusive effect of bots.

While these studies advise caution in interpreting bots' influence, emerging research suggests that bots powered by LLMs could have a more substantial impact on human opinion. This new powerful technology enables bots to be easily implemented, efficiently generate deceptive content, and actively interact with other users (Zhang et al., 2024). As this technology advances, the potential of bots to influence humans may increase, although the magnitude of the influence remains uncertain (Burtell & Woodside, 2023).

CASA Framework and Bots' Social Influence on Humans

The computers-as-social-actors (CASA) theory (Reeves & Nass, 1996) offers some insights for gauging the effect of bots on humans. Originally, the CASA theory focused on human-computer interactions by examining specific cues emitted by computers. Unlike traditional media (e.g., televisions, newspapers, radio, etc.), people perceive computers as independent sources rather than mere channels or mediums in the communication process (see review in Sundar & Nass, 2000; 2001; Hocevar et al., 2017). As independent sources, computers interact with humans eliciting mindless social responses from users in ways similar to human-to-human interactions. Specifically, people were found to apply social rules in their interactions with computers in responding to specific cues from computers.

Initially, Nass and Steuer (1993) proposed that cues related to the use of language (as communication tools), interactivity, the assignment of social roles typically played by humans, and human sounding speech were most likely to elicit automatic social responses. Then, later

studies identified and tested the cues specifically gender, ethnicity, reciprocal self-disclosure, and simple labels designated a computer as 'specialist,' all of which led users to apply social rules to computers, even when acknowledging that computers are as non-human (Leshner et al., 1998; Nass & Moon, 2000, e.g., Nass et al., 1997; Nass et al., 2000; Moon, 2000). For example, in a lab experiment, participants rated a male voiced computer as friendlier than a female voiced computer when evaluating their performance. Moreover, the same study found that a female voiced computer was perceived as more expert in matters of love and relationships compared to a male voiced computer. In other words, participants mindlessly gender-stereotyped computers based on the voice cue (Nass et al., 1997; Nass & Moon, 2000). These findings suggest that specific characteristics of computers affect people to apply mindless social responses to computers, treating computers as social actors akin to humans (Qiu & Benbasat, 2009, e.g., Nass et al., 1994, 1995; Nass & Moon, 2000).

In recent years, similar mindless reactions have been consistently observed when people interact with emerging technologies, such as voice activated navigation systems (Nass et al., 2005), the virtual assistant Alexa (Schneider, 2020), smartphones (Carolus et al., 2019), and in a study exploring how humans responded to a "lost" robot (Srinivvasan & Takayama, 2016). These results highlight that technologies (a broad category of computers) that possess specific features enable active social interactions. In other words, it is plausible that humans may be influenced by bots to a similar extent as they are by other humans. This influence may be amplified by the advent of more sophisticated "human-like" technologies, such as bots powered by LLMs. LLMs meet the components identified by Nass and Steuer (1993) as contributing to mindless responses among people, as mentioned earlier. These LLM-enhanced bots, which

mimic human behavior more closely and exceed human skills in various areas, could intensify the social influence exerted by bots in online interactions.

Hypotheses and Research Questions

Early studies raised concerns over bots, as they tend to distribute politically and emotionally triggering content as well as biased information (i.e. misinformation), and the increasing sophistication of bots raises concerns about tainted public discourse. This tendency indicates that bots' content will be distinct from that of humans. Moreover, bots are likely to generate more politically inciting (i.e. politicizing), biased (i.e. polarizing), and emotionally intense (i.e. negatively piqued) language compared to human users in a wide range of issues (e.g., Badawy et al., 2018; Bail et al., 2020, the U.S. Election; Broniatowski et al., 2018, vaccines; Xu & Sasahara, 2022, COVID-19 pandemic). In the context of GMOs, they also tend to display such tendencies in their tweets amid mixed views over the topic. Capturing this manifestation, the current dissertation breaks these tendencies down into four specific features–topical themes, politicization, polarization, and emotions (e.g., anger, fear, disgust, joy)–and assesses them respectively. In examining the linguistic features of bots' content compared to those of humans, I propose the following hypotheses:

H1: The topical themes of bots' tweets will differ from those of humans' tweets.

H2: Bots' tweets will be more likely than humans' tweets to exhibit 1) politicization, 2) polarization, and 3) negative emotions, while exhibiting less 4) neutrality and 5) positive emotion.

H2-1: Bots' tweets will exhibit more politicization than humans' tweets.

H2-2: Bots' tweets will exhibit more polarization than humans' tweets.

H2-3: Bots' tweets will exhibit more negative emotions, including a) anger, b) disgust, c) fear) than humans' tweets.

H2-4: Bots' tweets will exhibit less neutrality and positive emotion (i.e., joy) than humans' tweets.

Despite the high likelihood of bots containing distinct linguistic features compared to those of humans, there are only speculations to support the direct impact of bots on humans (e.g., Bail et al., 2020; Aldayel & Magdy, 2022). That is, the inquiry of whether bots' politicization, polarization, and emotions (e.g., anger, disgust, and fear) in their content directly influence the content of human users will be investigated. Thus, I pose the following research questions:

RQ1: Does the politicization of bots' tweets influence the politicization of humans' tweets?

RQ2: Does the polarization of bots' tweets influence the polarization of humans' tweets?

RQ3: D Do the negative emotions in bots' tweets—1) anger, 2) disgust, and 3) fear—influence the emotions in humans' tweets?

Previous research on the perceived communication quality of bots has shown that people often have positive perceptions of bots (Edwards et al., 2014; Edwards et al., 2016). Additionally, in some conversational contexts, interruptions made by artificial intelligence (AI) (e.g, LLM) have been found to improve relationships in human-human interactions (Hohenstein & Jung, 2019). However, when comparing content labeled with a human cue versus a bot cue, people tend to evaluate the content with a human label more positively, regardless of the actual source (Chu & Liu, 2024; Graefe et al., 2018; Karinshak et al., 2023). With regard to the quality of content produced by bots versus humans, some studies have compared AI-generated news articles with articles written by human journalists. Overall, people assessed both types of articles as equally descriptive, boring, and objective (Clerwall, 2014). In some cases, people rated AI-

generated texts as more effectively delivered and more believable (Graefe et al., 2018; Karinshak et al., 2023). For instance, when the source of the news was arbitrarily assigned, people often assessed AI-generated news as more credible and possessing greater journalistic expertise (Graefe et al., 2018). Moreover, people reported that AI-generated content seemed to have stronger arguments and more positive effects on attitudes compared to human-written content (Karinshak et al., 2023). However, when AI-generated stories were compared with human written stories, people found AI-stories similar or less engaging and similar or higher counterarguing from participants (Chu & Liu, 2024).

In sum, it remains uncertain whether individuals are capable of distinguishing between LLM-generated content and non-LLM-generated content. LLM-generated content can vary depending on the original input provided to the model. Although the content is produced by an LLM, the underlying source may differ, originating from either bots or humans. Specifically, it is unclear whether people can differentiate between different types of LLM-generated content (e.g., LLM-generated bot content and. LLM-generated human content) and non-LLM-generated content (e.g., human content and bot content). Additionally, it is uncertain whether LLM-generated content can be distinguished from content generated by (real) humans or (real) bots. Therefore, I propose the following research questions:

RQ4: Do people perceive LLM-generated content and non-LLM-generated content differently in terms of bot-likeness and human-likeness?

RQ4-1: Do people perceive human content and LLM-generated bot content differently in terms of (a) bot-likeness and (b) human-likeness?

RQ4-2: Do people perceive bot content and LLM-generated bot content differently in terms of (a) bot-likeness and (b) human-likeness?

RQ4-3: Do people perceive human content and LLM-generated human content differently in terms of (a) bot-likeness and (b) human-likeness?

RQ4-4: Do people perceive bot content and LLM-generated human content differently in terms of (a) bot-likeness and (b) human-likeness?

RQ5: Do people perceive LLM-simulated bot content and LLM-generated human content differently in terms of (a) bot-likeness and (b) human-likeness?

RQ6: Do people perceive bot content and human content differently in terms of (a) bot-likeness and (b) human-likeness?

Despite some research examining the effects of LLM-generated content versus non-LLM-generated content on perceptions of quality (Chu & Liu, 2024; Clerwall, 2014; Graefe et al., 2018; Karinshak et al., 2023), the impact of bots—specifically the extent to which they influence attitudes—remains unclear. Therefore, the following set of research questions is posed to compare LLM-generated content and non-LLM-generated content:

RQ7: Do LLM-generated content and non-LLM-generated content affect individuals' attitudes differently?

RQ7-1: Do LLM-generated bot content and human content affect individuals' attitudes differently?

RQ7-2: Do LLM-generated bot content and bot content affect individuals' attitudes differently?

RQ7-3: Do LLM-generated human content and human content affect individuals' attitudes differently?

RQ7-4: Do LLM-generated human content and bot content affect individuals' attitudes differently?

RQ8: Do LLM-generated human content and LLM-generated bot content affect individuals' attitudes differently?

RQ9: Do human content and bot content affect individuals' attitudes differently?

Overview of Studies

Topical Context: Genetically Modified Organisms

This study attempts to address existing gaps in research on social media bots, particularly their influence on public discussion regarding GMOs. GMOs have emerged as a technological advancement aimed at enhancing the quality and quantity of agricultural crops (Rathod & Hedaoo, 2022; Sohi et al., 2013). In the U.S., this biotechnology has led to an increase in crop yield, covering approximately 71.5 million hectares in 2019 (Catherine et al. 2024). Beyond quantity, GMOs also contribute to improved nutritional quality; for example, L-1 transgenic corn has shown increases in Vitamin C, beta-carotene, and folate compared to its non-GMO counterparts (Naqvi et al., 2009). Additionally, GMOs help reduce the use of pesticides and herbicides, thereby minimizing environmental hazards (Rathod & Hedaoo, 2022). However, public perception often associates GMOs with health and environmental risks (Catherine et al., 2024).

According to a public opinion survey conducted by the Pew Research Center (Funk, 2020), nearly half of Americans (48%) view GMOs as having positive or neutral health impacts, while 51% perceive them negatively. Notably, 30% of those who reported a neutral stance still expressed concerns about potential negative consequences of GMOs. Furthermore, another survey indicated that 54% of participants admitted to knowing very little or nothing at all about GMOs, and 25% claimed they had never heard of them (Hallman et al., 2013; Wunderlich &

Gatto, 2015). Discussions on Twitter (X) reflected similar trends, with 54% of posts being neutral and 32% negative, while only 14% expressed positive sentiment (Sohi et al., 2023).

Despite the consistent trends, studies by Jun et al. (2020) and Howell et al. (2018) revealed that public sentiment on Twitter (X) could shift significantly in response to real-world events, such as reports from the National Academies of Sciences, Engineering, and Medicine (NASEM) and the United States Department of Agriculture's (USDA) closure of public consultations on GMO regulations. Additionally, recent issues related to pandemics and climate change may have further influenced public perceptions of GMOs. These overlapping events highlight the complexities underlying public discussions about GMOs, which are shaped by various social dynamics, including the role of bots. This research aims to uncover how public discourse has evolved and the specific role that bots play in this context.

The current topic provides the context necessary to study the impact of bots' social influence on GMO discussions. To systematically explore the impact of bots on humans, the proposed dissertation conducted a three-part study.

Study 1: An Observational Study

Study 1 utilized observational data from Twitter (X) to analyze bots' direct influence on humans or the other way around. By examining linguistic components in social media posts, such as topical themes, politicization, polarization, and emotions, this study compared the nuances in content shared by bots and humans (H1, H2) and evaluated the direction of influence (RQ1-RQ3).

Studies 2 and 3: Controlled Online Experiments

The linguistic relationships identified in Study 1, however, are insufficient to directly claim a causal influence of bots on humans without establishing nonspuriousness. To address this gap, Studies 2 and 3 aim to investigate the causal influence of bots on human attitudes through a controlled online experiment. Specifically, in response to growing concerns about LLM-powered bots, these studies incorporate content generated by LLMs, including both LLM-powered bot content and LLM-powered human content, to assess whether people have ability to discern the source of different types of contents (RQ4-RQ6) and their effects on human attitudes (RQ7-RQ9). For comparison, content from bots and humans identified in Study 1 was used to evaluate the impact of LLM-generated content versus non-LLM-generated content. Given the complexity of the information ecosystem on social media, it is crucial to integrate diverse content types that are likely to appear on these platforms and assess their impact on one's attitudes. This second study simulated personas from the existing research (i.e., McFadden et al., 2024) using LLMs (i.e., Llama-3) to participate in the experiment. Further, Study 3 recruited actual human subjects to verify the results.

