Human-authored or AI-generated? Humans can usually tell the difference for political content on social media platforms

Sejin Paik[a, 1], **Tiago Ventura**[b], **Rebecca Ansell**[c], **Leticia Bode**[d], **Autumn Toney**[c], **and Lisa Singh**[a, b, c]

This manuscript was compiled on March 15, 2026

This study employs a set of behavioral experiments to examine how individuals perceive the humanness of social media posts about the 2024 U.S. presidential debate between Kamala Harris and Donald Trump. Posts presented in the experiment were authored either by humans or chatbots, varying the large language model (LLMs), the model architecture, the social media platform, and the political personas. Participants judged the randomly-paired posts on which they perceived to be AI-generated. Overall, human-authored posts were judged more human than AI outputs, but this separation narrowed for newer LLMs, with Llama 3.3 surpassing human-like judgments. The social media platform also mattered, with participants distinguishing human from AI on most platforms (YouTube, TikTok, Truth Social), but not on X/Twitter. Partisan identity also shaped evaluations, with Republican respondents more likely to view ideologically-aligned AI posts as human-authored. Our results indicate that audiences can still distinguish human from AI generated social media posts, but this ability weakens for newer LLMs, varies by platform, and reflects partisan bias.

Perceived humanness | AI-generated content (AIGC) | Large Language Models (LLMs) | Social media | Political communication |

Significance Statement

As AI chatbots become easily accessible and more powerful, understanding whether audiences can distinguish AI-generated from human-authored content is central to information integrity and democratic accountability. Using a set of behavioral experiments, we demonstrate that humans are still capable of identifying AI-content on social media across most platforms and models. Yet this ability is shaped by technological design, linguistic features, and motivated reasoning. This conditionality has real-world implications in that it reveals where human judgment serves as a viable defense against harmful or manipulated AI-generated political content, and where it cannot. Our findings provide practical insights calibrated to actual vulnerabilities rather than assumed universal detection failure. Understanding these boundaries is essential for building adaptive democratic safeguards that support informed political judgment in AI-satured information environments.

The rise of generative Artificial Intelligence (GenAI) represents a critical inflection point for democratic information systems (1). Previous technological shifts, such as the introduction of the internet and mobile technologies, fundamentally reshaped how information was distributed and accessed (2). GenAI differs in a fundamental way. Enabled by large language models (LLMs) (3, 4), texts that resemble human expression in their tone, syntax, and affective-style (5) can now be artificially mass-produced at an unprecedented scale (6), potentially making it difficult for humans to distinguish authentic political discourse from synthetic imitations.

With social media now serving as a dominant venue for online information consumption (7), both human-authored and AI-generated content (AIGC) may potentially circulate side by side (8). This raises fundamental questions about people's capacity to assess content authenticity, an ability with direct implications for political decision-making. If AIGC that amplifies certain viewpoints or creates false impressions of grassroots support cannot be distinguished from authentic participation, it may erode trust or increase misinterpretations in democratic discourse (9). Yet whether audiences actually lack the capacity to detect AIGC, or rather face specific, identifiable vulnerabilities under certain conditions, remains largely unknown.

These hypothesized concerns unfolded in the 2024 U.S. presidential election cycle. LLMs scaled the production of campaign materials (10), deepfake phone calls imitating candidates' voices were disseminated to voters (11), and AI-generated images and videos of political figures circulated on social media platforms (12) used to mock opponents or blur distinctions between satire and propaganda (13). The spread of low quality AIGC, or "AI slop" (14) contributed to emergent forms of misinformation (15). These incidents emphasize the urgency of understanding humans' detection and perception capacity, particularly as platform-based solutions prove inadequate. AI labeling practices have failed to scale (16) and their effectiveness remains contested (17, 18). Where platforms cannot reliably identify AIGC, audience judgments serve as a critical fallback mechanism for navigating authenticity.

To understand what shapes these authenticity judgments, we investigate audience perceptions of human-authored versus AI-generated posts in online political discourse through four research questions: (1) Can people distinguish human-

Author affiliations: [a]Massive Data Institute (MDI), Georgetown University, Washington, D.C. 20001; [b]McCourt School of Public Policy, Georgetown University, Washington, D.C. 20001; [c]Department of Computer Science, Georgetown University, Washington, D.C. 20007; [d]Communication, Culture, and Technology, Georgetown, University, Washington, D.C. 20057

S.P. directed the project end-to-end; S.P., T.V., R.A., L.B., A.T., and L.S. conceptualized, wrote, and edited the paper; T.V., S.P., and R.A. designed experiments and performed research; T.V. developed key analysis models; T.V., R.A., and S.P. curated and analyzed data; L.S., L.B., and T.V. acquired funding.

The authors declare no competing interests.

[1]To whom correspondence should be addressed. E-mail: sp1822georgetown.edu

authored content from AI-generated posts created using different LLMs across different social media platforms? (2) What semantic features influence people's evaluation of *perceived humanness*? (3) What individual traits influence people's capacity to distinguish human-authored from AI-generated posts? (4) How do people's political leanings shape their perception of humanness of AI-generated posts that adopt politically congruent versus incongruent personas?

These questions contribute to three core research threads. The first focuses on understanding AI detection in political contexts. Non-experts struggle to distinguish AI from human text (19, 20), often relying on surface cues like first-person pronouns or warm tones (21), and semantic cues signaling credibility or coherence (22). In political contexts, emotional tone, stylistic fluency, and conversational style serve as powerful cues of legitimacy and persuasiveness (23–25). As such, AIGC mimicking these markers may evade detection while shaping political attitudes (26). The second agenda aims to identify individual predictors of detection capacity. Individual traits such as political orientation and platform use shape how audiences evaluate synthetic content (27, 28). Third, we examine partisan congruence effects. When AIGC adopts political personas, alignment between audience and content ideology influences authenticity judgments (29, 30), with politically congruent content more likely to be perceived as human-authored.

This study consists of online surveys conducted in two-waves (combined, unique respondents = 1,122) with a collection of 19,500 unique pairwise comparisons using 2,386 social media posts (1,690 AI-generated, 696 human-authored) about the 2024 U.S. presidential debate between Kamala Harris and Donald Trump. This salient political event provided a consistent topical anchor in a high-stakes domain for AIGC (31). Using a pairwise design, participants evaluated human-authored* and AI-generated posts produced by LLMs and judged which was more likely AI-generated. The AI-generated posts were produced by six LLMs (GPT-4, GPT-4o, GPT-4o mini, Llama 2, Llama 3.2, Llama 3.3), imitating different platform discourse (X/Twitter, TikTok, YouTube, Truth Social), and taking on five political personas (Progressive, Liberal, Moderate, Conservative, Libertarian Populist).

We quantified perceptions of humanness by scaling pairwise responses with the Bradley–Terry (BT) statistical model to convert the responses into a novel *Perceived Humanness Indicator (PHI)* score, ranging from −1 (AI-authorship) to 1 (human-authorship). Pairwise comparisons provide robust scaling of latent perceptions – in our case, perceptions of humanness of social media posts – by leveraging relative rather than absolute judgments (in other words, by asking which of the two posts seems more AI-like rather than requiring an absolute rating). This approach has been widely used in social science to measure argument persuasiveness, ideology, and textual complexity (32, 33). This design enabled us to analyze how perceptions of humanness differed across model outputs, platform context, semantic features, and partisan alignment.