Table 1

Details of the Main Variables

Variable Name	Conceptual Definition	Operational Definition	Measurement Level	Units of Observations
Study 1				
Topics	Topics refer to frequently observed themes over time.	BERTopic generates topical themes based on closely clustered keywords.	Categorical	Aggregated themes of tweets across time
Bot status	Bot status indicates whether users are classified as bots or humans, with bots exhibiting human-mimicking behaviors that humans do not necessarily display.	Detecting bots with temporal information using LLMs (i.e., GPT-4)	Categorical	An Individual user
Politicization	Politicization refers to the extent to which an issue or event becomes personalized, with narratives shifting from broader economic, social, and political analysis to focus on competing actors, while also involving political figures who represent opposing factions, driving controversy within the political arena (Chinn et al., 2020).	Two dictionaries (i.e., *republican* and *democrat),* developed in Chinn et al. (2020) was implemented.	Continuous	The number of politicized words observed from an individual user

Table 1 (cont'd)

Polarization	Polarization refers to the degree of political bias, typically categorized as right-wing, left-wing, or centrist, that is often reflected in media (Baley et al. ,2020).	Pretrained models from hugging face (*bucketresearch/politicalBiasBERT*) was utilized to calculate the score of political bias in texts (Baly et al., 2020). The model initially adopted labels–left, center, and right–classifying and recoding using numeric values of -1, 0, or 1, respectively.	Categorical	Aggregated polarization of user's tweets
Emotions	Emotions in this study are based on Ekman's basic emotions, which include fear, anger, disgust, sadness, joy, and surprise. This study focuses on the expression of fear, anger, disgust, and joy, which are considered invariant across cultures and species (Ekman, 1992). Additionally, neutrality is included to account for objectivity in the linguistic features being examined.	Pretrained models from hugging face (*j-hartmann/emotion-english-distilroberta-base*) was utilized to measure the level of emotions (e.g., fear, anger, disgust, neutrality, joy) expressed in tweets on a scale from 0 to 1.	Continuous	Aggregated emotions of user's tweets

Table 1. (cont'd)

Study 2 and Study 3				
Attitude	Attitudes indicate one's stance towards GMO editing (McFadden et al., 2024).	Self-reported ratings of how likely individuals are to advocate for the position of the message, on a scale from 1 to 5.	Continuous	Individual participants
Bot likeness	Bot likeness refers to the extent to which people perceive the user of the message as bots.	Self-reported ratings of how likely individuals are to perceive the user of content created by bots, on a scale from 1 to 5.	Continuous	Individual participants
Human likeness	Human likeness refers to the extent to which people perceive the user of the message as bots.	Self-reported ratings of how likely individuals are to perceive the user of content created by humans, on a scale from 1 to 5.	Continuous	Individual participants

STUDY 1

Study 1 Research Design

Research Procedure

The dataset comprises 26,040,042 tweets which were extracted from July 2009 to September 2019. A list of hashtags and keywords (Table 2) was used to target and include posts relevant to GMOs, which were extracted through the valid application programming interface (API). Given the large size of the dataset, users were selected based on their activity levels. Specifically, the total number of posts by each user was calculated and 800 users from each quartile of activity was selected; in total 2,400 user accounts were selected, and the total of 1,449,994 tweets posted by these users were used for analysis. This approach maintains dataset representativeness while reducing its size. Before the analysis the data was cleaned in the procedure outlined in the following section. Once completed, the suggested variables were created based on the measurement description.

Table 2

Keywords and Hashtags used for Tweet Extraction

Keywords	Hashtags
gmo, gmos, gm food, gmfoods, gm foods, genetically modified, genetic modified, genetical modified, genetic modification, genetical modification, genetically modification, genetic engineering, genetical engineering, genetically engineering, genetical engineered, genetically engineered, transgenic, transgenics, transgenesis, transgenically, transgenes, transgene	#gmo, #gmos, #gmfood, #gmfoods, #gm_food, #gm_foods, #geneticallymodified, #genetically_modified, #geneticallymodifiedfood, #geneticallymodifiedfoods, #genetically_modified_food, #genetically_modified_foods, #geneticmodification, #genetic_modification, #geneticengineering, #geneticalengineering, #geneticallyengineering, #genetic_engineering, #geneticengineered, #genetically_engineered, #genetically, #transgenic, #transgenics, #transgenes, #transgene

Notes. Tweets written in English but published from all countries were extracted.

Data Processing

For topical analysis, tweets were preprocessed; however, the process was minimized as BERT is a transformer-based model which is already trained for contextualizing meanings between words. The extracted tweets were preprocessed using a package, *preprocessor*, which is available in Python. First, unnecessary features such as uniform resource locators (URLs), mentions, and reserved words (e.g., RT, FAV) were eliminated from the tweets; hashtags were kept while removing the sign (#) as they often capture meaningful content in tweets. Lastly, extra white spaces and punctuation were removed. Having cleaned the tweets of such content, all letters were changed to lower cases for uniformity. Except for topical analyses, raw full texts were provided.

Measurements

The conceptual and operational definitions for each variable are summarized in Table 1.

Bot Status. The present study used LLMs to detect bots, as tested in Heo et al. (2024). As recommended, temporal information, along with example cases, was provided to the LLMs to distinguish between bots and humans. Specifically, both correct and incorrect cases were presented as examples to enhance task performance. This approach was informed by multiple rounds of pilot studies, including coding by human experts (Appendix B). In the coding process, bots were assigned a value of 1, while humans were assigned a value of 0.

Topics. Latent topics were extracted using unsupervised topical modeling, specifically BERTopic (Grootendorst, 2022). Here, a topical modeling technique was applied, targeting bots, and human users, respectively.

Politicization. Two dictionaries (i.e., Republican and Democrat), developed by Chinn et al. (2020) were implemented to tally the frequency of words found in each dictionary (Appendix

A). Chinn et al. created these dictionaries as part of their exploration into politicization, specifically within COVID-19 news coverage. By analyzing the frequency of political party or political affiliation terms within various texts, their study demonstrates how such language can mark politicization, highlighting the subtle ways language may influence audience perception and contribute to politicization of an issue. This approach to tallying frequencies based on party-affiliated language provides a quantitative lens for assessing political discourse across diverse media channels and contexts.

Polarization. In investigating polarization, a pretrained model in Hugging Face, specifically *bucketresearch/politicalBiasBERT*, developed by Baly et al. (2020), was used to calculate the political bias score in texts. The model is designed to classify texts along a political spectrum, initially using labels–left, center, and right–and recoding them into numeric values of -1, 0, or 1, respectively. Baly et al. trained this model to recognize nuanced linguistic patterns and biases across political ideologies, allowing it to assess ideological leanings in a wide range of textual content. This numeric recoding facilitates a more standardized, quantitative analysis of political bias, enabling comparisons of language patterns across different political orientations.

Emotions. A sentence-level analysis will be conducted, which is an extension of the word–and lexical–level analysis that has been used in previous communication literature (Rudkowsky et al., 2018) using a pretrained model in Hugging Face (*j-hartmann/emotion-english-distilroberta-base*). Emotions, including anger, disgust, fear, joy, and neutrality, were evaluated on a scale from 0 (*no emotion*) to 1 (*extreme emotion*).

Table 3

Topical Themes of Bots and Humans

	Topic	Keywords
Bot		
	Gene editing technology	geneediting, im, think, those, genomeediting, crisprcas9, benefits, ag, my, talk
	GMO TweetZUP[1] trending	1h, tweetzup, trending, page, been, has, for, the, gmo, popped
	Non-GMO snack	everybody, perfect, glutendairy, snack, bars, kosher, healthy, keep, free, nongmo
	Non-GMO snack	ideal, everybody, glutendairy, snack, bars, kosher, healthy, keep, free, nongmo
	Non-GMO snack	ideal, everyone, glutendairy, snack, bars, kosher, healthy, keep, free, nongmo
	Non-GMO snack	perfect, everyone, glutendairy, snack, bars, kosher, healthy, keep, free, nongmo
	Non-GMO snack	ideal, everyone, glutendairy, snack, bars, kosher, healthy, keep, free, nongmo
	Non-GMO snack	ideal, everybody, glutendairy, snack, bars, kosher, healthy, keep, free, nongmo
	Non-GMO snack	everybody, perfect, glutendairy, snack, bars, kosher, healthy, keep, free, nongmo
	Non-GMO snack	perfect, everyone, glutendairy, snack, bars, kosher, healthy, keep, free, nongmo
Human		
	The GMO controversy	are, my, but, if, crops, gmos, not, so, we, organic
	AquAdvantage Salmon	salmon, frankenfish, fish, consumption, aquabounty, fda, approved, wild, approves, animal
	H.112—Protecting the right to know GMOs	vermont, vermonts, law, vt, effect, treading, passes, attorney, vts, h112
	Physicians for social responsibility supports mandatory labeling of GMOs	psr, responsibility, physicians, social, support, foods, for, labeling, of, gmo

[1]TweetZUP is a free tool that alerts users to trending topics and real-time events on Twitter (X).

Table 3 (cont'd)

A bill placing a moratorium on the cultivation of GMOs	maui, hawaii, moratorium, kauai, county, hawaiis, island, judge, mauis, hawaiians
Glyphosate and health related concerns	glyphosate, probable, carcinogen, deterioration, glyphosates, urine, genotoxic, classifies, roundup, geopolitics
Genetically engineered bacteria	newsgenetically, lives, save, bacteria, can, science, engineered, inoperable, listerine, anaerobic
non-GMO products	vegetariansafe, 5000mcg, biotin, nails, hair, skin, glutenfree, healthy, 100, here
GMOs fighting widespread bee-killing mites	bees, beekeepers, 'honey, 100000, bee, killing, neonicotinoid, permit, ants, flying
Conspiracy theories and public health beliefs	chemtrails, vaccines, fluoride, cdcwhistleblower, vaccine, skies, aluminum, radiation, sb277, water

Study 1 Results

After implementing the modified bot detection approach proposed by Heo et al. (2024), in a total of 2,400 user accounts, 1,742 user accounts (76%) were classified as bots whereas those of 448 (22%) were categorized as humans. The remaining 60 users (3%) were uncategorized; there was no particular distinction of this number of users, but this was attributed to a technical issue with GPT-4 when generating the response. A total of 1,449,994 tweets posted by 1,742 bots and 448 humans were used for the analysis.

In order to test H1, BERTopic was used to extract representative themes across tweets of each entity, and the 10 most dominant topics were selected in this report. In terms of bots' topics, the most prevailing topic included the keywords such as 'genediting,' 'genomeediting,'crisprcas9,'and 'benefits,' which referred to gene editing technologies; yet, the representative documents of this topic mostly questioned GMOs. Moreover, the same keywords were repeatedly shown in different topics; terms including 'glutendaily,' 'snack,' 'bars,' 'kosher,' 'healthy,' 'keep,' and 'nongmo' were repeatedly shown in seven other categories. The keywords in topics of bots implies the generic promotion of nonGMO foods. On the other hand,

topics of humans' tweets, there were more political and social controversies over GMOs. The most prevalent topic included terms such as 'if,' 'crops,' 'gmos,' 'not,' 'so,' and 'organic,' and relevant documents pointed out the GMO labeling issues and called out companies that support GMO foods. Similarly, other topics depicted social issues surrounding GMOs. Notably, the issue about genetically modified salmon approved by FDA was captured by the keywords like 'salmon,' 'frankenfish,' 'fish,' 'consumption,' 'aquabounty,' 'fda,' 'approved,' 'wild,' and 'approves,' 'animal.' Moreover, the issue of Maui County's moratorium on GMO crops in Hawaii was revealed with the terms including 'maui,' 'hawaii,' 'moratorium,' 'kauai,' 'county,' 'hawaiis,' 'island,' 'judge,' 'mauis,' and 'hawaiians.' That being said, the topics of tweets posted by humans and bots were discrete and bots were more likely to promote a nonGMO product whereas humans were more interested in discussing real life events related to GMOs (Table 3). Thus, the results support H1. Despite the distinct topics, it is important to note that some overlapping themes, such as questioning GMOs, suggest potential interactions between humans and bots.

To test the H2, the levels of politicization, polarization, and emotions including anger, disgust, fear, neutrality, and joy between bots and humans were compared. Mann-Whitney U tests were conducted due to the non-parametric nature of variables used in this test. Initially I posited that bots were more likely to post more politicized and polarized tweets. However, the results indicate that humans used more politicized and had more polarized languages in their contents compared to bots (politicization: $p = .0002$; polarization: $p < .001$, Table 4). In terms of negative emotions including anger, disgust, and fear, in alignment with the hypothesis, anger, disgust, and fear were shown to be higher among bots compared to humans (anger: $p < .001$; disgust: $p < .001$, fear: $p < .001$, Table 4). Furthermore, neutrality was higher for humans

(neutrality: $p<.001$, Table 4) but bots showed a higher level of joy compared to humans (joy: $p<.001$, Table 4). That is, H2 is partially supported: unexpectedly, humans were more likely to use polarized, politicized, neutral language, while expressing high anger, disgust, fear, and less joy compared to bots.

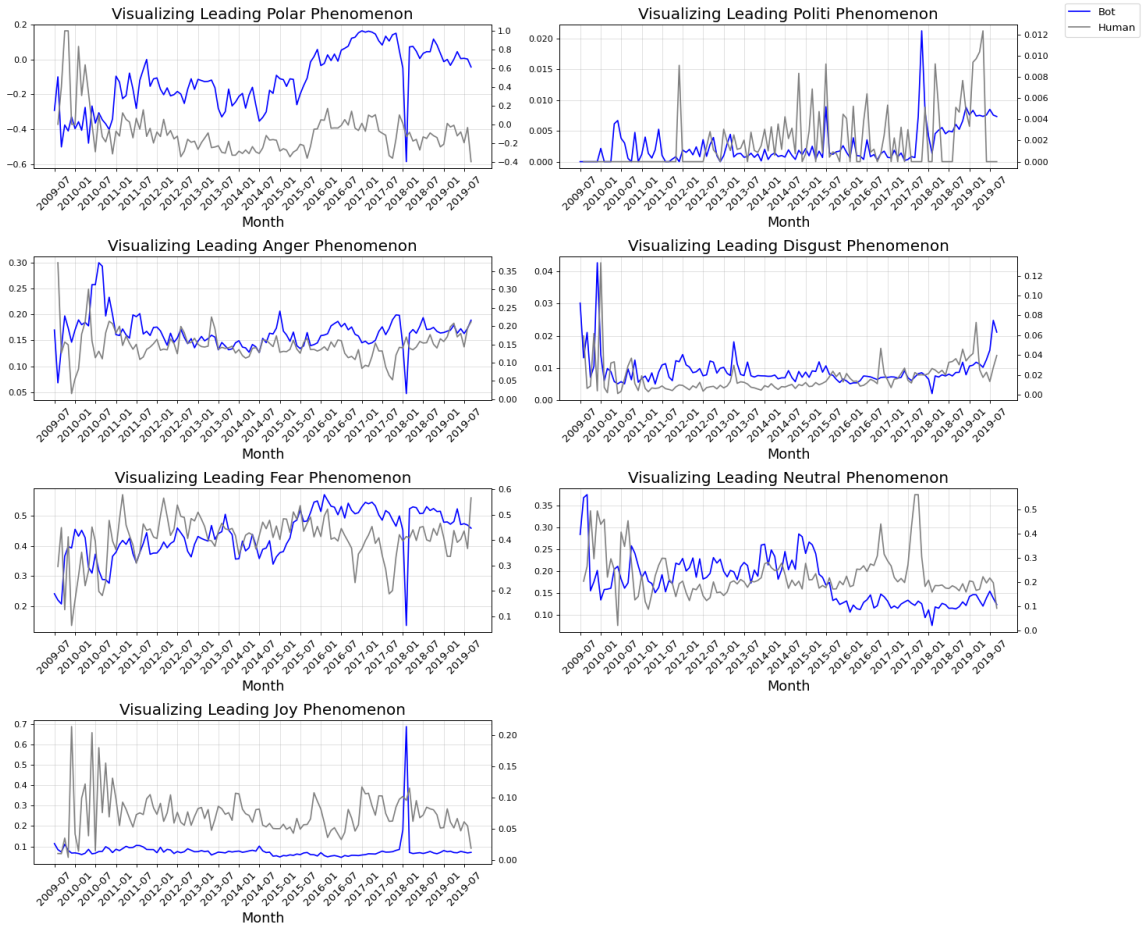
Table 4

Descriptive Statistics of Variables (Study 1)

	Polarization	Politicization	Anger	Disgust	Fear	Neutral	Joy
	M (SD)	*M* (SD)	*M* (SD)	*M* (SD)	*M* (SD)	*M* (SD)	*M* (SD)
Bot (*n*=1,250,401)	-0.09 (0.99)	0.00 (0.06)	0.15 (0.21)	0.01 (0.04)	0.46 (0.36)	0.16 (0.25)	0.10 (0.22)
Human (*n*=112,845)	-0.13 (0.99)	0.00 (0.05)	0.14 (0.21)	0.01 (0.08)	0.44 (0.37)	0.22 (0.29)	0.06 (0.15)
Bot + Human (*n*=1,363,246)	-0.09 (0.21)	-0.09 (0.99)	0.15 (0.21)	0.01 (0.04)	0.46 (0.36)	0.16 (0.26)	0.10 (0.22)
Range	-1 – 1	0 – 4	0 – 1	0 – 1	0 – 1	0 – 1	0 – 1

Figure 1

Visualization of Linguistic Feature Trends Over Time, Aggregated by Month, Comparing Bots and Humans



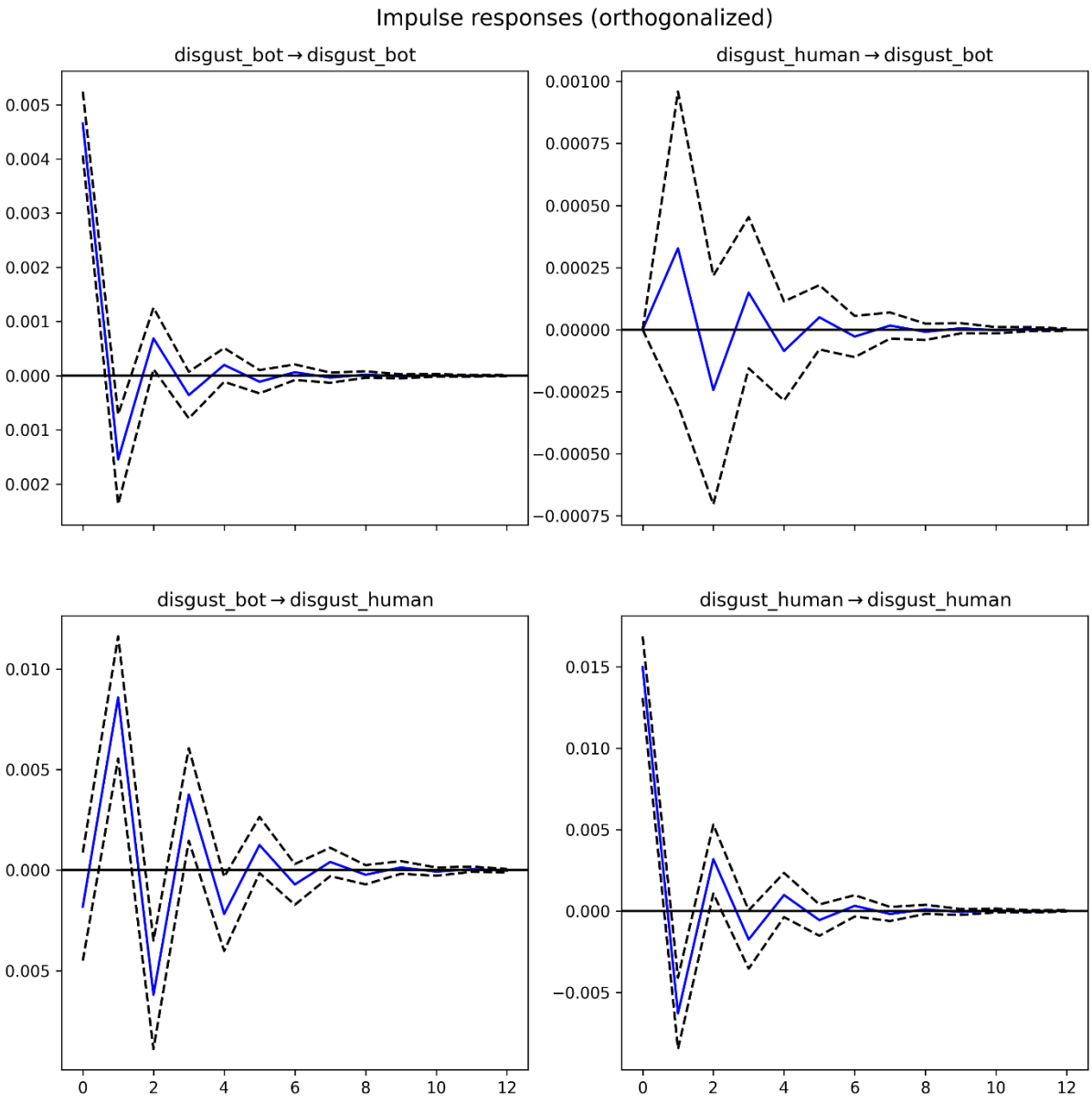
To evaluate the causal relationship posed by research questions (RQ1 through RQ3), vector autoregression (VAR) was used to run the granger causality tests as well as to generate impulse response functions (IRFs). Each linguistic score of bots and humans was aggregated by averaging the scores by month. Then the prediction of linguistic features from bots to humans and humans to bots were tested to infer a causal direction of the relationship. Granger causality tests specifically examine the *lagged impact* of each linguistic feature of one entity on another over time. Furthermore, IRFs are employed to measure the magnitude and significance of the temporal responses of the linguistic features of each entity on one another. While incorporating the longitudinal effect of variables in the social media environment, the delayed lag was selected

to 1. Such a decision was made given the short-lived discussion in social media (Zhang et al., 2023). Further, IRF was selected to 12 to predict the reactive dynamics of the effect of shocks over 12 months.

Figure 1 demonstrates how a monthly aggregated variable had changed over time. To specifically address the directional influence between bots and humans, Granger causality tests were conducted. When evaluating the directional relationship between politicization of bots and humans (RQ1), the results indicated that the politicization of bots Granger-caused the politicization of humans (p=0.03) (Table 5). Conversely, the influence from humans to bots was not statistically significant (p=1) (Table 5). In terms of polarization (RQ2), neither direction demonstrated statistical significance (bot to human: p=0.35; human to bot: p=1) (Table 5). Furthermore, with respect to the directional relationship of anger (RQ3a), the findings revealed that human anger Granger-caused bots anger (p=0.006) (Table 5), but not vice versa (p=1.00) (Table 5). In terms of fear (RQ3b), although bots influenced humans more than the reverse, the relationships were nonsignificant (bots to humans: p_{fear}=0.98; humans to bots: p_{fear}=0.12 (Table 5). For disgust (RQ3c), both directions were significant at the 0.05 significance level (bots to humans: p=0.00; humans to bots: p=0.01 (Table 5).

Figure 2

Impulse Response Functions (IRFs) of Disgust, Analyzing the Reciprocal Impact Between Bots and Humans



Further, IRFs were produced to confirm the treatment effect in future projection. The results showed that the increase in the lagged polarization and anger of bots decreased the level of those variables of humans, accordingly; yet these relationships were nonsignificant ($B_{polarization}$=-0.29, $p_{polarization}$=0.06; B_{anger}=-0.07, p_{anger}=0.58). Otherwise, the increase in the lagged politicization, fear, and disgust of bots increased the level of those corresponding variables of

humans; yet such relationships were also nonsignificant $B_{politicization}$=0.02, $p_{polarization}$=0.86; B_{fear}=0.02, p_{fear}=0.87) but disgust ($B_{disgust}$=1.68, $p_{disgust}$=0.00) (Figure 2).

The opposite direction (i.e., humans to bots) was also tested; the results showed that the increase in lagged politicization, polarization, and anger of humans decreased the level of those variables of bots, accordingly; yet these relationships were nonsignificant ($B_{politicization}$=-0.08, $p_{polarization}$=0.31; $B_{polarization}$=-0.03, $p_{polarization}$=0.52; B_{anger}=-0.05, p_{anger}=0.36). Otherwise, the increase in lagged fear and disgust of humans increased the level of those corresponding variables of bots; yet such relationships were also nonsignificant (B_{fear}=0.07, p_{fear}=0.24; $B_{disgust}$=0.02, $p_{disgust}$=0.31).

Table 5

Granger Causality Test (Based on VARs with Lag of 1)

Dependent Variable	Independent Variable	*p*-value
Bots → Humans		
Human Politicization	Bot Politicization	0.028*
Human Polarization	Bot Polarization	0.347
Human Anger	Bot Anger	0.264
Human Disgust	Bot Disgust	0.000***
Human Fear	Bot Fear	0.977
Humans → Bots		
Bot Politicization	Human Politicization	0.260
Bot Polarization	Human Polarization	0.351
Bot Anger	Human Anger	0.006**
Bot Disgust	Human Disgust	0.014*
Bot Fear	Human Fear	0.119

*p <.05, **p <.01, ***p <.001, two tailed

Study 1 Discussion

The present study examined the linguistic features of humans and bots and evaluated whether specific linguistic characteristics of one entity influenced those of the other. Previous studies have explored the provocative and often conflicting role of bots in online discussions, particularly during social and political events, where their presence can escalate conflicts and distort discourse. Despite differences in context, this study aimed to investigate the role of bots in comparison to humans in public discourse on Twitter (X) (the key results of Study 1 are summarized in Table 6).

First, the topical frames generated by humans and bots were compared to identify discrepancies in the content produced by each. The results revealed that bot-generated content tended to use more generic and promotional language than human-generated content. In contrast, human-generated content was more likely to reflect real-life events, such as those related to GMOs (e.g., FDA's approval of GMO salmon). Although bots also generated content referencing social events like Maui County's moratorium on genetically engineered crops and FDA's approval of GMO salmon, the majority of their content focused on promoting GMO-related products. Such results indicate the unique roles that bots and humans play in shaping public discourse. However, despite the fact that each entity generated distinct content, the presence of overlapping topics, specifically questioning GMOs, highlight their mutual interest that ispossibly shared in public discourse.

Next, specific linguistic features were targeted to examine whether bots expressed more politicization, polarization, and negative emotions such as anger, fear, and disgust. Additionally, the study investigated whether bots showed fewer neutral and positive emotions such as joy. Unexpectedly, humans' tweets were more likely to include polarized and politicized language.

However, negative emotions like anger, disgust, and fear were more prevalent in bot content compared to human content. Although joy was expected to be more prominent in human tweets, it was found to be higher in bot content, whereas neutrality was higher in human tweets. These findings confirm previous studies that have identified bots as highly emotionally charged (particularly with negative emotions), although they were found to express less politicized or polarized languages compared to humans. This discrepancy may be linked to the results on topical frames, which suggest that bots tend to generate more promotional and advertising-related content. Notably, however, bots were found to generate more angry, disgusted, and fearful content than humans, which may reflect the role of bots in amplifying perceived uncertainties, threats, or risks related to GMOs.

Further, the direct impact of bots on humans were assessed, particularly in terms of how bot content may affect human content. The results showed that politicization and disgust expressed by bots had longitudinal effects on human content creation. Notably, the disgust expressed by bots had significant long-term impacts on humans, suggesting a strong dynamic response, as indicated by the IRFs. This effect was more pronounced than that of the other variables, highlighting the potential role of bots in shaping human content over time.