The study yields three main contributions. First, to the best of our knowledge, we conduct the first large-scale, cross-platform evaluation of *perceived humanness* in online political discourse. Second, we propose a novel *Perceived Humanness Indicator (PHI)* score for text that estimates how human a post is assessed to be, where we define *perceived humanness* as the subjective impression that a piece of content was authored by a human. This definition captures how individuals interpret the linguistic properties and other cues that suggest humanness in written communication. This in turn helps us measure how audiences evaluate content authenticity. Third, our results demonstrate that audiences can distinguish AI-generated from human-authored political content across most conditions, yet this capacity varies systematically. Detection declines with newer, larger LLMs while remaining generally robust on most social media platforms. Certain semantic features, related to civility, emotion, and sentiment, influence perceived humanness. Individual differences also shape detection, with men and white participants showing higher accuracy than women and non-white participants, and Republicans selectively misclassify ideologically-aligned AI personas as human.

Results

We report results in four parts, mapping to our research questions: (1) discernment of human versus AI authorship based on LLMs and platforms, (2) semantic features on humanness judgment, (3) individual traits that predict detection capacity, and (4) political partisan alignment on humanness perceptions of AI-generated posts.

Individuals' ability to discern human-authored from AI-generated posts. We begin by examining participants' ability to distinguish between human-authored posts and AI-generated posts across six different LLMs using *PHI* scores that range from −1 (AI-authorship) to 1 (human-authorship). Figure 1, "A) Pooled Estimates" presents the pooled marginal means of *PHI* scores, with each point estimate representing the adjusted mean level of *perceived humanness* and the error bars indicating 90% and 95% confidence intervals (CIs). Values above zero indicate judgments leaning toward 'more human' while values below zero indicate judgments leaning toward 'more AI-like'.†

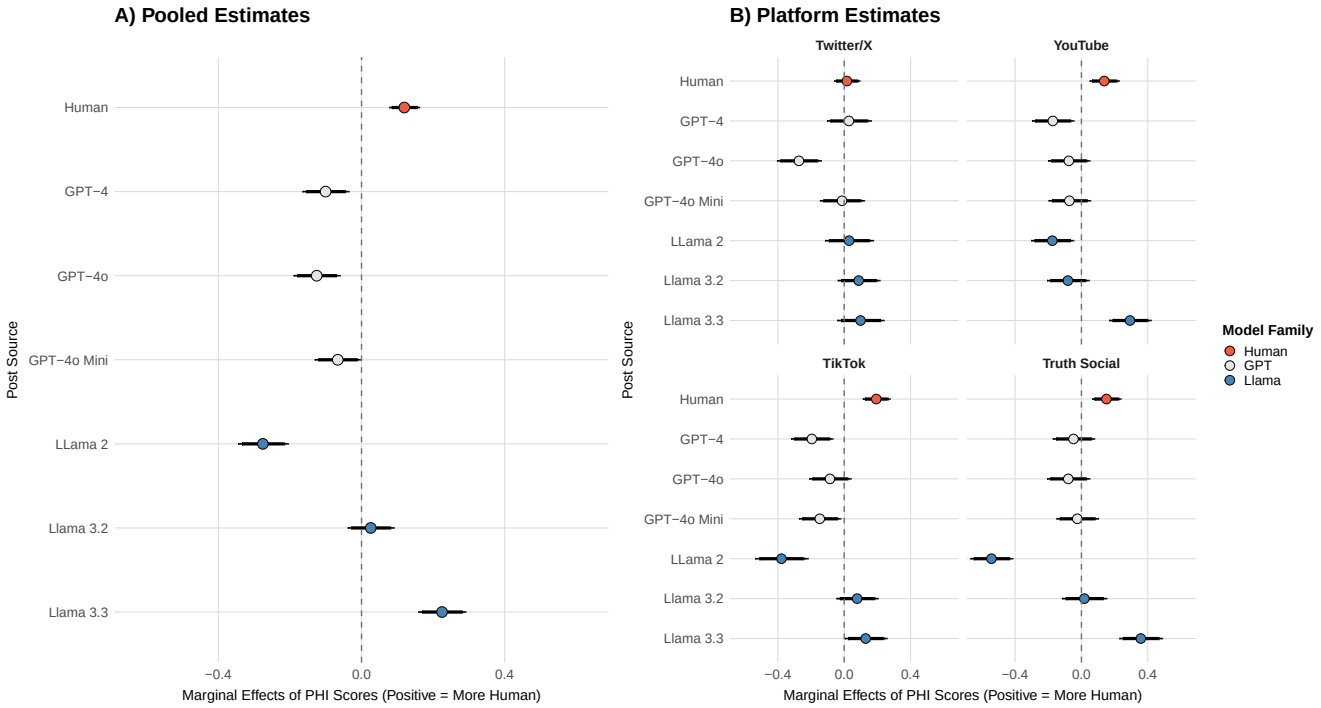
When comparing participants' judgement of humanness across the six LLMs, human-authored posts were judged as significantly more human (M = 0.12, 95% CI [0.08, 0.16]). Outputs from four (GPT-4, GPT-4o, GPT-4o mini, Llama 2) out of six LLMs were judged as significantly more AI-like, with mean PHI scores ranging from –0.28 to –0.07 (GPT-4: M = –0.10, 95% CI [–0.17, –0.04]; GPT-4o: M = –0.13, 95% CI [–0.19, –0.06]; GPT-4o mini: M = –0.07, 95% CI [–0.13, –0.00]; Llama 2: M = –0.28, 95% CI [–0.35, –0.21]). Posts generated by Llama 3.3 were the exception and judged as significantly more human-like (*Mean* = 0.23, 95% CI [0.16, 0.29]), indicating that participants frequently misclassified posts from the more recent Llama model. Llama 3.2 was the only model not significantly different from zero. Pooling from all social media platforms, results show participants are able to distinguish between human-authored and AI-generated posts.

We now describe how *PHI* scores varied across the social media platforms. Figure 1, B) Platform Estimates presents

*We used a bot detector to determine if the posts were likely authored by a bot or not, and found only 2 were classified as likely bots. See *SI Appendix* for more details.

†As AI-generated posts comprised the majority of stimuli in our dataset, the zero line captures participants' perceptual midpoint rather than an equal distribution of human-authored or AI-generated posts.

Fig. 1. Pooled- and Platform-Level Estimates of *Perceived Humanness*



unpooled marginal means with 95% and 90% confidence intervals of the *PHI* scores across four different platforms – Twitter/X, Youtube, Truth Social, TikTok.