Table 6

Summary of Results for Hypotheses and Research Questions in Study 1

Hypotheses and Research Questions	Summary of Results
H1: The topical themes of bots' tweets will differ from those of humans' tweets.	H1 is supported; Bots' tweets and humans' tweets exhibited distinct topical issues.
H2. Bots' tweets will be more likely than humans' tweets to exhibit 1) politicization, 2) polarization, and 3) negative emotions including a) anger, b) disgust, c) fear, while exhibiting less 4) neutrality and 5) positive emotion (i.e., joy).	H2-1 is not supported; Humans' tweets exhibited a greater level of politicization compared to bots' tweets. H2-2 is not supported; Humans' tweets exhibited a greater level of polarization compared to bots' tweets.H2-3 is supported. Bots' tweets showed a higher level of anger, disgust, and fear than humans' tweets. H2-4 is partially supported. Bots' tweets showed a lower level of neutrality but a higher level of joy than humans' tweets.
RQ1: Does the politicization of bots' tweets influence the politicization of humans' tweets?	A Granger causality test indicates that the politicization of bots' tweets influenced that of humans. However, the IRF results did not further support the directional influence.
RQ2: Does the polarization of bots' tweets influence the polarization of humans' tweets?	A Granger causality test indicates that the polarization of bots' tweets did not influence that of humans (or vice versa). The IRF results also did not support either directional influence.
RQ3: Do the negative emotions in bots' tweets— 1) anger, 2) disgust, and 3) fear—influence the emotions in humans' tweets?	RQ3-1: A Granger causality test indicates that the anger in humans' tweets influenced bots' tweets. However, the IRF results did not further support this directional influence. RQ3-2: A Granger causality test indicates that the disgust in bots' tweets influenced humans' tweets. The IRF results also supported this directional influence. RQ3-3: A Granger causality test indicates that the fear in bots' tweets does not influence humans' tweets or vice versa. Additionally, the IRF results did not further support this directional influence.

STUDY 2

Study 2 Research Design

Stimuli

To evaluate the impact of LLM-powered bots, GPT-4 was used to create a bot content with five tweets. In parallel, to address the growing trend of humans using LLMs to generate content, comparable LLM-generated human content was also created. Additionally, tweets from bot and human accounts identified in Study 1 were compiled, with five tweets selected from each category (human or bot). In each stimulus, five tweets were presented. A compilation of tweets was provided as a reference to facilitate the content creation process, using the few-shot prompting technique. Importantly, the length of tweets was limited to 280 characters, in accordance with X guidelines. Additionally, the prompt instructed GPT-4 to include URL to verify its ability to reference an external source. However, due to potential confounding effects, the URLs[2] were kept consistent across conditions. Details of the instructions are included in Appendix C, and the stimuli are provided in Appendix A.

Research Procedure

An online experiment was conducted using personas developed from data provided by McFadden et al. (2024). This approach, validated by Hewitt et al. (2024), demonstrated a strong correlation with actual treatment effects. While Hewitt et al. (2024) utilized GPT-4 in their experiments, the present study employed Llama-3. This decision was made because GPT-4 had already been used for content creation, potentially leading to data contamination through

[2] This study chose to use repeated URL links across conditions for two reasons. First, the links were not interactively used by participants, as they were provided to the LLMs. Second, when creating the stimuli using GPT-4, it was found that GPT-4 was capable of incorporating URL links, even when relevant to the context. To maintain uniformity, this study decided to embed the same URL links in each tweet, using them consistently across all conditions.

memorization (e.g., Li et al., 2024; Jiang et al., 2024). Initially, participant personas were created using Llama-3, based on demographic data from McFadden et al. (2024). The dataset, provided by the first author of the paper, included demographic information such as race, age, education, gender, partisanship, income, location, region, and previous experience in fields related to food, agriculture, health, or medicine. In alignment with the method implemented by Hewitt et al. (2024) to stimulate persona, a list of demographic information was provided to Llama-3 which used it to construct a persona for each participant that would then respond to the questions after being exposed to a stimulus. (an actual script is available in Appendix D). The created personas were randomly assigned to one of the experimental conditions, where they were exposed to a stimulus and asked to indicate their stance on the issue presented.

Participants

As described earlier, the present study generated personas based on previous research. This section outlines the demographics of participants from McFadden et al. (2024), which the present study utilized for Llama-3. In total, 3,125 participants responded to the survey. More than half of participants identified as female (55%). The average age was 44.39 (*SD*=19.31). Of all the participants 71% were White, followed by Black (14%) and Hispanic (7%); additionally, 36% of them reported to be identified as Democrats and 28% as Republicans (see Table 8 for details).

Measurements

The conceptual and operational definitions for each variable are summarized in Table 1.

Bot-likeliness. Participants will be asked to measure the perceived bot-likeness of the message on a 0-100 scale (0 = *not at all* to 100 = *very much*).

Human-likeness. Participants will be asked to measure the perceived human-likeness of the message on a 0-100 scale (0 = *not at all* to 100 = *very much*).

Attitudinal Change. The same question used by McFadden et al. (2024) was employed to evaluate participants' stance on the safety of gene editing in the context of food and agriculture. The question asked, "What is your opinion about the safety of gene editing in the context of food and agriculture?" Participants responded using a 5-point Likert scale, ranging from "extremely unsafe" (1) to "extremely safe" (5).

Participants' initial survey answers were subtracted from their responses from the present experiment to assess any changes in their attitudes. The higher scores indicate the extent to which participants changed their response in accordance with the given message.

Study 2 Results

In the experiment, personas were generated using demographic information from McFadden et al. (2024) with Llama-3, and these personas responded to the questions. In total, 3,125 personas were created, and corresponding responses were recorded. Occasionally, Llama-3 did not generate responses for specific questions; in these cases, available responses were considered and included in the analysis. To address the research questions, the study employed planned contrast analysis, which enabled comparisons between specific conditions. As the research questions required comparisons between two conditions, contrast weights of -1 and 1 were assigned to the selected conditions, with a weight of 0 assigned to the remaining conditions. The details are included in Table 9.

RQ4 examines the impact of LLM-generated content compared to non-LLM-generated content on the extent to which people perceive the content as being created by bots or humans. When comparing the LLM-generated bot condition to the human condition (RQ4-1a),

participants perceived LLM-generated bot content as more bot-like than human content ($t(1190.53) = 16.57$, $p = 0.00$). Conversely, when comparing the LLM-generated bot condition with the human condition (RQ4-1b), participants in the LLM-generated bot condition perceived the content as more human-like than in the human condition ($t(1293.04) = 2.71$, $p = 0.01$).

Next, the LLM-generated bot condition was compared to the bot condition (RQ4-2a), revealing that participants in the LLM-generated bot condition reported higher levels of bot-likeness compared to those in the bot condition, though the effect was not significant ($t(1361.23) = 1.77$, $p = 0.08$). When comparing the LLM-generated bot condition with the bot condition (RQ4-2b), the results showed that the bot condition had a higher mean score for human-likeness compared to the LLM-generated bot condition ($t(1530.37) = -20.39$, $p = 0.00$).

The comparison of bot-likeness between the LLM-generated human condition and the human condition (RQ4-3a) revealed that the human condition showed a higher level of bot-likeness compared to the LLM-generated human condition ($t(1005.08) = -2.02$, $p = 0.04$). However, when comparing the LLM-generated human condition with the human condition (RQ4-3b), the mean score for human-likeness in the LLM-generated human condition was significantly higher than in the human condition ($t(1432.47) = 66.23$, $p = 0.00$).

Lastly, when comparing the LLM-generated human condition with the bot condition (RQ4-4a), participants in the bot condition perceived the content as more bot-like than those in the LLM-generated human condition ($t(1425.42) = -30.05$, $p = 0.00$). When comparing perceived human-likeness between the LLM-generated human condition and the bot condition (RQ4-4b), participants in the LLM-generated human condition perceived the content as more human-like compared to those in the bot condition ($t(1543.08) = 31.30$, $p = 0.00$).

RQ5a examines whether there is a discrepancy in the perception of bot-likeness between LLM-generated human content and LLM-generated bot content. The results indicated that participants in the LLM-generated bot condition perceived the content as more bot-like compared to those in the LLM-generated human condition ($t(1312.16) = 28.01$, $p = 0.001$). RQ5b explores whether participants perceive the human-likeness of LLM-generated human content differently from LLM-generated bot content. The results showed that LLM-generated human content was perceived as more human-like compared to LLM-generated bot content ($t(1499.02) = -50.23$, $p = 0.00$).

RQ6a investigates whether there is a perceived difference between the bot condition and the human condition. The results indicated that the bot condition was perceived as more bot-like compared to the human condition ($t(1048.14) = 16.23$, $p = 0.001$). RQ6b examines whether there is a difference in the perception of human-likeness between the bot condition and the human condition. The results showed that participants in the bot condition perceived the content as more human-like compared to those in the human condition ($t(1385.17) = 28.19$, $p = 0.00$).

RQ7 examines the comparative effects of content generated by LLM-generated bots versus non-LLM-generated content on attitude change regarding gene editing in food and agriculture. First, the LLM-generated bot condition was compared with the human condition (RQ7-1). The results indicated that the human condition led to a greater attitude change compared to the LLM-generated bot condition ($t(1530.56) = -4.11$, $p = 0.001$).

Next, the LLM-generated bot condition was compared with the bot condition (RQ7-2). The results showed a negligible difference in attitude change between the two conditions, which was not significant ($t(1558.51) = 0.04$, $p = 0.97$).

The attitude change induced by the LLM-generated human condition was compared with the human condition (RQ7-3). The results showed a slightly higher attitude change for the human condition compared to the LLM-generated condition, but the difference was not significant ($t(1491.15) = -1.54$, $p = 0.12$).

The LLM-generated human condition was compared with the bot condition (RQ7-4), and the results showed that LLM-generated human content induced a greater attitude change compared to bot content ($t(1541.88) = 2.93$, $p = 0.00$).

RQ8 explores the comparative impact of LLM-generated bot content versus LLM-generated human content. The results indicated that LLM-generated human content led to a greater attitude change among participants compared to LLM-generated bot content ($t(1550.63) = -2.93$, $p = 0.00$).

Lastly, RQ9 examines the difference in attitude change between the human condition and the bot condition. The results showed that participants in the human condition experienced a greater attitude change compared to those in the bot condition ($t(1542.14) = -4.10$, $p = 0.00$).

Study 2 Discussion

Expanding on Study 1, which identified the influence of bots on humans on Twitter, Study 2 primarily focused on examining individuals' ability to discern LLM-generated content from non-LLM-generated content, as well as the extent to which they are influenced by LLM-generated content compared to non-LLM-generated content (the key results of Study 2 are summarized in Table 7).

An interesting finding emerged regarding LLM-generated human content, which was perceived as less likely to be created by bots compared to bot content but more likely to be created by bots compared to human content; LLM-generated human content was, however, more

likely to be created by humans compared to both human content and bot content. Additionally, LLM-generated bot content was more likely to be recognized as bot-generated compared to human content, but was equivalently perceived as being bot-generated compared to bot content; LLM-generated bot content was more likely to be perceived as human-generated compared to human content, but less likely to be perceived as human-generated compared to bot content. These findings suggest that LLM-generated human content may be perceived as more human-like and less bot-like compared to non-LLM-generated content, potentially deceiving individuals and influencing their perceptions, thereby affecting attitude change. Yet, people were generally not equipped to accurately identify the source of the content. In terms of attitude change, LLM-generated bot content had an impact as equivalent as bot content, and its impact was significantly smaller compared to human content. However, although LLM-generated human content was not as impactful as human content, it had a greater impact on attitude change compared to bot content. The results also indicate the LLM-generated human content could affect people's attitude as much as human content.

Additionally, this study evaluated how effectively individuals distinguished between LLM-generated bot content and LLM-generated human content, as well as the impact of each type of content on attitude change. Interestingly, participants were more likely to perceive LLM-generated bot content as being created by bots, and less likely to perceive it as human-generated, compared to LLM-generated human content. Moreover, participants' attitudes were more strongly swayed by LLM-generated human content than by LLM-generated bot content. These findings underscore the potential risk of LLM-generated human content, which may deceive individuals more effectively than LLM-generated bot content.

Finally, the present study compared bot content and human content to assess how well individuals can discriminate between the two types, as well as how each type influences attitude change. Bot content was more likely to be perceived as bot-generated compared to human content, but it was still seen as more human-like than human content. Despite this perceptual ambiguity, human content led to a greater attitude change than bot content.

Table 7

Summary of Results for Hypotheses and Research Questions in Study 2

Research questions	Summary of results
RQ4-1: Do people perceive human content and LLM-generated bot content differently in terms of (a) bot-likeness and (b) human-likeness?	LLM-generated bot content was perceived as more likely to be created by bots than human content. However, LLM-generated bot content was more likely to be perceived as created by humans compared to human content.
RQ4-2: Do people perceive bot content and LLM-generated bot content differently in terms of (a) bot-likeness and (b) human-likeness?	Bot content was not perceived differently from LLM-generated bot content in terms of bot-likeness. However, bot content was more likely to be perceived as created by humans compared to LLM-generated bot content.
RQ4-3: Do people perceive human content and LLM-generated human content differently in terms of (a) bot-likeness and (b) human-likeness?	LLM-generated human content was perceived as more likely to be created by bots than human content. However, LLM-generated human content was perceived as more likely to be created by humans than human content.
RQ4-4: Do people perceive bot content and LLM-generated human content differently in terms of (a) bot-likeness and (b) human-likeness?	Bot content was perceived as more likely to be created by bots than LLM-generated human content. Additionally, LLM-generated human content was perceived as more likely to be created by humans than bot content.
RQ5: Do people perceive LLM-generated bot content and LLM-generated human content differently in terms of (a) bot-likeness and (b) human-likeness?	LLM-generated bot content was perceived as more likely to be created by bots than LLM-generated human content. Additionally, LLM-generated human content was perceived as more likely to be created by humans than LLM-generated bot content.

Table 7 (cont'd)

RQ6: Do people perceive bot content and human content differently in terms of (a) bot-likeness and (b) human-likeness?	Bot content was perceived as more likely to be created by bots than human content. However, bot content was perceived as more likely to be created by humans than human content.
RQ7-1: Do LLM-generated bot content and human content affect individuals' attitudes differently?	Human content led to a greater level of attitude change compared to LLM-generated bot.
RQ7-2: Do LLM-generated bot content and bot content affect individuals' attitudes differently?	There was no difference in attitude change between LLM-generated bot content and bot content.
RQ7-3: Do LLM-generated human content and human content affect individuals' attitudes differently?	There was no difference in attitude change between LLM-generated human content and human content.
RQ7-4: Do LLM-generated human content and bot content affect individuals' attitudes differently?	LLM-generated human content led to a greater level of attitude change compared to bot content.
RQ8: Do LLM-generated human content and LLM-generated bot content affect individuals' attitudes differently?	LLM-generated human content led to a greater level of attitude change compared to LLM-generated bot content.
RQ9: Do human content and bot content affect individuals' attitudes differently?	Human content led to a greater level of attitude change compared to bot content.

Table 8.
Sample Demographics in Study 2 and Study 3

	Study 2 *M* (SD) or *n* (%)	Study 3 *M* (SD) or *n* (%)
Age	44.39 (19.31)	45.24 (16.62)
Gender		
Female	1713 (54.82)	701 (49.09)
Male	1412 (45.18)	682 (47.76)
Other		45 (3.15)
Race		
White	2227 (71.26)	939 (65.76)
Black	446 (14.27)	169 (11.83)
Hispanic	219 (7.01)	137 (9.59)
Non-Black /Non-White/ Non-Hispanic	233 (7.46)	183 (12.82)
Partisanship		
Democrats	1126 (36.03)	664 (46.50)
Republicans	864 (27.58)	302 (21.15)
Non-Republican/ Non-Democrat	1137 (36.38)	462 (32.35)
Education		
Less than high school degree	118 (3.78)	3 (0.21)
High school degree (high school diploma or equivalent including GED)	787 (25.18)	171 (11.97)
Some college but no degree	892 (28.54)	298 (20.87)
Associate's degree in college	344 (11.01)	168 (11.76)
Bachelor's degree in college	638 (20.42)	537 (37.61)
Graduate or Professional degree (MS, PhD, JD, MD)	346 (11.07)	251 (17.58)

Table 8 (cont'd)

Income		
Less than $10,000	217 (6.94)	68 (4.76)
$20,000 - $29,999	354 (11.33)	103 (7.21)
$30,000 - $39,999	332 (10.62)	129 (9.03)
$40,000 - $49,999	263 (8.42)	119 (8.33)
$50,000 - $59,999	335 (10.72)	144 (10.08)
$60,000 - $69,999	218 (6.98)	109 (7.63)
$70,000 - $79,999	274 (8.77)	118 (8.26)
$80,000 - $89,999	146 (4.67)	82 (5.74)
$90,000 - $99,999	174 (5.57)	91 (6.37)
$100,000 - $149,999	388 (12.42)	232 (16.25)
Area		
Suburban	1570 (50.24)	754 (52.80)
Urban	842 (26.94)	426 (29.83)
Rural	713 (22.82)	248 (17.37)
Region		
South (Delaware, Maryland, Washington DC, Virginia, West Virginia, Kentucky, North Carolina, South Carolina, Tennessee, Georgia, Florida, Alabama, Mississippi, Arkansas, Louisiana, Texas, Oklahoma)	1339 (42.85)	545 (38.17)
Midwest (Ohio, Michigan, Indiana, Wisconsin, Illinois, Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebraska, Kansas)	668 (21.38)	273 (19.12)
Northeast (Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania)	575 (18.40)	246 (17.23)
West (Montana, Idaho, Wyoming, Colorado, New Mexico, Arizona, Utah, Nevada, California, Oregon, Washington, Alaska, and Hawaii)	543 (17.38)	364 (25.49)

Table 9

Planned Contrast Analysis Weights, Condition Means (and Standard Deviations)

	LLM-generated bot content	LLM-generated human content	Bot	Human
Contrast weight				
Contrast 1	1	0	0	-1
Contrast 2	1	0	-1	0
Contrast 3	0	1	0	-1
Contrast 4	0	1	-1	0
Contrast 5	1	-1	0	0
Contrast 6	0	0	1	-1
Study 2				
M (*SD*)				
Bot-likeness	61.45 (25.24)	28.78 (18.41)	59.32 (20.08)	32.11 (38.59)
Human-likeness	13.49 (24.25)	70.31 (20.13)	37.21 (21.52)	10.72 (14.94)
Attitude change	0.55 (1.15)	0.71 (1.07)	0.55 (1.19)	0.81 (1.33)
n	781	781	781	782
Study 3				
M (*SD*)				
Bot-likeness	48.80 (27.31)	47.64 (27.03)	48.82 (26.54)	48.28 (28.05)
Human-likeness	58.42 (26.55)	57.93 (26.26)	57.01 (25.91)	57.46 (27.26)
Attitude change[a]	0.19 (0.78)	0.20 (0.62)	0.20 (0.67)	0.21 (0.73)
n	377	365	349	337

[a]The impact of attitude change for Study 3 was tested using multinomial logistic regression.

STUDY 3

Study 3 Research Design

Study 3 aimed to replicate Study 2 with human participants in order to verify the findings generated by Llama-3 in Study 2. Although other LLMs (e.g., GPT-4) have produced rigorous results in previous research (Hewitt et al., 2024), Llama-3 had not been actively employed in prior studies. In Study 3, by recruiting human participants, we were able to both confirm the effectiveness of the results generated by Llama-3 and replicate the findings from Study 2.

Stimuli

In Study 2, only textual information was used as input for the LLMs. However, for Study 3, which involved human participants, the stimuli were designed as tweet-formatted images. While creating realistic tweet stimuli, additional external cues were incorporated. To control for confounding factors, these cues—such as username, user handle, profile image, post date and time, number of likes, views, retweets, quotes, and saves—were kept constant across conditions (see Figure S-4).

Research Procedure

The research procedure in Study 3 was slightly modified to accommodate actual human participants. Participants were recruited via the online platform Prolific, and upon completion of the study, each participant received a $1 compensation. The study was advertised on Prolific, and individuals interested in participating voluntarily joined. To minimize bias, the true aim of the study was concealed at the outset in order to assess the genuine impact of LLM-generated and non-LLM-generated messages.

After consenting to participate, participants were first asked to complete a set of questions designed to assess their pre-existing stance on gene editing in the context of food and

agriculture. These questions were mixed with bogus items to prevent participants from identifying the study's true purpose. Participants were then randomly assigned to one of the experimental conditions and asked to respond to the same attitude questions about gene editing, as well as questions regarding bot-likeness and human-likeness. Finally, participants completed demographic questions before the experiment concluded. At the end of the study, a debriefing message was provided to explain the true purpose of the experiment.

Participants

As mentioned, participants were recruited via Prolific. An initial power analysis result pointed to having 1,424 samples based on the small effect size of AI on attitudes reported by Huang and Wang. (2023), but I oversampled in a case of the poor quality of data. After removing responses that did not complete (at least) 60% of questions and did not pass an attention check question, in total, 1,428 responses were left. Less than half of participants identified as female (49%). The average age was 45.24 (*SD*=16.62). Of all the participants, 66% were White, followed by Black (12%) and Hispanic (10%); additionally, 46% of them reported to be identified as Democrats and 21% as Republicans (see Table 8 for details).

Measurements

The conceptual and operational definitions for each variable are summarized in Table 1.

Bot-likeness. The same item was used in Study 2.

Human-likeness. The same item was used in Study 2

Attitude Change. The same measure used in Study 2 was applied in Study 3 to calculate attitude change. The key difference, however, was that in Study 3, attitudes toward gene editing in the field of agriculture were measured both before and after exposure to the stimuli, as there was no pre-measurement of attitudes in advance.

Table 10

Multinomial Logistic Regression Model of Attitude Changes after Exposure to the Stimuli

Independent variables	Dependent variable category	
	Negative change (v. no change)	Positive change (v.no change)
	b (SE)	*b (SE)*
(Reference group: Human content)		
LLM-generated bot content	0.18 (0.33)	0.11 (0.19)
LLM-generated human content	0.22 (0.33)	0.16 (0.19)
Bot content	-0.08 (0.36)	0.11 (0.19)
AIC	2046.97	

$^{\dagger}p < .1$, $*p < .05$, $** p < .01$, $***p < .001$, two-tailed.

Study 3 Results

In reporting the results for Study 3, a planned contrast analysis was conducted to address RQs 4 through 6, following the same methodology used in Study 2.

RQ4 explores the difference in bot-likeness and human-likeness perceptions between the LLM-generated conditions and non-LLM-generated conditions. The difference in perceived bot-likeness was examined between the LLM-generated bot content and the human content (RQ4-1a). The results showed a minimal difference between these conditions ($t(1420) = 0.72$, $p = 0.47$). The LLM-generated bot condition was compared with the human condition to test the difference in perceived human-likeness (RQ4-1b), and the results were non-significant ($t(1419) = 1.37$, $p = 0.17$).

The LLM-generated bot condition was also compared with the bot condition (RQ4-2a), and the results revealed no significant differences between the conditions ($t(1420) = -0.02$, $p = 0.98$). However, when comparing the LLM-generated bot condition to the bot condition in terms of perceived human-likeness (RQ4-2b), the results showed that the LLM-generated bot content was perceived as more human-like compared to the bot content ($t(1419) = 2.02$, $p = 0.04$).

When comparing the LLM-generated human condition with the human condition (RQ4-3a), participants perceived them almost equally as bot-like ($t(1420) = -0.89$, $p = 0.38$). Further, the difference in perceived human-likeness between the LLM-generated human condition and the human condition was minimal (RQ4-3b) ($t(1419) = 0.68$, $p = 0.50$).

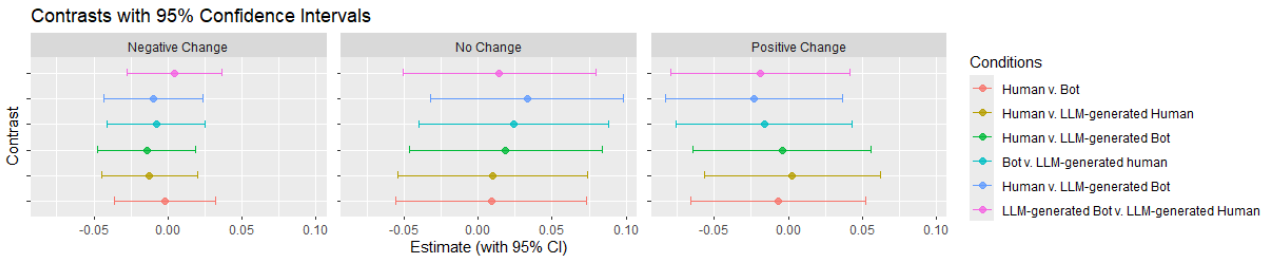
Regarding the comparison between the LLM-generated human condition and the bot condition (RQ4-4a), only a negligible shift was found in perceived bot-likeness ($t(1420) = -1.63$, $p = 0.47$), and when comparing the LLM-generated human condition with the bot condition for human-likeness (RQ4-4b), the results were also non-significant ($t(1419) = 1.32$, $p = 0.19$).

RQ5a explores the difference in perceived bot-likeness between the LLM-generated bot condition and the LLM-generated human condition, with the result showing no significant difference ($t(1420) = 1.61$, $p = 0.11$). RQ5b further examines the perceived human-likeness of the LLM-generated human condition versus the LLM-generated bot condition, revealing no significant difference ($t(1419) = 0.69$, $p = 0.49$).

RQ6a investigates whether there is a discrepancy in perceiving the content as bots, specifically between the bot condition and the human condition, but no significant difference was observed ($t(1420) = 0.74$, $p = 0.46$). RQ6b examines the difference in perceived human-likeness between the human condition and the bot condition, with no significant difference found ($t(1419) = -0.64$, $p = 0.52$).

Figure 3

The Contrasted Estimated Marginal Means for Each Comparison of the Experimental Groups



For the analysis of attitude change (RQ7-RQ9), a multinomial regression was conducted, as the dependent variable was not normally distributed. Approximately 74% of participants showed no change in attitude, 5% moved in the opposite direction of the content, and 21% shifted their attitude in the direction suggested by the content. Multinomial regression was applied to compare the conditions where participants shifted their attitudes (either positively or negatively) versus those who showed no change. The “no change” group was used as the reference to compare “positive change” (i.e., change in alignment with the content) and “negative change" (i.e., change in opposition to the content). The results of the multinomial

regression are presented in Table 10. Additionally, to compare the impact of each experimental condition with one another, the *emmeans* package was used. This package allows for contrasting the estimated marginal means of each experimental condition across groups, where the attitudes of the groups were changed either positively, negatively, or not at all. Figure 3 displays the contrasted estimated marginal means for each comparison of the experimental groups.

RQ7 evaluated attitude change induced by the LLM-generated conditions versus the non-LLM-generated condition. The LLM-generated bot condition was compared with the human condition (RQ7-1), and the results showed no significant difference in all three categories of attitude change (positive change: estimate =-0.02 SE=0.03, $p = 0.99$; negative change: estimate = -0.01, SE=0.02, $p = 0.99$; no change: estimate = 0.02, SE=0.03, $p = 0.99$).

The impact of the LLM-generated bot condition was compared with the bot condition (RQ7-2), and the results showed no significant effect on attitude change for either the positive, negative change, no change groups (positive change: estimate =0.003 SE=0.03, $p = 1.00$; negative change: estimate = -0.01, SE=0.02, $p = 0.99$; no change: estimate = 0.01, SE=0.03, $p = 1.00$).

When the LLM-generated human condition was compared with the human condition (RQ7-3), there was no significant difference in attitude change for either the positive, negative, or no change groups (positive change: estimate =0.003 SE=0.03, $p = 1.00$; negative change: estimate = -0.01, SE=0.02, $p = 0.99$; no change: estimate = 0.01, SE=0.03, $p = 1.00$).

The LLM-generated human condition was compared with the bot condition (RQ7-4), the effect was not statistically significant in none of the groups (positive change: estimate =-0.004 SE=0.03, $p = 1.00$; negative change: estimate = -0.01, SE=0.02, $p = 0.98$; no change: estimate = 0.02, SE=0.03, $p = 0.99$).