Three of the four platforms displayed noticeable separation in participants' ability to distinguish posts as human or AI. YouTube showed the clearest distinction. Human-authored posts were judged as significantly more human (M = 0.13, 95% CI [0.04, 0.22]), while most LLM-generated posts were correctly classified as AI, especially posts from GPT-4 (M = -0.17, 95% CI [-0.29, -0.04]) and Llama 2 (M = -0.17, 95% CI [-0.30, -0.04]). Llama 3.3 was the one LLM whose YouTube comments were identified as more human-like than human posts. TikTok and Truth Social showed similar patterns as YouTube. Twitter/X showed the weakest differentiation between human- and AI-generated posts. Among Twitter/X model outputs, only GPT-4o was significantly judged as more AI-like (M = 0.27, 95% CI [-0.40, -0.14]). For the rest, human posts as well as the other LLM outputs were not statistically different from zero. In sum, our unpooled results are robust across most platforms. YouTube, TikTok, and Truth Social demonstrated a clear human-AI separation with five of six models perceived as more AI-like and one perceived as more human-like than the human posts. Twitter/X was a notable anomaly with minimal reliable discernment.

Semantic features of *perceived humanness* at the post-level. We now investigate how semantic features of the human-authored and AI-generated posts correlate with participants' perceptions of humanness. We define "semantic features" as measurable linguistic cues that encode what is said (denotation) and how it is said (connotation) (34).[‡]

Figure 2 presents the marginal effects for each semantic feature with corresponding 90% and 95% confidence intervals. Results show that in the first semantic category of civility, both offensive language (M = 0.12, 95% CI [0.05, 0.18]) and the presence of hateful speech (M = 0.17, 95% CI [0.01, 0.33]) were significant positive predictors of perceived humanness. Irony was not statistically different from zero. In the second semantic category, emotion, all three categories, optimism (M = 0.17, 95% CI [0.09, 0.25]), sadness (M = 0.17, 95% CI [0.05, 0.28]), and joy (M = 0.16, 95% CI [0.09, 0.23]), are statistically significant predictors of perceived humanness, showing participants often associate emotion with human-generated content. Thirdly, in the sentiment category, negative sentiment is positively associated with humanness (M = 0.16, 95% CI [0.01, 0.23]), and positive sentiment was negatively related to humanness (M = -0.08, 95% CI [-0.15, -0.01]), suggesting humans tend to perceive overly positive posts as being generated by AI chatbots[§]

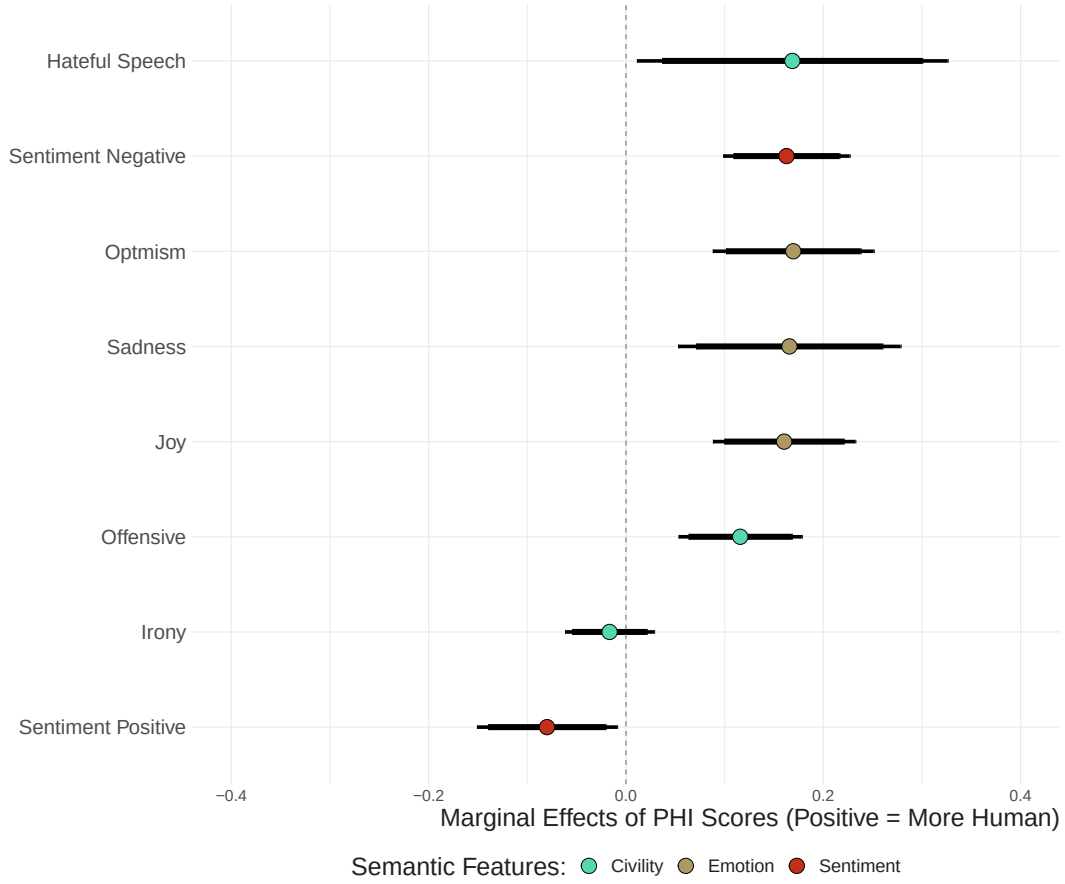
Individual traits that influence ability to identify AI-posts. We further examined how individual characteristics of the raters predict the ability to correctly identify AI-generated posts when paired against human-authored posts.[¶]

[‡] For every post, we used TweetEval (35) to classify whether the post contains the following three semantic categories: Civility (Irony, Hate, Offensive), Emotion (Joy, Optimism, Sadness), and Sentiment (Positive, Negative).

[§] As a robustness check, we carried out this semantic analysis for each platform (Twitter/X, YouTube, TikTok, and Truth Social). We found that the pattern of coefficients was broadly consistent with these aggregated results.

[¶] The experiment randomized all post pairings (human-human, AI-AI, or human-AI). For this analysis, we filtered to include only pairs containing one human-authored and one AI-generated post.

Fig. 2. Post-Level Semantic Features Associated with *Perceived Humanness*



Note: The horizontal bars represent the corresponding 90% and 95% confidence intervals.

With regard to demographics, only gender and race were significant predictors. Males were more likely than females to correctly identify AI posts *(*β = 0.021*,* 95% CI [0.00, 0.04]*)*, and White participants likewise showed a positive effect *(*β = 0.03, 95% CI [0.008, 0.067]*)* (plot A of Figure 3). All other demographic characteristics (age, education, hispanic, and political affiliation) were not significantly different from zero.

For the attitudinal covariates, higher offline news consumption was associated with significantly lower accuracy in identifying AI posts *(*β = -0.015, 95% CIs [-0.028,-0.002]*)*, meaning that individuals who rely more heavily on traditional, non-digital news sources are potentially worse at distinguishing AI posts from human-authored ones (plot B of Figure 3). Lower trust in news, both online and offline, was negative, but not statistically significant. Online News Consumption, News Cynicism, Interest in Politics, and Conspiracy Beliefs variables were also insignificant, with effects near zero.