For RQ8, which examines the effect of the LLM-generated human condition versus the LLM-generated bot condition on attitude change, both conditions exhibited equivalent non-significant effects across groups (positive change: estimate =-0.01 SE=0.03, p = 1.00; negative change: estimate = -0.002, SE=0.02, p = 1.00; no change: estimate = 0.01, SE=0.03, p = 1.00).

Finally, for RQ9, which investigates the effect of the human condition versus the bot condition on attitude change, both conditions exhibited equivalent non-significant effects across groups (positive change: estimate =0.02 SE=0.03, p = 0.99; negative change: estimate = 0.005, SE=0.02, p = 1.00; no change: estimate = 0.01, SE=0.03, p = 1.00).

Study 3 Discussion

While targeting human participants, Study 3 aimed to replicate the findings from Study 2. Although most results were not fully reproduced, some aligned with those from the previous study. When participants were asked to evaluate whether the content they received appeared to be generated by humans or bots, they found no significant difference between LLM-generated content and non-LLM-generated content. However, there was one exception: when comparing LLM-generated bot content with bot content, participants reported that the LLM-generated bot content was more likely to be created by humans. Even though this finding was not reported in Study 2, the current result showed the difficulty of identifying the source of messages with the given content. Additionally, different types of content did not affect participants' attitudes, which did not replicate the results from Study 2, where LLM-generated human content and human content were found to cause a greater attitude change compared to other types of content. (the key results of Study 2 are summarized in Table 11).

Overall, the mean scores for perceptions of bot-likeness and human-likeness ranged narrowly from 47.64 to 48.82, indicating no significant differences across the conditions.

Similarly, the mean perception of human-likeness ranged from 57 to 58 across the conditions. This suggests that participants were generally unable to distinguish between LLM-generated and non-LLM-generated content, except when comparing LLM-generated bot content to bot content, where they perceived the LLM-generated bot content as more human-like.

When comparing the LLM-generated bot content to other LLM-generated content, no significant differences were observed in terms of perceived bot-likeness or human-likeness. This pattern was also consistent when comparing the bot content with the human content.

Regarding the impact of LLM-generated content on attitude change, no statistically significant differences were observed in any of the comparisons (non LLM-generated content). This is quite consistent with the results from Study 2, even though Study 2 reported the significant effect of attitude change of LLM-generated content over non LLM-generated content.

Finally, when comparing the effects of LLM-generated content to one another, the extent to which participants changed their attitudes was nearly identical regardless of the type of LLM-generated content they were exposed to. Similarly, no significant difference in attitude change was observed between the bot content and the human content. These results indicate that varied types of content did not necessarily lead to attitude change.

Table 11

Summary of Results for Hypotheses and Research Questions in Study 3

Research questions	Summary of results
RQ4-1: Do people perceive human content and LLM-simulated bot content differently in terms of (a) bot-likeness and (b) human-likeness?	Human content and LLM-generated bot content were not perceived differently in terms of bot-likeness and human-likeness.
RQ4-2: Do people perceive bot content and LLM-simulated bot content differently in terms of (a) bot-likeness and (b) human-likeness?	Bot content and LLM-generated bot content were not perceived differently in terms of bot-likeness. However, LLM-generated bot content was perceived as more likely to be created by humans than bot content.

Table 11 (cont'd)

RQ4-3: Do people perceive human content and LLM-generated human content differently in terms of (a) bot-likeness and (b) human-likeness?	Human content and LLM-generated human content were not perceived differently in terms of bot-likeness or human-likeness.
RQ4-4: Do people perceive bot content and LLM-generated human content differently in terms of (a) bot-likeness and (b) human-likeness?	Bot content and LLM-generated human content were not perceived differently in terms of bot-likeness or human-likeness.
RQ5: Do people perceive LLM-generated bot content and LLM-generated human content differently in terms of (a) bot-likeness and (b) human-likeness?	LLM-generated bot content and LLM-generated human content were not perceived differently in terms of bot-likeness or human-likeness.
RQ6: Do people perceive bot content and human content differently in terms of (a) bot-likeness and (b) human-likeness?	Bot content and human content were not perceived differently in terms of bot-likeness or human-likeness.
RQ7-1: Do LLM-generated bot content and human content affect individuals' attitudes differently?	There was no difference in attitude change between LLM-generated bot content and human content.
RQ7-2: Do LLM-generated bot content and bot content affect individuals' attitudes differently?	There was no difference in attitude change between LLM-generated bot content and bot content.
RQ7-3: Do LLM-generated human content and human content affect individuals' attitudes differently?	There was no difference in attitude change between LLM-generated human content and human content.
RQ7-4: Do LLM-generated human content and bot content affect individuals' attitudes differently?	There was no difference in attitude change between LLM-generated human content and bot content.
RQ8: Do LLM-generated human content and LLM-generated bot content affect individuals' attitudes differently?	There was no difference in attitude change between LLM-generated human content and LLM-generated bot content.
RQ9: Do human content and bot content affect individuals' attitudes differently?	There was no difference in attitude change between bot content and human content.

GENERAL DISCUSSION

This dissertation investigates the impact of bots on humans. In Study 1, social media data (tweets) were used to examine the linguistic features of both bots and humans, focusing on how these entities differ in their linguistic styles and whether bots' linguistic features influence human language use. First, the findings from topic modeling indicate that the tweets of humans and bots were distinct. However, there was an overlapping topic questioning GMOs, suggesting a shared interest between both bots and humans. The subsequent analysis was conducted to compare the difference of linguistic features between bots and humans, as well as to infer a causal relationship between bots and humans; the results showed that anger, disgust, fear, and joy were found to be more prevalent among bots compared to humans, while polarization, politicization, and neutrality were pronounced among humans compared to bots. Moreover, some features such as politicization and disgust directly influenced human linguistic behaviors.

To further investigate the influence of bots, Studies 2 and 3 explored recent advancements in technology, particularly LLMs, which enable bots to generate content that more closely resembles human-written text. This technological development creates opportunities for bots to be perceived as humans, potentially altering the information ecosystem on social media platforms. However, the extent to which people can identify whether a message is written by a human or a bot, and whether this distinction influences their attitudes, remains unclear. And studies 2 and 3 sought to investigate these effects. The results from Studies 2 and 3 consistently demonstrated that participants were unable to accurately distinguish the source of content. In Study 2, LLM-generated human content had a stronger impact on participants' attitudes compared to both LLM-generated bot content and bot content. Additionally, human content led to a greater attitude change than LLM-generated bot content and (non-LLM) bot content.

However, this effect was not observed in Study 3. The following sections break down these results and explore their broader implications.

In Study 1, an observational study was conducted to analyze the topics and linguistic features in tweets from both bots and humans. The results revealed distinct differences in the content produced by humans and bots. Humans focused more on real-life events, whereas bots tended to post promotional and generic content. However, there were some topics shared between bots and humans, which indicate a shared interest and possible further interactions. In terms of linguistic features, humans were more likely to use politicized, polarized, and neutral language, while bots displayed higher levels of anger, disgust, fear, and joy. This pattern aligns with previous literature, which indicates that bots use emotionally charged language, with negative emotions such as anger, disgust, and fear being more pronounced than humans (Bail et al., 2020; Badawy et al., 2018; Keller & Klinger, 2019; Shao et al., 2017). This finding emphasizes the malicious role of bots in insinuating negativity. Yet, humans still displayed more politicized, polarized, and neutral languages, which could be attributed to the generally low levels of these variables, specifically politicization averages of 0.002 for bots and 0.003 for humans. That being said, it is important to note that these differences could be attributed to large sample sizes and require caution for further interpretation.

Furthermore, Study 1 identified a direct influence between bots and humans in terms of language use, with bots' politicization and disgust impacting human discourse. This influence was evident over time, especially with regard to disgust, which was further reinforced by the IRFs. Although politicization was less pronounced among bots compared to humans, the politicized tweets of bots still had a significant influence on those of humans, presenting a potential risk for shaping human attitudes. Moreover, among the various negative emotions,

bots' disgust demonstrated a direct influence on that of humans. These findings align with previous literature that emphasizes the possible nefarious impact bots can have on human attitudes (Aldayel & Magdy, 2022; Bail et al., 2020), and they were confirmed longitudinally.

In Study 2, a controlled online experiment was conducted to assess the impact of bots on human attitudes, as well as the extent to which people could discern whether content was generated by bots or humans. Given the rise of LLMs, there was concern regarding the potential impact of LLM-generated content on users. As such, in addition to the human and bot content tested in Study 1, this experiment also included content generated by GPT-4. The experiment was conducted using Llama-3, with personas derived from demographic information provided by McFadden et al. (2024). The results were somewhat mixed when participants identified the source of the content. Yet, LLM-generated human content was seen as more human-like than other counterparts and perceived as less bot-like than LLM-generated bot content and bot content. These findings suggest that LLM-generated human content may have a greater impact on individuals than traditional bot content and even LLM-generated bot content. Nonetheless, as seen in other comparisons, it is likely that people overall were unable to identify the source of content, and such results are consistent with previous studies suggesting that people struggle to differentiate AI-generated content from human content (Clerwall, 2014; Graefe et al., 2018; Karinshak et al., 2023). Furthermore, when examining attitude change, LLM-generated human content also had a greater influence on attitude change compared to bot content and LLM-generated bot content. This finding implies that LLM generated content, specifically LLM-generated human content that mimics human language could have a more significant impact on individuals than traditional bot content (Burtell & Woodside, 2023).

Study 3 aimed to replicate Study 2 with real human participants. While the majority of significant findings from Study 2 did not replicate in Study 3, some trends were consistent with the earlier results. Overall, the variation in perceptions of bot-likeness and human-likeness was minimal, with means for bot-likeness ranging from 47.64 to 48.82 on a 0-100 scale, and human-likeness ranging from 57.01 to 58.42 on a 0-100 scale. The only notable difference was that LLM-generated bot content was more likely to be perceived as human-generated than bot content. These findings suggest that participants struggled to distinguish between content created by bots and humans in alignment with Study 2. Yet, in Study 2, participants misattributed the source of the content whereas in Study 3, there was a weak tendency to classify the content as either human-generated or bot-generated. This lack of differentiation poses a potential risk, as inaccurate perceptions of content origin could affect how users process and respond to the information.

Interestingly, while most participants did not change their attitudes in response to the stimuli, 21% (301 out of 1,428) of participants did shift their stance in alignment with the message. Although the effect was not statistically significant, this finding suggests that certain individuals may still be influenced by LLM-generated content as well as non-LLM generated content. However, it is inconclusive since the non-significant results in Study 3 can be attributed to confounding variables that could not be controlled. Hence, it is important to be cautious when interpreting these results. Additionally, it is possible that the effects were attenuated since the study's design captured participants' attitudes immediately after exposure. The future directions that could address these limitations are presented in the following section.

Table 12.
A Summary of Results

Summary
Study 1 demonstrated the role of bots in public discourse on Twitter (X) regarding GMOs, in contrast to humans. Results showed that bots' posts were more generic and commercial compared to those of humans. In terms of linguistic features, bots elicited more anger and fear than humans whereas politicization and polarization were more pronounced among humans. When analyzing the direct impact of bots' linguistic features on humans, politicization in bots was found to directly influence politicization in humans, and disgust in bots also directly influenced disgust in humans. Study 2 and Study 3 were conducted to further explore the causal influence of bots. Both studies incorporated the latest technology, specifically LLMs, which potentially enhance the impact of bots. Online experiments were conducted to compare the impact of LLM-empowered content versus non-LLM-empowered content on individuals' attitudes as well as their ability to identify the source of the content. While Study 2 found that participants perceived LLM-empowered human content as more human-generated and less bot-generated than the other content types, it also found that LLM-empowered human content was more effective at changing attitudes compared to traditional bot content and LLM-generated bot content. In contrast, Study 3 found that all content influenced attitudes to a similar degree, and participants had difficulty identifying the correct source.

Overall, the results are in alignment with the CASA theory (Reeves & Nass, 1996), as computers, specifically AI-enhanced bots in the present context, may induce social responses from users who mindlessly apply social rules while interacting with them. Although attitude changes prompted by the contents were inconsistent across the present studies, it is notable that participants were not proficient at identifying the correct source of the content, which implies the potential influence of bots on social media platforms. The CASA theory has been applied to numerous technologies, and recent development in artificial intelligence—specifically those that possess language capabilities, facilitate interactivity, fill social roles, and operate with human-sounding speech (Nass & Steuer, 1993)—could make AI-enhanced bots a powerful source of influence, promoting human-machine communication that closely resembles human-human communication. As more people adopt this new technology, it is imperative to continue studies

exploring its potential benefits and risks for individuals and society. These efforts could help raise awareness among individuals and inform policymakers and governments to enact proper interventions and/or regulations regarding AI use particularly on social media.

CONCLUSION

This dissertation explores the impact of bots on online public discourse, specifically focusing on human users' language and attitudes in response to their possible online interactions. The findings suggest that bots and humans exhibited distinct topics, though there was some overlap regarding the reliability of GMO technology. A similar topic shared between bots and humans indicated common interests and the potential for further interactions. Moreover, bots were more likely to display negatively charged emotions in their tweets, and specific features, such as politicization and disgust, influenced the linguistic characteristics of humans' tweets, particularly over time. This highlights the importance of user vigilance, as well as governmental interventions and/or regulations, in mitigating the potential disruptions bots can cause in public discourse and the influence they may exert on humans' opinions. As the present research implies the potential risk posed by bots, it is imperative to conduct more research to explore the effective and legitimate forms of interventions that could mitigate the harm. Importantly, the present results align with prior research that primarily focused on politically charged topics, demonstrating that the presence of bots can affect public discourse on social media platforms, regardless of the topic.

Furthermore, the comparison between LLM-generated content and non-LLM-generated content demonstrates the potential for LLM-powered bots to deceptively influence users, as people are often unable to distinguish between human-like and machine-generated content on social media. As bots become more sophisticated with advances in technology, they are increasingly capable of influencing human attitudes in ways that are almost indistinguishable from human interaction. Given the proliferation of such technologies on social media,

understanding their relational impact on humans becomes even more critical (Guzman & Lewis, 2019).

Limitations

This research is not without its limitations. Despite efforts to design the studies rigorously, several trade-offs were made to balance cost and efficiency. First, only a selected number of users were included in Study 1. This decision was made to reduce the computational costs associated with bot detection and subsequent analyses. However, the users were randomly selected based on their level of activity (e.g., number of posts), and tweets were retrieved accordingly. While an effort was made to use a representative sample, it will be important to incorporate a full dataset in future research.

Second, pretrained models used to assess polarization and detect discrete emotions may introduce limitations in interpreting results. Although these models are initially fine-tuned through supervised learning to capture specific concepts, transitioning to unsupervised learning can lead to unintended biases or deviations from the original task, potentially distorting the intended interpretation. In contrast, a dictionary-based approach was employed to analyze politicization, which can enhance interpretability by focusing on specific, predefined terms. However, this approach struggles to capture the contextual nuances and subtleties of meaning that pretrained models are better equipped to handle. As such, future researchers should be aware of the inherent limitations of both approaches.

Regarding the online experiments, the third limitation is the stimuli used cannot be claimed to have been perfectly designed. The classification of messages as bot-generated or human-generated, as determined by human coders and a bot detection tool, showed only limited consensus. As a result, only those messages for which there was agreement between at least two

human coders and the bot detection tool were included in the analysis. As outlined in Appendix B, achieving agreement among human coders in bot detection proved to be a challenging task. Nonetheless, the present study aimed to include only messages that had reached a reasonable level of consensus.

Fourth, the use of Llama-3 to generate personas and conduct experiments may have influenced the findings, given that the method has not been widely validated in prior research. In contrast, Hewitt et al. (2024) used GPT-4 and found a high correlation between actual results and those derived from LLM-simulated personas. Although their study did not employ Llama-3, the state-of-the-art performance of Llama-3 across various fields (Dubey et al., 2024) supports its inclusion in the current research, despite the limited validation of the method.

Fifth, the non-significant results of Study 3 could be attributed to confounding factors. Despite efforts to control the cues provided in the stimuli, some of them may have been perceived as specific indicators for identifying content as either human or bot-generated. The present research focused primarily on the content, but it is important to recognize that source cues might have influenced participants, leading them to heuristically perceive and process the message. Therefore, it is crucial to specifically control for such variables and investigate their effects.

Lastly, Study 3 was unable to replicate the results from Study 2. The primary difference between the two studies was the use of human participants in Study 3, but it is important to note that the observed differences could also be attributed to the short time lag between the pre and post-attitude measures. In Study 2, pre-existing attitudes were drawn from McFadden et al. (2024), whereas in Study 3, participants' existing attitudes were measured immediately prior to exposure to the stimulus. Given the findings of previous persuasion studies, it is difficult to

change one's attitude in a short period, and the brief interval between the pre and post-exposure measurements in Study 3 may have influenced the results.

Future Directions

The present research project demonstrated the impact of bots on humans across three separate studies. Several key findings warrant further exploration. As noted in the limitations, although the current study selected users from a dataset, future research could expand by including the entire dataset, incorporating all users who participated in the discussions. Although the current findings were derived from the representative samples of long-term public discourse surrounding GMOs, by analyzing the whole dataset, a more comprehensive understanding of the influence of bots on humans would be provided.

Furthermore, as noted earlier, unsupervised and supervised methods can yield different results. While recognizing the trade-offs associated with each approach, it is important to attempt replication of the results using alternative methods to verify their robustness. For example, comparing results from both dictionary-based and pre-trained models can help verify their consistency. In social science research, supervised methods are often preferred for their ability to maintain the interpretability of results. However, combining different methods can contribute to methodological advancements and provide a more comprehensive understanding.

As discussed earlier, the current research utilized specific language models for distinct purposes, such as GPT-4 for content creation and Llama-3 for the online experiment. This decision was made to avoid data contamination resulting from overreliance on a single model. However, this choice limits the generalizability of the findings. In future research, different language models should be tested to better understand the precise impact of LLMs on the influence of bots and shaping individuals' attitudes. Additionally, employing various LLMs in

online experiments could further expand our understanding of their potential in social science research. Moreover, regarding the results from Study 2 and Study 3, further investigation is needed to understand how individuals' perceptions of the source of content affect their evaluation of messages and attitudes. This point is exemplified by studies conducted by Wischnewski et al. (2021) and Wischnewski et al. (2024). The present studies highlight the challenge of detecting bots, especially as technological advancements make identification even more difficult. The issue arises when people mistakenly perceive users as bots, and the potential impact of such misperceptions. Wischnewski et al. (2021) found that people are more likely to assume users with opinion-incongruent posts are bots, a tendency mediated by perceived credibility. In other words, individuals who identify posts that align with their own opinions are more likely to evaluate those posts as credible and attribute them to humans rather than bots. While the experimental conditions in this study compared opinion-congruent and opinion-incongruent content, other factors could bias users' perceptions, and it is important to identify conditions that may amplify these biases.

Further, the cognitive processing of messages from either bots or humans remains an important area for future research. In the current study, it is difficult to assess the amount of cognitive effort people devote to processing each message, particularly in relation to (perceived) source cues, which may influence the extent to which messages affect attitude change (Chaiken & Ledgerwood, 2012; Chaiken & Maheswaran, 1994). If people engage more deeply with LLM-generated messages than with non-LLM-generated content, this could amplify the impact on attitude change. While social media messages are often short-lived, it is critical to understand how cognitive processing might vary depending on the type of content, as this could shape long-term effects.

Moreover, the effects of repeated exposure to messages should also be examined. Given the transient nature of attention specifically on social media, users' attention spans tend to be short. However, repeated exposure to the same messages over time may influence attitudes in ways that are not immediately evident (Bornstein, 1989; Schmidt & Eisend, 2015). Yan et al. (2023) reported the effects of people's exposure to bots increased individuals' perpetual bias, leading them to overestimate others' vulnerability to bots and decreased their self-efficacy in recognizing bots. Although this experiment also relied on relatively short exposure of the message, it still showed a significant impact on subsequent evaluations. That being said, investigating the effects of repeated or extended exposure on message acceptance or rejection would be a valuable avenue for further research (e.g., Skurka & Keating, 2024).

Lastly, while this dissertation focused on GMOs, it would be insightful to replicate this research in different contexts. Exploring how bots influence perceptions and attitudes in other topical areas could offer broader implications for understanding the dynamics of online discourse.

REFERENCES

Aldayel, A., & Magdy, W. (2022). Characterizing the role of bots' in polarized stance on social media. *Social Network Analysis and Mining, 12*(1), 30. https://doi.org/10.1007/s13278-022-00858-z

Badawy, A., Ferrara, E., & Lerman, K. (2018, August). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 258-265). IEEE.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., & Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences, 117*(1), 243–250. https://doi.org/10.1073/pnas.1906420116

Baly, R., Martino, G. D. S., Glass, J., & Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. *arXiv preprint arXiv:2010.05338*. https://doi.org/10.48550/arXiv.2010.05338

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin, 106*(2), 265-289. https://doi.org/10.1037/0033-2909.106.2.265

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health, 108*(10), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Burtell, M., & Woodside, T. (2023). Artificial influence: An analysis of AI-driven persuasion. *arXiv preprint arXiv:2303.08721*. https://doi.org/10.48550/arXiv.2303.08721

Caldarelli, G., De Nicola, R., Del Vigna, F., Petrocchi, M., & Saracco, F. (2020). The role of bot squads in the political propaganda on Twitter. *Communications Physics, 3*(1), 81. https://doi.org/10.1038/s42005-020-0340-4

Carolus, A., Muench, R., Schmidt, C., & Schneider, F. (2019). Impertinent mobiles-Effects of politeness and impoliteness in human-smartphone interaction. *Computers in Human Behavior*, *93*, 290-300. https://doi.org/10.1016/j.chb.2018.12.030

Catherine, K. N., Mugiira, B. R., & Muchiri, N. J. (2024). Public perception of genetically modified organisms and the implementation of biosafety measures in Kenya. *Advances in Agriculture*, *2024*(1), 5544617. https://doi.org/10.1155/2024/5544617

Chaiken, S., & Ledgerwood, A. (2012). A theory of heuristic and systematic information processing. In P. A. M. Van Lange, E. T. Higgins, & A. W. Kruglanski (Eds.), *Handbook of theories of social psychology* (Vol. 1, pp. 246-266). Sage Publications.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology, 66*(3), 460–473. https://doi.org/10.1037/0022-3514.66.3.460

Chinn, S., Hart, P. S., & Soroka, S. (2020). Politicization and polarization in climate change news content, 1985-2017. *Science Communication*, *42*(1), 112-129. https://doi.org/10.1177/107554701990029

Chu, H., & Liu, S. (2024). Can AI tell good stories? Narrative transportation and persuasion with ChatGPT. *Journal of Communication*, *74*(5), 347-358. https://doi.org/10.1093/joc/jqae029

Clerwall, C. (2017). Enter the robot journalist: Users' perceptions of automated content. In *The Future of Journalism: In an Age of Digital Media and Economic Uncertainty* (pp. 165-177). Routledge.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., …Zhao, Z. (2024). The Llama 3 Herd of Models (arXiv:2407.21783). *arXiv.* https://arxiv.org/abs/2407.21783

Edwards, C., Beattie, A. J., Edwards, A., & Spence, P. R. (2016). Differences in perceptions of communication quality between a Twitterbot and human agent for information seeking and learning. *Computers in Human Behavior*, *65*, 666-671. https://doi.org/10.1016/j.chb.2016.07.003

Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2014). Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior*, *33*, 372-376. https://doi.org/10.1016/j.chb.2013.08.013

Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

Ferrara, E. (2020). # covid-19 on twitter: Bots, conspiracies, and social media activism. *arXiv preprint arXiv: 2004.09531*. https://arxiv.org/abs/2004.09531

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Funk, C. (2020, March 18). About half of U.S. adults are wary of health effects of genetically modified foods, but many also see advantages. *Pew Research Center*. https://www.pewresearch.org/short-reads/2020/03/18/about-half-of-u-s-adults-are-wary-of-health-effects-of-genetically-modified-foods-but-many-also-see-advantages/

Gorwa, R., & Guilbeault, D. (2020). Unpacking the Social Media Bot: A Typology to Guide Research and Policy. *Policy & Internet*, *12*(2), 225–248. https://doi.org/10.1002/poi3.184

Graefe, A., Haim, M., Haarmann, B., & Brosius, H. B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism, 19*(5), 595-610. https://doi.org/10.1177/1464884916641269

Grootendorst, M. (2022). BERTopic: Neural Topic Modeling With a Class-Based TFIDF Procedure. *arXiv:2203.05794v0571*. https://doi.org/10.48550/arXiv.2203.05794

Groshek, J. (2011). Media, instability, and democracy: Examining the Granger-causal relationships of 122 countries from 1946 to 2003. *Journal of Communication, 61*(6), 1161-1182. https://doi.org/10.1111/j.1460-2466.2011.01594.x

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society, 22*(1), 70-86. https://doi.org/10.1177/146144481985869

Hallman, W. K., Cuite, C. L., & Morin, X. K. (2013). *Public perceptions of labeling genetically modified foods* (Working Paper 2013–1). Rutgers, The State University of New Jersey, School of Environmental and Biological Sciences. http://humeco.rutgers.edu/documents_PDF/news/GMlabelingperceptions.pdf

Hartmann, J. (2022). *Emotion English DistilRoBERTa-base*. Hugging Face. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/

Heo, R.J., Heo, J., Lee, S., & Peng, T.Q. (2024). *Leveraging large language models to detect bots: Utilizing meta, temporal, and social interaction information* [Manuscript submitted for publication]. Department of Communication. Michigan State University

Hewitt, L., Ashokkumar, A., Ghezae, I., & Willer, R. (n.d.). *Predicting results of social science experiments using large language models* [Working paper]. https://docsend.com/view/ity6yf2dansesucf

Hocevar, K. P., Metzger, M., & Flanagin, A. J. (2017). Source credibility, expertise, and trust in health and risk messaging. In *Oxford research encyclopedia of communication*. https://doi.org/10.1093/acrefore/9780190228613.013.287

Hohenstein, J., & Jung, M. (2020). AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior, 106*, 106190. https://doi.org/10.1016/j.chb.2019.106190

Howell, E. L., Wirz, C. D., Brossard, D., Jamieson, K. H., Scheufele, D. A., Winneg, K. M., & Xenos, M. A. (2018). National Academies of Sciences, Engineering, and Medicine report on genetically engineered crops influences public discourse. *Politics and the Life Sciences, 37*(2), 250-261. https://doi.org/10.1017/pls.2018.12

Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. *Journal of Communication, 73*(6), 552-562. https://doi.org/10.1093/joc/jqad024

Jiang, M., Liu, K. Z., Zhong, M., Schaeffer, R., Ouyang, S., Han, J., & Koyejo, S. (2024). Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*. https://doi.org/10.48550/arXiv.2401.06059

Jun, I., Zhao, Y., He, X., Gollakner, R., Court, C., Munoz, O., Bian, J., Capua, I., & Prosperi, M. (2020). Understanding perceptions and attitudes toward genetically modified organisms on Twitter. *International Conference on Social Media and Society*, 291–298. https://doi.org/10.1145/3400806.3400839

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1-29. https://doi.org/10.1145/3579592

Keller, T. R., & Klinger, U. (2019). Social bots in election campaigns: Theoretical, empirical, and methodological implications. *Political Communication*, *36*(1), 171-189. https://doi.org/10.1080/10584609.2018.1526238

Leshner, G., Reeves, B., & Nass, C. (1998). Switching channels: The effects of television channels on the mental representations of television news. *Journal of Broadcasting & Electronic Media*, *42*(1), 21-33. https://doi.org/10.1080/08838159809364432

Li, Y., Guo, Y., Guerin, F., & Lin, C. (2024, November). An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 528-541). https://doi.org/10.18653/v1/2024.findings-emnlp.30

McFadden, B. R., Rumble, J. N., Stofer, K. A., & Folta, K. M. (2024). US public opinion about the safety of gene editing in the agriculture and medical fields and the amount of evidence needed to improve opinions. *Frontiers in Bioengineering and Biotechnology*, *12*, 1340398. https://doi.org/10.3389/fbioe.2024.1340398

Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research*, *26*(4), 323-339. https://doi.org/10.1086/209566

Naqvi, S., Zhu, C., Farre, G., Ramessar, K., Bassie, L., Breitenbach, J., ... & Capell, T. (2009). Transgenic multivitamin corn through biofortification of endosperm with three vitamins representing three distinct metabolic pathways. *Proceedings of the National Academy of Sciences*, *106*(19), 7762–7767. https://doi.org/10.1073/pnas.0901412106

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81-103.https://doi.org/10.1111/0022-4537.00153

Nass, C., & Steuer, J. (1993). Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research*, *19*(4), 504-527. https://doi.org/10.1111/j.1468-2958.1993.tb00311.x

Nass, C., Isbister, K., & Lee, E. J. (2000). Truth is beauty: Researching embodied conversational agents. *Embodied Conversational Agents*, *2000*, 374-402.