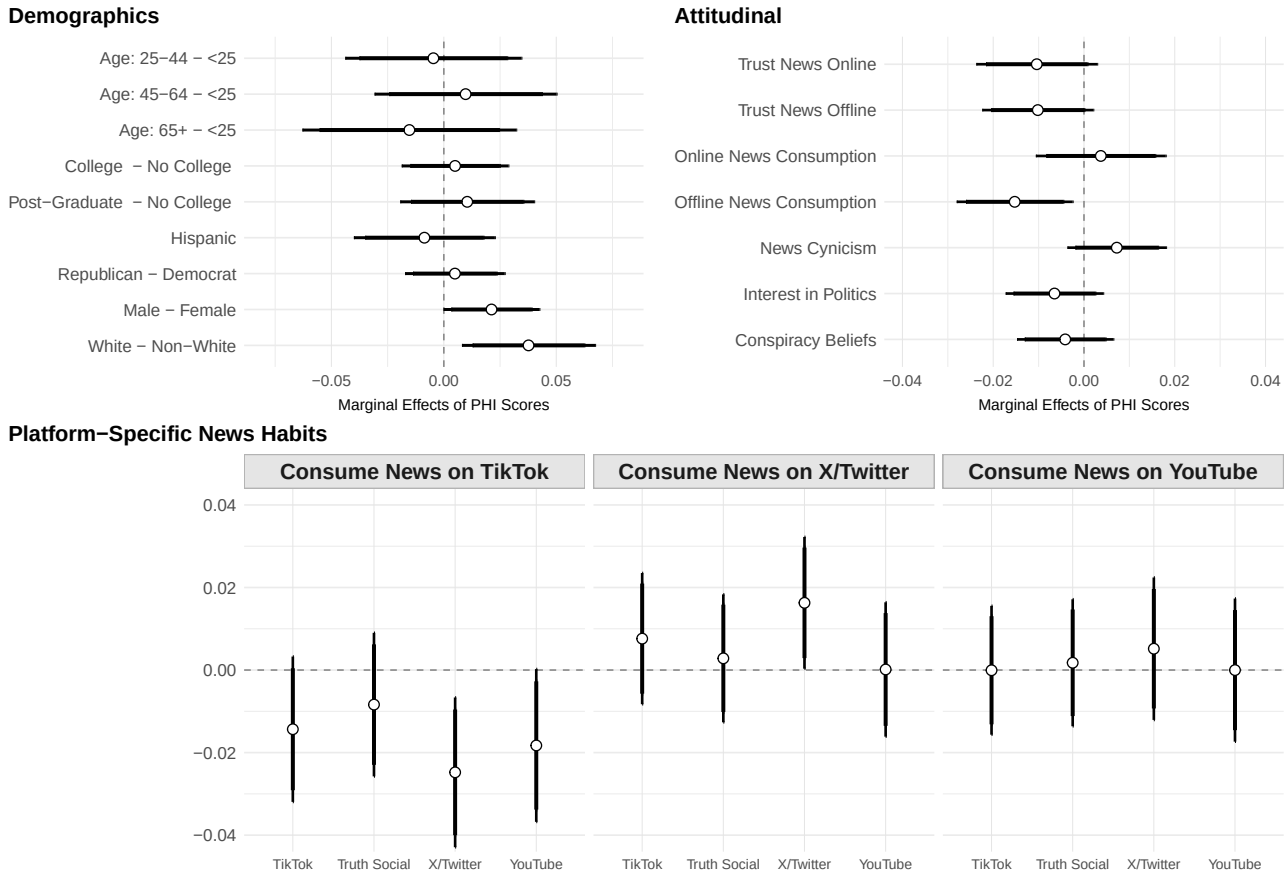
Lastly, platform-specific news habits revealed that people who consume news on X/Twitter were significantly more likely to identify AI authorship when evaluating X/Twitter posts *(*β = -0.024, 95% CI -0.006, -0.042]), whereas the analogous effect did not emerge for YouTube. In contrast, TikTok news consumers were significantly less likely to identify AI posts on other platforms such as X/Twitter and YouTube. (plot C of Figure 3). Together, the results indicate modest but meaningful individual differences centered on gender, race, and news consumption habits.

Political personification of LLMs and participants' partisan alignment. Finally, we explore whether partisan congruence between respondents and the political persona of AI-generated posts shapes humanness judgments.[‖]

Partisan asymmetries in *perceived humanness* emerged primarily among Republicans. They were less likely to label Conservative and Populist personas as AI-generated, treating those ideologically-aligned voices as more human. Republicans were also more likely to identify posts from Liberal and Progressive personas as being AI-generated. In the most extreme cases, Republicans were five percentage points (β=0.057, CI 95% [0.028, 0.085]) more likely to identify a progressive post as AI-generated compared to Democrats, and eight percentage points (β=0.081, CI 95% [0.052, 0.111]) and four percentage points (β=0.042, CI 95% [0.014, 0.071])

[‖] For this analysis, we consider only AI-AI pairs, and examine how participants' party identification interacted with the political persona of the posts.

Fig. 3. Individual-Level Predictors of Correctly Identifying GenAI Posts



less likely to identify a post from a Libertarian Populist and Conservative personas, respectively, as AI-generated (see Figure 4). We did not observe strong partisan effects among Democrats.

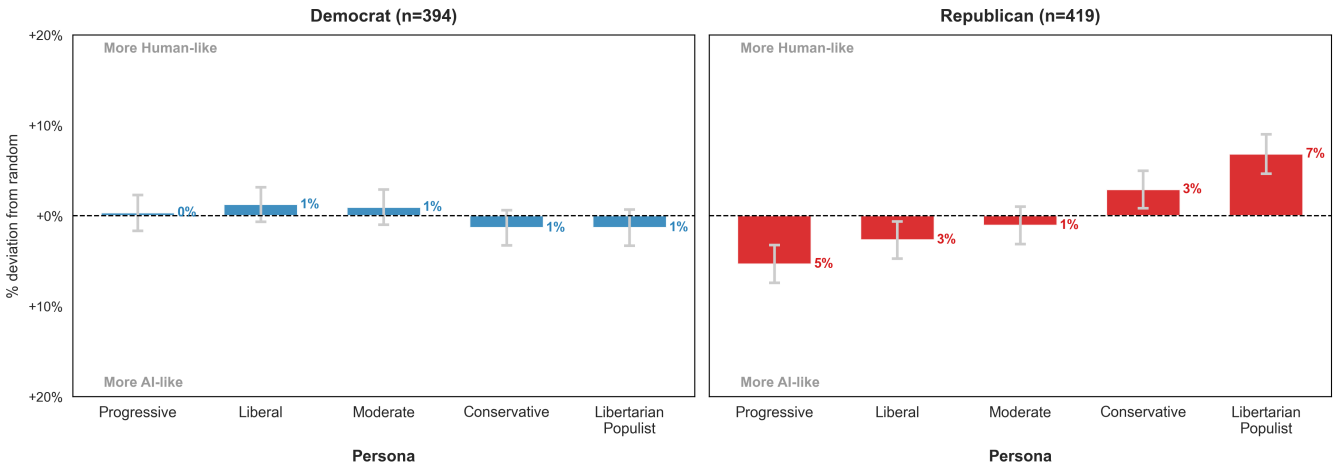
Discussion

While warnings about AIGC's threat to democratic discourse have proliferated, with 80% of countries holding elections in 2024 reporting GenAI incidents and nearly 70% of those judged harmful to election integrity (36), our study reveals a more complex reality. By examining *perceived humanness* of posts about the 2024 Harris-Trump presidential debate across multiple LLMs, social media platforms, semantic features, and individual traits, we find that humans retain substantial capacity to detect AI-generated political content. This result challenges assumptions of inevitable detection failure that have shaped recent policy debates (37, 38). However, this capacity is neither uniform nor guaranteed. Human judgment succeeds across most conditions while declining in certain circumstances, varying systematically with model architecture, platform constraints, and individual characteristics.

We first find that participants often distinguished human-authored posts from AI-generated ones, with human content rated as more human and several models (GPT-4 variants, Llama 2) judged as more AI-like. Yet newer models blurred this boundary, with Llama 3.3 frequently misclassified as human. The progression from Llama 2 (reliably detected) to Llama 3.3 (misclassified as human) occurred in only about 18 months, demonstrating that model inputs and architecture can shape perceptual distinctiveness, and that all of these patterns may shift quickly as models change. Meta's incorporation of Facebook and Instagram data into Llama's training corpus (39) further reveals how the injection of critical and relevant data could increase human-like stylistic fluency and reduce discernibility for content formats such as short social media posts. Similar patterns in UK election contexts show newer models like Llama 3 and Gemini achieving above-human levels of perceived humanness, with participants unable to distinguish AI-generated from human-written content over 50% of the time (28).

Beyond LLM variations, our results also reveal contextual variation across social media platforms and news consumption habits on these platforms. Participants were largely able to distinguish human-authored from AI-generated posts on YouTube, TikTok, and Truth Social, yet on X/Twitter these differences collapsed and judgments approached chance. This pattern is notable given X/Twitter's prominence as a venue for real-time political conversation and breaking news (40). Prior research has shown that the brevity and stylistic homogeneity of tweets make them difficult to differentiate

Fig. 4. Partisan Congruence Effects in AI-AI Pairs of Politically-Personified LLMs



along dimensions such as credibility or platform source cues (41, 42), a limitation that likely extends to AI detection.

These platform differences extend to reflect underlying variation in the semantic cues audiences use to judge authenticity. Our semantic results reveal that audiences distinguished between specific, nuanced emotional expressions and generic sentiment signals. Specific emotions, whether positive (optimism, joy) or negative (sadness), increased perceived humanness, as did norm-violating content such as offensive language and hate speech. In contrast, generic positive sentiment was associated with AI-like perception. This pattern suggests a mechanism for misclassification. Humans appear to equate vivid, situated emotional content and norm-violating language with human authorship, while stereotyping AIGC as uniformly upbeat and generically positive (43).

Individual characteristics modestly moderated detection accuracy. While men and white participants showed slightly higher accuracy than women and non-white participants, these differences operated within a context where most demographic groups demonstrated above-chance detection. Among attitudinal covariates, higher offline news consumption was associated with significantly lower accuracy in detecting AI-generated posts when paired against human-authored content. These effects echo existing patterns where individual trait asymmetries shape confidence and accuracy in evaluating the authorship of online information (28). Partisan motivation in labeling content as AI-generated appeared primarily among Republicans, who were more likely to label ideologically-aligned (Conservative or Populist) AI personas as human and ideologically-opposed (Progressive) personas as AI-generated. Such findings mirror public opinion data from the 2024 election, which show that concern about AIGC misinformation was patterned by partisanship and demographics, rather than direct exposure to synthetic content (44) and that judgments of algorithmic systems often follow trust-based heuristics rather than technical knowledge (45).

We pause to note that this study has several limitations. First, while the use of the 2024 U.S. presidential debate provided a consistent topical anchor, the focus on a single, highly salient political event prevents us from generalizing to other issue domains and lower-salience contexts. Additionally, while human-authored posts were sampled directly from the four platforms, we could not verify authorship with complete certainty. Our data collection took place (first wave in fall 2024) soon after the release of the tested LLMs (2023-2024), during a period when large-scale AI-generated activity on social media was beginning to emerge yet remained difficult to quantify. Lastly, despite a large and diverse sample, limitations of online survey experiments, such as the artificiality of pairwise comparisons evaluated through a survey platform, should temper strong claims about how judgments operate in fully naturalistic environments. Future research could expand beyond single-event contexts to examine how perceived humanness evolves across issue domains and across different platform environments with distinct stylistic conventions. Combining behavioral experiments with trace data could capture how judgments interact with engagement and spread dynamics in naturalistic settings. Beyond judgments of *perceived humanness*, future work could also test other more contextually-tailored downstream outcomes such as persuasion or trust.

The conditional nature of human's capability to correctly detect authorship necessitates interventions calibrated to specific contexts where detection fails, rather than one-size-fits-all approaches. Where human detection is weak in content from newer models like Llama 3.3 and X/Twitter-specific posts, technical interventions can be applied such as embedding watermarking when posts are generated, building tools to help social media users easily check the authenticity and veracity of political media they engage with, and encouraging platform policies to require identity verification for political accounts during electoral periods. For individual vulnerabilities, where offline news consumers show lower AI detection accuracy and demographic gaps exist, targeted media literacy for these groups, rather than generic AI awareness measures for a general-purpose audience, becomes essential. Finally, for policymakers, regulatory frameworks must remain adaptive and context-specific as model capabilities and human detection strategies coevolve.

Our findings reveal that human capacity to distinguish AI-generated from human-authored political content is substantial but conditional, succeeding across most contexts while declining systematically with newer models, platform constraints, and partisan motivations. For now, human judgment remains a viable defense in assessing content authenticity. This capacity matters not because all AIGC is malicious, but because recognizing when content may be AI-generated enables more informed evaluation of its credibility, origin, and intent. As AI outputs converge with human discourse through platform-specific training, current detection cues will likely erode. New distinguishing features may emerge or evaluation strategies may adapt, but accurate identification will increasingly require humans to draw on both content-level and source-level signals. The future of democratic discourse in AI-saturated information environments depends on preserving human agency across the full chain of political communication, from production to dissemination to consumption. At stake is not only how AI-generated political information is created, but whether citizens can recognize it as such and maintain the informed, independent judgment that democratic legitimacy requires.

Materials and Methods

We divide our methodology into four components: survey design, data collection, explanation of the perceived humanness indicator score, and analysis.

Survey design. To evaluate how individuals assess the *perceived humanness* of human-authored vs. AI-generated political social media content, we conducted two U.S. based online surveys (combined N = 1,122), centered on social media posts on the 2024 U.S. presidential debate. The first survey was launched in March 2025 (N = 559 passed all attention checks from an initial total of 601) featuring X/Twitter and YouTube content, and the second in June 2025 (N = 563 passed all attention checks from an initial total of 606) featuring Truth Social and TikTok content. The surveys were designed and deployed on Qualtrics (46) and participant recruitment and distribution occurred on Connect (47). Details on participants' demographic breakdown are included in *SI Appendix*. Both surveys received approval from Georgetown University's Institutional Review Board.[**]

Participants were asked questions related to their demographics, political identity, trust in media, conspiracy beliefs, cynicism towards the news, and news consumption habits on specific-platforms.[††] Then, participants completed 30 pairwise evaluations split into two 15-item blocks per survey wave. In each evaluation, participants were asked to indicate which of the two posts (A or B) was more likely to have been generated by an AI chatbot. Within platform, each item was paired multiple times against others to enable robust rank estimation; pair types (human–AI, human–human, AI–AI) and order were fully randomized. Across both survey waves, we collected responses to 19,500 unique pairwise comparisons using 2,386 social media posts (1,690 AI-generated, 696 human-authored). Three attention checks were embedded throughout the survey to verify annotator engagement. Only respondents passing all checks were retained. Questionnaire and attention check wordings, accompanying descriptive statistics, and details on pairwise construction are available in *SI Appendix*.

Pairwise comparisons provide robust scaling of latent perceptions. In our case, perceptions of humanness of social media posts was measured by leveraging relative rather than absolute judgments (simply put, by asking which of the two posts seems more AI-like rather than requiring an absolute rating). Recent research has also shown that pairwise comparisons reduce bias and measurement error compared to traditional annotation methods (48), and relative annotation methods lead to more reliable labels by reducing individual biases in scoring (49).

Data collection. In order to construct our pairwise evaluation task, we collected the following datasets: 1) human-authored posts from four social media platforms – X/Twitter, YouTube, Truth Social, and TikTok, and 2) AI-generated posts from six LLMs – GPT-4, GPT-4o, GPT-4o Mini, Llama 2, Llama 3.2, Llama 3.3. Table 1 shows the final count of posts collected for each platform and LLM combination used in our surveys. The content of both human-authored and AI-generated posts focus on the September 10th, 2024 presidential debate between Kamala Harris and Donald Trump.

Human-authored posts. For X/Twitter, we collected posts using hashtags *#debatenight* and *#debate2024* during a 30-hour window from September 9, 2024 at 11:57 PM to September 11, 2024 at 5:26 AM ET, capturing both real-time reactions as well as pre- and post-debate commentary. Truth Social posts were collected from 7:00 PM to 11:59 PM ET on September 10, 2024, using the same hashtags. YouTube and TikTok video comments were scraped from ten selected videos per platform discussing the debate. While we cannot guarantee that the human-authored posts collected for this study was not AI-generated or purely human-authored, we took steps to verify that the posts originated from human-operated accounts rather than bots. We also conducted automated bot detection analysis on our X/Twitter human-authored post dataset using the Botometer API () (see *SI Appendix* for details).

AI-generated posts. To emulate the human-authored posts collected above, we prompted six LLMs for our AI-generated posts dataset. We used both the GPT family of closed models (GPT-4 (50), GPT-4o (51), GPT-4o mini (52), and the LLama family of open models (Llama 3.3 (53), Llama 3.2 (53), and Llama 2 (54)). In both cases, we consider models with different numbers of parameters.

Prompting followed both a general and platform-tailored template. For all four platforms, we prompted each LLM to create 20 posts mimicking the voice and opinions of five different political personas – Progressive, Liberal, Moderate, Conservative, Libertarian Populist. Full description and prompt wording for the five political personas are in *SI Appendix*. For platform-specific guidance, X/Twitter and Truth Social prompts included the full debate transcript and LLMs were instructed to produce posts that would contain between 70-279 characters with variations in punctuations, emojis, and hashtags related to the provided debate transcript. YouTube and TikTok prompts included video transcripts that

[**] STUDY00008554: Labeling Humanness and Misinformation Data

[††] The sample was balanced across demographic categories including gender, age, race, and political affiliation to ensure diversity and representativeness. Full question wording is available in the *SI Appendix*, Table S3 - S6

Platform	Human	GPT-4	GPT-4o	GPT-4o Mini	Llama 2	Llama 3.2	Llama 3.3	Total
X Posts	203	67	67	66	56	74	60	593
YouTube Comments	153	75	75	75	73	74	75	600
TT Posts	179	75	75	75	47	76	73	600
TS Comments	161	75	72	75	73	66	71	593
Total	696	292	289	291	249	290	279	2386

Table 1. Final counts of posts/comments by platform and model

we manually collected from these platforms (more details in *SI Appendix*) and LLMs were asked to generate video comment-like posts between 5-50 characters related to the given transcripts. Additional prompt details and justifications for platform-specific instructions are in *SI Appendix*.

Data preprocessing. Both the human-authored and AI-generated posts underwent systematic cleaning to ensure consistency and user privacy protection. For all platform posts, user mentions were anonymized by replacing usernames with "@USER" tokens, and URLs were replaced with non-functional placeholder links. Language filtering was applied to retain English-only content across platforms using LangDetect (55). The AI-generated content also required additional processing to ensure authenticity and diversity. We implemented a uniqueness constraint requiring the first five words of each generated post to be distinct, preventing repetitive outputs. Malformed emoji tokens (e.g., ":happy", ":cry") generated by language models were removed to maintain realistic formatting. Additional data cleaning details are in *SI Appendix*.

A. The Perceived Humanness Indicator (PHI) Score. We converted our survey's pairwise responses into continuous humanness scores using the Bradley-Terry (BT) model (56), which estimates relative item strength from pairwise comparison outcomes. For posts i and j, the model defines the probability that annotators perceive post i as more human-like as:

$$P(i \succ j) = \frac{e^{\gamma_i}}{e^{\gamma_i} + e^{\gamma_j}} \quad [1]$$

where γ_i and γ_j are latent humanness scores estimated from all comparisons. Scores were optimized iteratively until predicted outcomes aligned with observed judgments and then normalized to $[-1, 1]$. We term these normalized scores the Perceived Humanness Indicator (PHI): the closer a post's *PHI* to -1, the more AI-like it appeared to annotators, while the closer to 1, the more human-like it appeared. This continuous measure avoids limitations of binary scoring by capturing subtle linguistic distinctions in perceived human-authorship authenticity.

B. Analysis. Using our data collection and survey responses, we conducted the following analyses.

Marginal Effects of Pooled- and Platform-Level Assessments of PHI . Using our dataset of posts and their *PHI* scores, we assessed differences by post source by estimating the marginal effects of these scores, both pooled across all platforms and separately by platform. The pooled estimates summarize the overall effect of post source on humanness ratings. Platform-specific estimates allow us to examine how the *perceived humanness* of each source varies across platforms. Together, these analyses quantify how human and different LLMs compare in terms of *perceived humanness*, both overall and within each social media contexts.

Semantic Feature Extraction and Analysis. . To investigate how semantic features influence *perceived humanness*, we applied TweetEval's language markers classifiers to all of our collected social media posts to generate binary labels indicating the predicted presence or absence of semantic features (35). These binary indicators specify whether each semantic characteristic was detected in the text. We organized semantic features across three categories: Civility (Irony, Hate, Offensive), Emotion (Joy, Optimism, Sadness), and Sentiment (Positive, Negative). *SI Appendix* contains details on our category breakdown. The semantic feature analysis takes advantage of the pairwise design since the BT model estimation yields post-level PHI scores. We ran a standardized Ordinary Least Squares (OLS) regression at the post-level predicting the *PHI* score, conditional on the semantic features of the posts. This approach allowed us to perform downstream analyses of how specific semantic expressions contribute to perceptions of humanness in social media content – regardless of their authorship.

Individual-Level Predictors of Correctly Identifying GenAI Posts. To analyze individual differences in AI detection ability, we estimated mixed-effects logistic regression models predicting the correct identification of AI-generated posts in randomized, human-AI post pair comparisons. Our dependent variable was a binary indicator of whether participants correctly identified the AI-generated post when presented within these pairs.

We estimated three model specifications: (1) demographics (age, education, gender, race, political affiliation), (2) attitudinal (news trust and consumption on/offline, news cynicism, political interest, and conspiracy beliefs), and (3) interaction models examining whether platform-specific news consumption habits affected detection accuracy differently across social media platforms. The interaction models tested whether participants' familiarity of consuming news on specific platforms (measured through news consumption frequency on Facebook, X/Twitter, TikTok, and YouTube) influenced their ability to detect AI-generated content from those same platforms. To control for baseline differences among respondents, as well as effects from the ordering of the pairs, all models employ random intercepts at the participant-level and at the order in which the participant saw a given pair in the survey (57). All continuous predictors were standardized to facilitate interpretation of effect sizes. More information on the construction of our regression and interaction models can be found in *SI Appendix*.

Partisan Effects in AI-AI Pairs of Politically-Personified LLMs. To examine whether partisan identities influenced participants' perceptions, we analyzed AI-AI post pair comparisons where participants chose between AI-posts generated with different political personas (Progressive, Liberal, Moderate, Conservative, and Libertarian Populist). We modeled the probability of participants selecting a post as AI (0/1) using a Linear Probability Model (LPM) conditional on participants' party identification (Democrat vs. Republican), the political persona assigned to the post, and their interaction. For each pair, we recorded the annotator's party identification and which persona's post was selected as more human-like.

ACKNOWLEDGMENTS. This research was supported by the Tech Public Policy (TPP) Grant at the McCourt School of Public Policy and Project Liberty. We also thank the Massive Data Institute (MDI) at Georgetown University for institutional support, including research infrastructure, staff time, and shared resources that made this work possible. The study, *Generative AI, Humanness, and Misinformation in the 2024 U.S. Presidential Election*, was funded through the TPP program. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are grateful to our team's research manager, Rebecca Vanarsdall, who manages the research project and our MDI technical team for their technical support.

1. Veda C. Storey, Wei Thoo Yue, J. Leon Zhao, and Roman Lukyanenko. Generative artificial intelligence: Evolving technology, growing societal impact, and opportunities for information systems research. *Information Systems Frontiers*, 2025. . URL https://link.springer.com/article/10.1007/s10796-025-10581-7. Open access; Published: 25 February 2025.
2. Manudeep Bhuller, Tarjei Havnes, Jeremy McCauley, and Magne Mogstad. How the internet changed the market for print media. Working Paper 2023-21, Becker Friedman Institute for Economics, University of Chicago, Chicago, IL, February 2023. URL https://bfi.uchicago.edu/wp-content/uploads/2023/02/BFI_WP_2023-21.pdf.
3. Leonardo Banh and Gero Strobel. Generative artificial intelligence. *Electronic Markets*, 33 (1):1–17, 2023. . URL https://doi.org/10.1007/s12525-023-00680-1.
4. Michael Wessel, Martin Adam, Alexander Benlian, Ann Majchrzak, and Ferdinand Thies. Generative AI and its transformative value for digital platforms. *Journal of Management Information Systems*, 42(2):346–369, 2025. . URL https://doi.org/10.1080/07421222.2025.2487315.
5. Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), March 2024. ISSN 2157-6904. . URL https://doi.org/10.1145/3641289.
6. Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. The generative ai paradox: "what it can create, it may not understand". *arXiv*, oct 2023. . URL https://arxiv.org/abs/2311.00059.
7. Nic Newman, Richard Fletcher, Craig T Robertson, A Ross Arguedas, and Rasmus Kleis Nielsen. *Reuters Institute digital news report 2024*. Reuters Institute for the study of Journalism, 2024.
8. Yiluo Wei and Gareth Tyson. Understanding the impact of ai-generated content on social media: The pixiv case. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 6813–6822, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. .
9. Victoria Asbury-Kimmel, Keng-Chi Chang, Katherine T. McCabe, Kevin Munger, and Tiago Ventura. The effect of streaming chat on perceptions of political debates. *Journal of Communication*, 71(6):947–974, 12 2021. . URL https://doi.org/10.1093/joc/jqab041.
10. Shane Goldmacher. A campaign aide didn't write that email. a.i. did. *The New York Times*, March 2023. URL https://www.nytimes.com/2023/03/28/us/politics/artificial-intelligence-2024-campaigns.html.
11. Holly Ramer. Consultant on trial for ai-generated robocalls mimicking biden says he has no regrets. AP News, 2025. URL https://apnews.com/article/new-hampshire-election-ai-robocalls-ce944b35271ff2c6df39a434d04d4212.
12. Giulio Corsi, Bill Marino, and Willow Wong. The spread of synthetic media on x. *Harvard Kennedy School (HKS) Misinformation Review*, June 2024. . URL https://misinforeview.hks.harvard.edu/article/the-spread-of-synthetic-media-on-x/.
13. Stuart A. Thompson. How president trump uses a.i., October 22 2025. Retrieved from https://www.nytimes.com/2025/10/22/briefing/how-president-trump-uses-ai.html.
14. Nesrine Malik. With 'AI slop' distorting our reality, the world is sleepwalking into disaster. The Guardian, April 2025. URL https://www.theguardian.com/commentisfree/2025/apr/21/ai-slop-artificial-intelligence-social-media. Opinion/Comment.
15. Jingnan Huo. Ai images of hurricanes and disasters are being weaponized as propaganda. https://www.npr.org/2024/10/18/nx-s1-5153741/ai-images-hurricanes-disasters-propaganda, October 2024. Accessed: 2025-09-07.
16. Alexios Mantzarlis and Nasha Dutta. Tech platforms promised to label ai content. they're not delivering, October 23 2025. URL https://indicator.media/p/tech-platforms-fail-to-label-ai-content-c2pa-metadata. An Indicator audit of AI content labeling practices across major social media platforms.
17. Chloe Wittenberg, Ziv Epstein, Adam J. Berinsky, and David G. Rand. Labeling ai-generated content: Promises, perils, and future directions. Topical policy brief, MIT AI Policy Forum, MIT Schwarzman College of Computing, Cambridge, MA, November 2023. URL https://computing.mit.edu/wp-content/uploads/2023/11/AI-Policy_Labeling.pdf.
18. Edson C. Jr. Tandoc, Jia Yao Lim, and Shangyuan Wu. Man vs. machine? the impact of algorithm authorship on news credibility. *Digital Journalism*, 8(4):548–562, 2020. . URL https://www.tandfonline.com/doi/abs/10.1080/21670811.2020.1762102.
19. Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 7282–7296. Association for Computational Linguistics, 2021. . URL https://www.aclweb.org/anthology/2021.acl-long.565/.
20. Nils C. Köbis and Lilian D. Mossink. Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in Human Behavior*, 114:106553, 2021. ISSN 0747-5632. . URL https://doi.org/10.1016/j.chb.2020.106553.
21. Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11): e2208839120, 2023. . URL https://www.pnas.org/doi/10.1073/pnas.2208839120.
22. Alex Wasdahl. Machine credibility: How news readers evaluate ai-generated content. *InterActions: UCLA Journal of Education and Information Studies*, 19(1), September 2024. . URL https://escholarship.org/uc/item/59p6r0tn.
23. Jason Weismueller, Paul Harrigan, Kristof Coussement, and Tina Tessitore. What makes people share political content on social media? the role of emotion, authority and ideology. *Computers in Human Behavior*, 129:107150, 2022. . URL https://www.sciencedirect.com/science/article/pii/S0747563221004738.
24. Susan Bickford. Emotion talk and political judgment. *The Journal of Politics*, 73(4): 1025–1037, October 2011. . URL https://doi.org/10.1017/S0022381611000740.
25. R. M. Schuetzler, G. M. Grimes, and J. Scott Giboney. The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems*, 37(3):875–900, 2020. . URL https://www.tandfonline.com/doi/full/10.1080/07421222.2020.1790204?utm_source=researchgate.net&medium=article.
26. Eun Go and S. Shyam Sundar. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97: 304–316, 2019. .
27. Natalie Jomini Stroud. Media use and political predispositions: Revisiting the concept of selective exposure. *Political Behavior*, 30(3):341–366, 2008. . URL https://link.springer.com/article/10.1007/s11109-007-9050-9.
28. Angus R Williams, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. Large language models can consistently generate high-quality content for election disinformation operations. *PloS one*, 20(3):e0317421, 2025. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0317421.
29. Sarah Kreps, R. Miles McCain, and Miles Brundage. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117, 2022. . URL https://www.cambridge.org/core/journals/journal-of-experimental-political-science/article/abs/all-the-news-thats-fit-to-fabricate-aigenerated-text-as-a-tool-of-media-misinformation/40F27F0661B839FA47375F538C19FA59.
30. Yuxia Wang, Rui Xing, Jonibek Mansurov, Giovanni Puccetti, Zhuohan Xie, Minh Ngoc Ta, Jiahui Geng, Jinyan Su, Mervat Abassy, Saad El Dine Ahmed, Kareem Elozeiri, Nurkhan Laiyk, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Ryuto Koike, Masahiro Kaneko, Artem Shelmanov, Ekaterina Artemova, Vladislav Mikhailov, Akim Tsvigun, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. Is human-like text liked by humans? multilingual human detection and preference against ai. *arXiv preprint arXiv:2502.11614*, 2025. . URL https://arxiv.org/abs/2502.11614. Revised May 23, 2025; originally submitted February 17, 2025.
31. Galen Stocking, Luxuan Wang, Samuel Bestvater, and Regina Widjaya. How News Influencers Talked About Trump and Harris during the 2024 Election, February 2025. URL https://www.pewresearch.org/short-reads/2025/02/06/how-news-influencers-talked-about-trump-and-harris-during-the-2024-election/.
32. Peter John Loewen, Daniel Rubenson, and Arthur Spirling. Testing the power of arguments in referendums: A bradley–terry approach. *Electoral Studies*, 31(1):212–221, 2012.
33. David Carlson and Jacob M Montgomery. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review*, 111 (4):835–843, 2017.
34. Chandni Maggo and Puneet Garg. From linguistic features to their extractions: Understanding the semantics of a concept. In *Proceedings of the 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, pages 427–431. IEEE, July 2022. .

35. Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1644–1650, 2020.
36. International Panel on the Information Environment. Press release: Global trend of widespread role of genai in national elections in 2024, May 2025. URL https://ipie.webflow.io/news/press-release-global-trend-of-widespread-role-of-genai-in-national-elections-in-2024. Accessed 2025-09-09; Press release summarizing the technical report "The Role of Generative AI Use in 2024 Elections Worldwide: Searching for Solutions.".
37. Jaiv Doshi, Ines Novacic, Curtis Fletcher, Mats Borges, Elea Zhong, Mark C. Marino, Jason Gan, Sophia Mager, Dane Sprague, and Melinda Xia. Sleeper social bots: a new generation of ai disinformation bots are already a political threat. *arXiv*, August 2024. . URL https://arxiv.org/abs/2408.12603.
38. Sarah Kreps and Doug Kriner. How ai threatens democracy. *Journal of Democracy*, 34(4): 122–131, October 2023. . URL https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy/.
39. Laura Tingle. Meta ai confirms it uses facebook and instagram posts to train its models, including australian social media data, July 17 2025. URL https://www.theguardian.com/australia-news/2025/jul/17/meta-ai-facebook-instagram-personal-information-social-media-posts-learn-australian-concepts.
40. Jocelyn McCullough. Musk's saga reveals how core twitter is to u.s. politics. Axios, April 15 2022. URL https://www.axios.com/2022/04/15/musks-twitter-bid-political-implications. Accessed: 2025-09-09.
41. Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. .
42. Gordon Pennycook and David G. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019. . URL https://www.pnas.org/doi/10.1073/pnas.1806781116.
43. Jason M. Chein, Sofia A. Martinez, and Anthony R. Barone. Human intelligence can safeguard against artificial intelligence: individual differences in the discernment of human from ai texts. *Scientific Reports*, 14(1):25989, oct 2024. . URL https://www.nature.com/articles/s41598-024-76218-y.
44. Hanyang Yan, Jennifer Jones, and Young Mie Kim. The origin of public concerns over ai supercharging misinformation in the 2024 u.s. presidential election. *Harvard Kennedy School (HKS) Misinformation Review*, 6(2), 2025. . URL https://misinforeview.hks.harvard.edu/article/the-origin-of-public-concerns-over-ai-supercharging-misinformation-in-the-2024-u-s-presidential-election.
45. Y. Wang. Factors related to user perceptions of artificial intelligence (ai)-based content moderation on social media. *Computers in Human Behavior*, 148:107799, 2023. ISSN 0747-5632. . URL https://doi.org/10.1016/j.chb.2023.107799.
46. Qualtrics. Qualtrics. https://www.qualtrics.com/, 2025. Accessed October 28, 2025.
47. Connect Research Platform. Connect: Rapid Data Collection for Behavioral Science Research, 2024. URL https://connect.georgetown.edu. Accessed April 2, 2025.
48. Hasti Narimanzadeh, Arash Badie-Modiri, Iuliia G. Smirnova, and Ted Hsuan Yun Chen. Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), October 2023. . URL https://doi.org/10.1145/3610183.
49. Lise De Bruyne, Orphee De Clercq, and Veronique Hoste. Annotating affective dimensions in user-generated content. *Language Resources and Evaluation*, 55:1017–1045, 2021. . URL https://doi.org/10.1007/s10579-020-09524-2.
50. OpenAI. Gpt-4 technical report. https://openai.com/research/gpt-4, 2023. Accessed March 2025.
51. OpenAI. Gpt-4o: Openai's most advanced multimodal model. https://openai.com/index/gpt-4o, 2024. Accessed March 2025.
52. OpenAI. Gpt-4o mini. https://platform.openai.com/docs/models/gpt-4o-mini, 2024. Accessed: 2025-10-23.
53. Meta AI. Llama 3: Next-generation open-weight language models. https://ai.meta.com/blog/meta-llama-3/, 2024. Accessed March 2025.
54. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama 2: Open foundation and fine-tuned chat models. https://ai.meta.com/llama/, 2023.
55. Nakatani Shuyo. Language detection library for java. https://github.com/shuyo/language-detection, 2010. Accessed March 2025.
56. Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.
57. Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2007.