Nass, C., Jonsson, I. M., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005, April). Improving automotive safety by pairing driver emotion and car voice emotion. In *CHI'05 extended abstracts on Human factors in computing systems* (pp. 1973-1976). https://doi.org/10.1145/1056808.1057070

Nass, C., Moon, Y., & Green, N. (1997). Are computers gender-neutral? Gender stereotypic responses to computers. *Journal of Applied Social Psychology*, *27*(10), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, C. (1995, May). Can computer personalities be human personalities?. In *Conference Companion on Human Factors in Computing Systems* (pp. 228-229).

Nass, C., Steuer, J., & Tauber, E. R. (1994, April). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72-78).

Qiu, L., & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of management information systems*, *25*(4), 145-182. https://doi.org/10.2753/MIS0742-1222250405

Rathod, D., & Hedaoo, R. P. (2022). Assessment of knowledge and attitudes on genetically modified foods among students studying life sciences. *Cureus*, *14*(12). https://doi.org/10.7759/cureus.32744

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2-3), 140-157. https://doi.org/10.1080/19312458.2018.1455817

Satariano, A., & Mozur, P. (2023, February 7). The people onscreen are fake. The disinformation is real. *The New York Times*. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

Savage, S., Monroy-Hernandez, A., & Höllerer, T. (2016, February). Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 813-822).

Schmidt, S., & Eisend, M. (2015). Advertising repetition: A meta-analysis on effective frequency in advertising. *Journal of Advertising*, *44*(4), 415-428. https://doi.org/10.1080/00913367.2015.1018460

Schneider, F. (2020). How users reciprocate to Alexa. In C. Stephanidis et al. (Eds.), *HCI International 2020—Late breaking posters* (pp. 376–383). Springer International Publishing. https://doi.org/10.1007/978-3-030-60700-5_48

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 96, 104. http://arxiv.org/abs/1707.07592

Skurka, C., & Keating, D. M. (2024). How repeated exposure to persuasive messaging shapes message responses over time: a longitudinal experiment. *Human Communication Research*, hqae008. https://doi.org/10.1093/hcr/hqae008

Sohi, M., Pitesky, M., & Gendreau, J. (2023). Analyzing public sentiment toward GMOs via social media between 2019-2021. *GM Crops & Food, 14*(1), 1-9. https://doi.org/10.1080/21645698.2023.2190294

Srinivasan, V., & Takayama, L. (2016, May). Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4945-4955). https://doi.org/10.1145/2858036.2858217

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences, 115*(49), 12435–12440. https://doi.org/10.1073/pnas.1803470115

Stocking, G., & Sumida, N. (2018, October 15). *Social media bots draw public's attention and concern*. Pew Research Center. https://www.journalism.org/2018/10/15/social-media-bots-draw-publics-attention-and-concern/

Sundar, S. S., & Nass, C. (2000). Source orientation in human-computer interaction: Programmer, networker, or independent social actor. *Communication Research, 27*(6), 683-703. https://doi.org/10.1177/009365000027006001

Sundar, S. S., & Nass, C. (2001). Conceptualizing sources in online news. *Journal of Communication, 51*(1), 52-72. https://doi.org/10.1111/j.1460-2466.2001.tb02872.x

Swanson, N. R., & Granger, C. W. (1997). Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association, 92*(437), 357-367. https://doi.org/10.1080/01621459.1997.10473634

Wischnewski, M., Bernemann, R., Ngo, T., & Krämer, N. (2021, May). Disagree? You must be a bot! How beliefs shape twitter profile perceptions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-11). https://doi.org/10.1145/3411764.3445109

Wischnewski, M., Ngo, T., Bernemann, R., Jansen, M., & Krämer, N. (2024). "I agree with you, bot!" How users (dis) engage with social bots on Twitter. *New Media & Society, 26*(3), 1505-1526. https://doi.org/10.1177/14614448211072307

Wunderlich, S., & Gatto, K. A. (2015). Consumer perception of genetically modified organisms and sources of information. *Advances in Nutrition, 6*(6), 842-851. https://doi.org/10.3945/an.115.008870

Xu, W., & Sasahara, K. (2022). Characterizing the roles of bots on Twitter during the COVID-19 infodemic. *Journal of Computational Social Science, 5*(1), 591–609. https://doi.org/10.1007/s42001-021-00139-3

Yan, H. Y., Yang, K. C., Shanahan, J., & Menczer, F. (2023). Exposure to social bots amplifies perceptual biases and regulation propensity. *Scientific Reports, 13*(1), 20707. https://doi.org/10.1038/s41598-023-46630-x

Zhang, Y., Shah, D., Pevehouse, J., & Valenzuela, S. (2023). Reactive and asymmetric communication flows: Social media discourse and partisan news framing in the wake of mass shootings. *The International Journal of Press/Politics*, *28*(4), 837-861. https://doi.org/10.1177/19401612211072793

Zhang, Y., Sharma, K., Du, L., & Liu, Y. (2024, May). Toward Mitigating Misinformation and Social Media Manipulation in LLM Era. In *Companion Proceedings of the ACM on Web Conference 2024* (pp. 1302-1305). https://doi.org/10.1145/3589335.3641256

APPENDIX A: SUPPLEMENTARY TABLES AND FIGURES

Table S-13
Description of Key Terms

Terms	Descriptions
Username (handle)	It is a public identifier of users that is used to log in to your account and is visible when sending and receiving replies and Direct Messages.
Retweet (RT)	Retweeting involves sharing someone else's tweet with the given user's own followers
Followers	Followers refer to accounts that follow the user.
Followees	Followers refer to accounts that the user follows.
Mention	Mentioning enables a user to include other people's usernames anywhere in the user's own tweet.
Hashtags	Self-assigned topic categories for tweets that people include alongside their posts.
FAV	Favorite (FAW) is a feature that allows users to bookmark tweets and show appreciation for them
BERTopic	Unsupervised topical modeling that is used to extract latent topics from documents
VAR Model	Vector Autoregression (VAR) models are used to analyze the dynamic relationships between multiple time series variables and how changes in one variable may influence others over time (Swanson & Granger, 1997).
IRF	Impulse Response Functions (IRFs) show how the variables in a model react over time to a shock in one or more of the model's variables (Swanson & Granger, 1997; Zhang et al. 2023).

Table S-13 (cont'd)

Granger Causality	Granger causality tests estimate whether lags of one variable can be used to predict another variable longitudinally (Groshek 2011)

Table S-14
Dictionaries for Politicalization

	A list of words
Republicans	republican, republicans, gops, gop, conservatives, conservative
Democrats	democrat, democrats, liberal, liberals

Figure S-4
Stimulus for Each Condition

LLM-generated bot

Figure S-4 (cont'd)

LLM-generated human

Figure S-4 (cont'd)

Bot

Figure S-4 (cont'd)

Figure S-4 (cont'd)

Human

Figure S-4 (cont'd)

APPENDIX B: PILOT TEST RESULTS FOR BOT DETECTION

Although the preliminary study suggested the potential of incorporating temporal information alongside random examples (Heo et al., 2024), the present study conducted a pilot experiment to validate the proposed method. In the first pilot study, a separate dataset consisting of 800 user accounts (200 users from each quartile of activity) was extracted from the original dataset. Using the proposed method with GPT-4, which included temporal information and random examples, the results indicated a skewed distribution: 79% of decisions identified as bots, 2% as humans, and 19% as undecidable. This skewed tendency raised concerns, prompting further verification before proceeding with the method.

In the subsequent pilot study for the present research, 100 random users were selected from the final dataset of 2,400 users. Consistent with the prior study, GPT-4 was employed to detect bots using meta-information, text, and temporal information. To improve task performance, an ensemble approach was adopted, employing majority voting based on decisions from each information type. Additionally, for a few-shot prompting, random cases were introduced. Given that the initial inclusion of temporal information with random cases led to unexpected results, I also incorporated both correct and incorrect examples, anticipating that this would help contextualize the decision-making process. Specifically, two example cases were provided for each decision, following a confusion matrix: 1) identify bots as bots, 2) identify bots as humans, 3) identify humans as bots, and 4) identify humans as humans. To establish ground truth, an expert (the author) classified each user as either a bot or a human.

The results revealed generally low F1 scores across all conditions. However, the condition in which temporal information was provided alongside random examples yielded the highest F1 score (0.44), followed by the condition with temporal information and correct/incorrect cases (F1 = 0.43). Other combinations of information were not effective, and the ensemble approach did not surpass the performance of the few-shot prompting conditions (see Table S-15).

Given the low F1 scores, an additional analysis was performed using the widely-used bot detection tool, Botometer, to compare its output against the ground truth labels. The Botometer produces a score ranging from 0 to 5, with two thresholds—2.5 and 3—being used to classify users as bots. However, the results revealed that Botometer's classification also failed to generate high F1 scores, with a score of 0.32 for the 3.0 threshold. Moreover, no true positives or false negatives were detected, further illustrating the challenges in bot detection.

These results highlight the inherent complexity of bot detection. To assess the level of agreement among experts, three experts (including the author) were tasked with coding 52 user accounts as either bots or humans. The average pairwise percent agreement among experts was 70%, underscoring the challenge of bot classification, even among trained professionals. Despite the limitations of the current bot detection method, I proceeded with the use of GPT-4, temporal information, and example cases. Although temporal information combined with random examples yielded the highest F1 scores, this approach resulted in a high number of false positives and a low rate of true negatives. To address this issue, I incorporated alternative examples (both

correct and incorrect cases), which helped reduce false positives and improve the identification of true positives. This revised approach was ultimately adopted for the main study.

Table S-15
The Performance Metrics of Bot Detection Tools

	Accuracy	Precision	Recall	F1
Meta	0.65	0.33	0.36	0.35
Text	0.56	0.33	0.53	0.41
Temporal	0.42	0.24	0.54	0.33
Social interaction	0.46	0.30	0.60	0.40
Ensemble (majority voting)	0.60	0.33	0.40	0.36
Temporal + Random examples	0.33	0.30	0.86	0.44
Temporal + Correct and incorrect examples	0.45	0.31	0.69	0.43
Botometer (threshold 3)	0.72	0.50	0.23	0.32
Botometer (threshold 2.5)	0.85	0	.	.

APPENDIX C: AN INSTRUCTION FOR SIMLUATING TWEETS USING GPT-4

Here’s a compilation of tweets about genetically modified organisms (GMOs). The goal is to create five new tweets opposing GMO issues based on this compilation. It should be realistic. When including links, please ensure they come from valid sources and support the content of the tweet. Aim to match the average length of the example tweets provided, but feel free to vary your tweets' lengths within the 280-character limit

APPENDIX D: AN INSTRUCTION FOR PERSONA SIMULATION USING LLAMA-3

Demographic details of this person: Ethnicity/Race: [race], Age: [#], Education: [education], Gender: [gender], Partisanship: [partisanship], Income: [income], Area: [area], Region: [region], Experience in a food or agricultural field: [# of years], Experience in a health or medical field: [# of years].

You will be asked to read a message. Please read it carefully and then answer the following questions. Please respond by indicating the number that best reflects your opinion.

Message:

[INSERTED STIMULUS BASED ON THE ASSIGNED CONDITION]

Question 1: What is your opinion about the safety of gene editing in the context of food and agriculture?
1. Extremely safe
2. Somewhat safe
3. Neither safe nor unsafe
4. Somewhat unsafe
5. Extremely unsafe
Question 1 Answer:

Question 2: What is your opinion about the safety of gene editing in the context of health and medical applications?
1. Extremely safe
2. Somewhat safe
3. Neither safe nor unsafe
4. Somewhat unsafe
5. Extremely unsafe
Question 2 Answer:

Question 3: Please indicate the extent to which you think this message was likely created by a human, using a scale from 0 to 100 (0 = not at all, 100 = very much).
Question 3 Answer:

Question 4: Please indicate the extent to which you think this message was likely created by a bot, using a scale from 0 to 100 (0 = not at all, 100 = very much).
Question 4 Answer: