Human-authored or AI-generated? Humans show limited ability to tell the difference for political content on social media

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

This manuscript was compiled on August 13, 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 (LLM), the model architecture, the social media platform, and the political personas. Participants judged randomly-paired posts on which they perceived to be AI-generated. We find that human-authored posts were on average perceived as more human-like than AI outputs generated from five of the six LLMs in this study. However, their ability to detect AI is limited. They do so only slightly above chance in human-AI pairs, and this ability weakens for newer LLMs, varies by platform, and reflects partisan bias. Only posts generated with Llama 3.3 scored, on average, higher on perceived humanness than human-authored posts. The social media platform also mattered, with participants distinguishing human from AI on most platforms (YouTube, TikTok, Truth Social), but not on X/Twitter. Incivility and negative emotions are associated with high humanness perceptions, and partisan identity also shaped evaluations, with Republican respondents more likely to view ideologically mis-aligned AI posts as AI-generated.

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. Using a set of within-subjects online behavioral experiments, we demonstrate that humans show a limited ability to correctly identify AI content on social media, performing only slightly above chance. This ability is shaped by model development, platform context, linguistic features, and motivated reasoning. These boundaries have real-world implications, as they reveal that human judgment serves only as a limited defense against manipulated AI-generated political content. Furthermore, our findings offer practical insights into where human judgment fails rather than assuming universal detection failure, with implications for building democratic safeguards that support informed political judgment in AI-saturated information environments.

The rise of generative artificial intelligence (GenAI) represents a turning point for information systems in democracies worldwide (1, 2). Previous technological shifts, such as the introduction of the internet and mobile technologies, reshaped how information was distributed and accessed (3, 4). GenAI differs in a fundamental way. Enabled by large language models (LLMs), AI-generated texts that resemble human expression in their tone, syntax, and affective style (5) can now be artificially mass-produced at a scale not previously possible (6), potentially making it difficult for humans to distinguish authentic political discourse from synthetic content.

Both human-authored and AI-generated content (AIGC) may potentially circulate side by side on social media, which now serves as a dominant venue for online information consumption (7). 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 communication, it may erode trust or increase misinterpretations in democratic discourse (8). Yet whether audiences actually lack the capacity to detect AIGC, or rather face specific, identifiable vulnerabilities under certain conditions, remains largely unknown.

Media reports from the 2024 U.S. presidential election cycle describe how these hypothesized concerns can manifest. LLMs scaled the production of campaign materials (9), deepfake phone calls imitating candidates' voices were disseminated to voters (10), and AI-generated images and videos of political figures circulated on social media platforms (11) used to mock opponents or blur distinctions between satire and propaganda. Some argue that the spread of low-quality AIGC, or "AI slop" (12), contributed to emergent forms of misinformation. These incidents emphasize the urgency of understanding humans' detection and perception capacity, particularly as platform-based solutions prove inadequate. AI labeling practices struggled to scale (13) and their effectiveness remains contested (14). Where platforms cannot reliably identify AIGC, audience judgments serve as a critical detection mechanism for navigating authenticity. How well this works likely varies across contexts, potentially affected by the nature of the content, the platform environment, and the characteristics of the audience making the judgment.

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. and T.V. directed the project end-to-end; T.V., L.B., and L.S. conceptualized the project; S.P., T.V., R.A., L.B., A.T., and L.S. authors 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

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-authored content from AI-generated posts created using different LLMs across different social media platforms? (2) What emotive features influence people's evaluation of 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 the following core research threads. The first focuses on understanding how people distinguish AI-generated from human content in political contexts. They struggle to distinguish AI from human text (15, 16), often relying on surface cues like first-person pronouns or warm tones (17), and semantic cues signaling credibility or coherence (18). In political contexts, emotional tone, stylistic fluency, and conversational style serve as powerful cues of legitimacy and persuasiveness (19, 20). As such, AIGC mimicking these markers may evade detection while shaping political attitudes (21). 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 (22). Third, we examine partisan congruence effects. When AIGC adopts political personas, alignment between audience and content ideology influences authorship judgments (23), with politically congruent content more likely to be perceived as human-authored.

This study comprises two within-subject online behavioral experiments, in which participants completed multiple pairwise comparison tasks (combined, unique respondents: 1,122 and 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 (24). The dataset that we draw the pairs from consists of two components. Human-authored posts were collected directly from four social media platforms (X/Twitter, YouTube, TikTok, and Truth Social).[1] Specifically, we analyze the text of posts from X/Twitter and Truth Social, and comments from YouTube and TikTok video posts.[2] AI-generated posts were produced by the authors via controlled prompting of six LLMs (GPT-4, GPT-4o, GPT-4o mini, Llama 2, Llama 3.2, Llama 3.3). These posts were designed to imitate discourse across the four platforms and took on five political personas (Progressive, Liberal, Moderate, Conservative, Libertarian Populist). Using a pairwise design, participants evaluated which of two posts they perceived as more human-like.

We quantified perceptions of humanness by scaling pairwise survey responses about which social media post was more likely to be AI-generated, using the Bradley–Terry (BT) statistical model. The resulting BT scores were then converted 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). Recent work confirms that aggregating pairwise comparisons yields better calibrated scalar measurements than direct scoring (26). This approach has also been widely used in social science to measure argument persuasiveness, ideology, and textual complexity (27–29). This design enabled us to analyze how perceptions of humanness differed across model outputs, platform context, emotive 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 show limited ability to distinguish AI-generated from human-authored political content. Pooling all models together, *perceived humanness* is higher and statistically significant when comparing human-authored posts and AI-generated content. However, this finding has some important caveats. Identifying AI-generated content for political social media content is far from easy, as the aggregated detection rate, i.e., when participants identify an AI post as AI-generated when paired with another human post, is on average 58.6%, only slightly above chance. Across models, for five out of six frontier AI models, *perceived humanness* is higher for Human posts compared to AI models. Posts from Llama 3.3 are the only AI-generated content more likely to be classified as human, even when compared to human posts, indicating that discernment declines for these larger, newer LLMs. Across platforms, AI detection rates vary. Participants are most accurate on TikTok and least accurate on X/Twitter, where perceived humanness does not differ significantly between human-authored and AI-generated posts. Certain semantic features related to civility, emotion, and sentiment increase *perceived humanness*. Individual differences and media habits also correlate with better detection, with men and white participants showing higher accuracy in identifying AI-generated posts than women and non-white participants, those who consume more news offline showing worse detection, and partisan identity leading participants to 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) emotive language features on humanness judgment, (3) individual traits that predict

[1] A potential concern is whether human-authored posts contain AI-generated content. Automated detection of AI-generated text in short-form social media content remains a largely unsolved problem (25). We therefore acknowledge that we cannot fully verify the human origin of all collected posts, even though the data is collected directly from the platforms. To assess the presence of artificial accounts in our data collection, we examined the authors of our posts on X/Twitter for bot activity using the Botometer API. 2 of 79 evaluated accounts exceeded the threshold; all posts were retained. See *SI Appendix* S2.1 for details.

[2] For consistency, we refer to all content as posts throughout the remainder of the manuscript.

detection capacity, and (4) political partisan alignment on humanness perceptions of AI-generated posts.

Individuals' ability to distinguish 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. The scores are built using a Bradley-Terry model based on responses from all pairwise tasks. We estimate these scores per platform, and normalize them to range from −1 to 1; the closer a post's *PHI* score is to -1, the more AI-like it appeared to survey respondents, while the closer it is to 1, the more human-like it appeared. Figure 1A (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 95% confidence intervals (CIs). The comparison between marginal means across post sources allows for the relative identification of which group of posts was, on average, consistently considered AI-generated in our surveys.[3]

Descriptively, human-authored posts have a positive average *PHI* score; therefore, they are judged more often in the pairwise task as posts produced by humans compared to the median AI post in our distribution ($M = 0.12$, 95% CI [0.08, 0.16]). Outputs from four (GPT-4, GPT-4o, GPT-4o mini, Llama 2) out of six LLMs have negative marginal means, judged more frequently as AI-like compared to the median AI post, 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]), while Llama 3.2 is centered around the median ($M = 0.026$, 95% CI [-0.039, 0.090]). Posts generated by Llama 3.3 were the exception and the only among the AI-generated posts with a positive average *PHI* score ($M = 0.23$, 95% CI [0.16, 0.29]). Most importantly, when combining all AI-generated posts, the difference in *PHI* score between human and AI-generated posts is negative and statistically significant at conventional levels (β = -0.169, $p < 0.05$). Per model differences, the average differences between the *PHI* scores of Human posts and five other models are statistically significant (Human vs GPT-4, GPT-4o, GPT-4o mini, Llama 2, and Llama 3.2). This indicates that participants in our experiment consistently perceived human-authored posts as being more human-like compared to posts from these five models. The exception here is Llama 3.3, which participants rated as more human-like than posts from the platforms. Llama 3.3's posts have a statistically significantly higher average *PHI* score than the marginal means for the human-authored posts and all other models. In the SI Appendix, we present the inferential tests comparing human posts with each of the six AI models (SI Appendix Table S12) and pooled comparisons between human-authored and all AI-generated posts (SI Appendix Table S13).

We now describe how *PHI* scores varied across the social media platforms. Figure 1B (Platform Estimates) presents unpooled marginal means of the *PHI* scores across four different platforms – X/Twitter, Youtube, Truth Social, TikTok. Three of the four platforms displayed noticeable separation in participants' perceptions about the humanness of human-authored or AI-generated posts. Scores indicate YouTube showed the clearest distinction (See SI Appendix Table S13). Human-authored posts were judged as significantly more human ($M = 0.13$, 95% CI [0.04, 0.22]), while most LLM-generated posts were consistently judged as more AI-like, especially posts from GPT-4o (M = -0.17, 95% CI [-0.29, -0.04]) and Llama 2 (M = -0.17, 95% CI [-0.30, -0.04]). TikTok and Truth Social scores showed similar patterns as YouTube. Scores for X/Twitter showed the weakest differentiation between Human and AI-generated posts. Among X/Twitter model outputs, we observe no statistically significant difference when comparing human posts to most AI models, except for GPT-4o. In sum, our unpooled results are robust across most platforms. YouTube, TikTok, and Truth Social demonstrated a detectable human-AI separation with the majority of AI models. X/Twitter was a notable anomaly with minimal reliable discernment. The platform results hold when considering pooled effects across all AI-generated posts (SI Appendix, S12).

To assess the robustness of our findings, we also measure the ability to identify AI content considering responses' accuracy only in human-AI pairings (SI Appendix Figures S5, S6). We call this measure AI detection rate, i.e., when participants correctly identify an AI post as AI-generated, and correctly identify a human-authored post as not AI-generated, when considering pairs that contain one AI post and one human post. Across all platforms and all models, the AI detection rate is on average 58.6% (95% CI [57.8,59.3]). Participants have the highest accuracy rate when classifying TikTok posts, with 64.8% (95% CI [63.3,66.3]). The detection rate for the other platforms is as follows: 58.7% (95% CI [57.2,60.2]) for Truth Social, 57.1% (95% CI [55.5,58.5]) for Youtube, and only 54.2% (95% CI [52.8,55.7]) for X/Twitter posts. We also use this measure to validate the *PHI* scores. SI Appendix Figure S7 shows that in the lower/upper bounds of the *PHI* scores, participants also show the highest (when *PHI* score = -1) and lowest (when *PHI* score = +1) AI detection rate. In conclusion, the AI detection rate aligns with findings from the *PHI* scores, with both measures showing participants distinguish human from AI content, at rates above chance.

Emotive language features of *perceived humanness* at the post-level. Because emotions are fundamental to the human experience, are expressed through natural language in measurable ways, and play a prominent role in political communication on social media (19), we investigate their relationship with *perceived humanness*. Prior research highlights that emotional expression differentiates human-authored from AI-generated text, indicating that emotive language may function as a cue in perceptions of whether content was authored by a human (30). We use the gold standard benchmark for emotive expression classification, TweetEval (31), implemented with RoBERTa as the language model classifier due to its high performance.[4] For every post, we classify whether the post contains the following three features: Civility (Irony, Hate, Offensive), Emotion (Joy, Optimism, Sadness), and Sentiment (Positive, Negative).

[3] Our sample contains more AI-posts than human posts. Therefore, the zero of the PHI scores represents the median post of our distribution, and falls inside the AI distribution. This point does not represent a perceptual boundary between human and AI posts. For this reason, our analysis focuses on the extent to which the marginal means of human-authored posts differ from those of AI-generated posts.

[4] Details on the TweetEval framework and gold-standard benchmarks are provided in the *Materials and Methods*

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

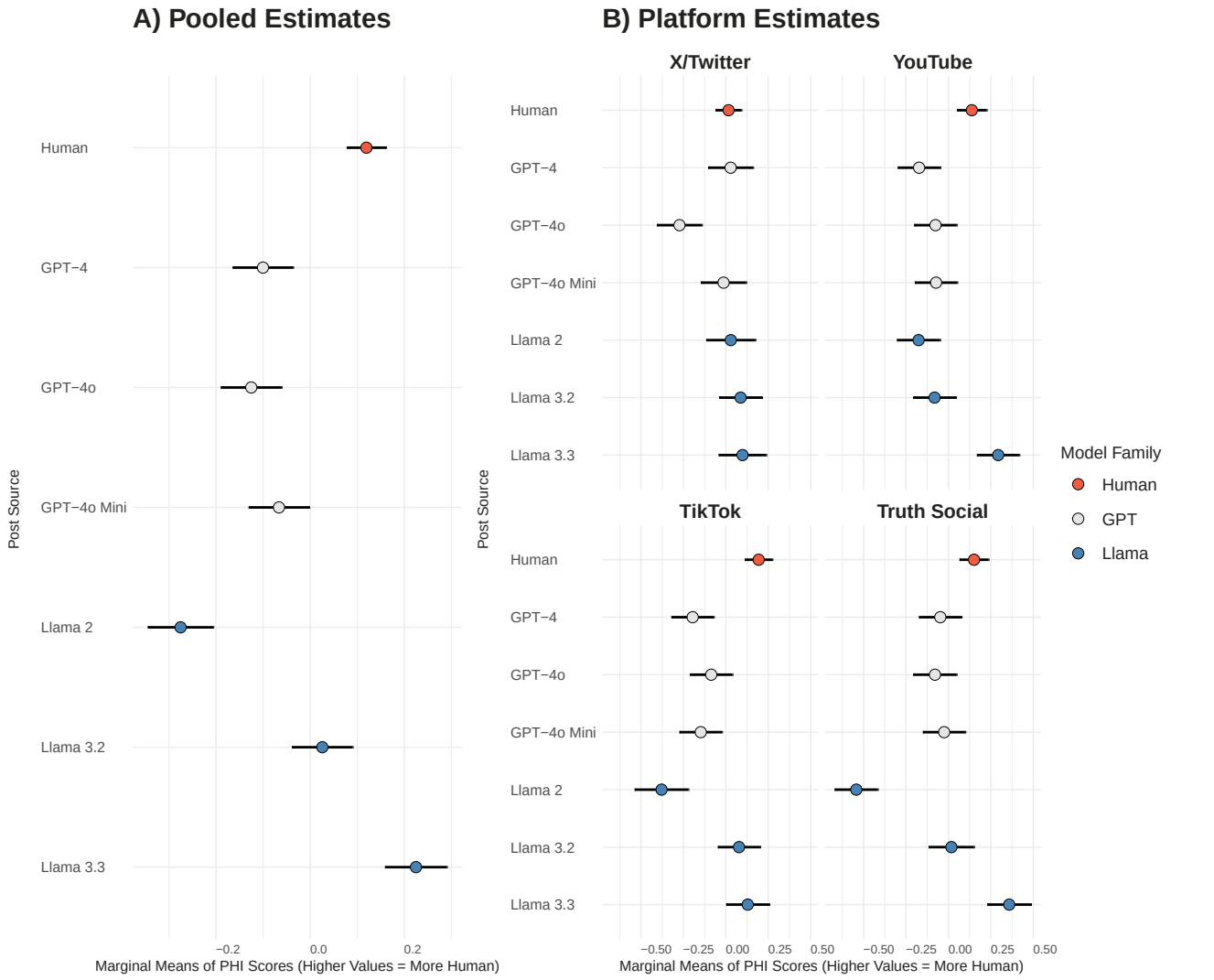
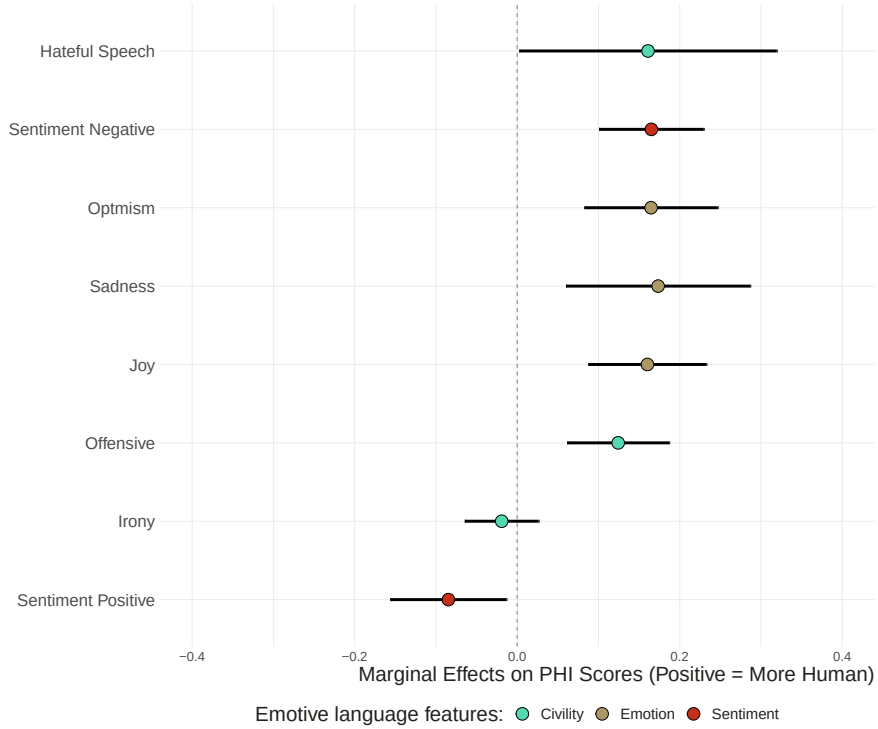


Figure 2 presents the marginal effects for each emotive feature with corresponding 95% confidence intervals. Results show that both offensive language ($\beta = 0.12$, 95% CI [0.05, 0.18]) and the presence of hateful speech ($\beta = 0.17$, 95% CI [0.01, 0.33]) were significant positive predictors of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

In the SI materials, we carried out the same emotive expression analysis for each platform. We found that the pattern of coefficients was broadly consistent with these aggregated results (SI Appendix Figure S12). We also consider heterogeneity of the marginal effects for emotive language features conditional on post source type (human or AI) and find no statistically significant variation (SI Appendix Figure S10). Lastly, we examine measurement stability using an alternative classifier. Figures S8 and S9 in the SI Appendix present results using binary and continuous probability labels from our primary classifier (TweetEval, implemented via the TweetNLP library (32)) and from BERTweet (33); results are robust to classifier choice.

Individual traits that influence ability to identify AI-posts. We further examine how individual characteristics of the respondents predict the ability to correctly identify AI-generated posts when paired against human-authored posts. For this analysis, we filtered to include only pairs containing one human-authored and one AI-generated post. 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-

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



AI pair comparisons. Our dependent variable was a binary indicator of whether participants correctly identified the AI-generated post when presented within these pairs. Point estimates are converted to average marginal effects (AME) in the probability scale.[5]

With regard to demographics, only gender and race were significant predictors. Males were more likely than females to correctly identify AI posts *(AME* = 0.021*,* 95% CI [0.00, 0.04]*)*, and White participants likewise showed a positive effect *(AME* = 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 *(AME* = -0.015, 95% CI [-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). Trust in news (online and offline), online news consumption, news cynicism, interest in politics, and conspiracy beliefs were not statistically significant, 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 *(AME* = 0.016, 95% CI [0.002, 0.033]), 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. We note the effect sizes are small, with the largest individual-level difference at approximately three percentage points, and this analysis should be treated as exploratory rather than as a test of pre-registered hypotheses.

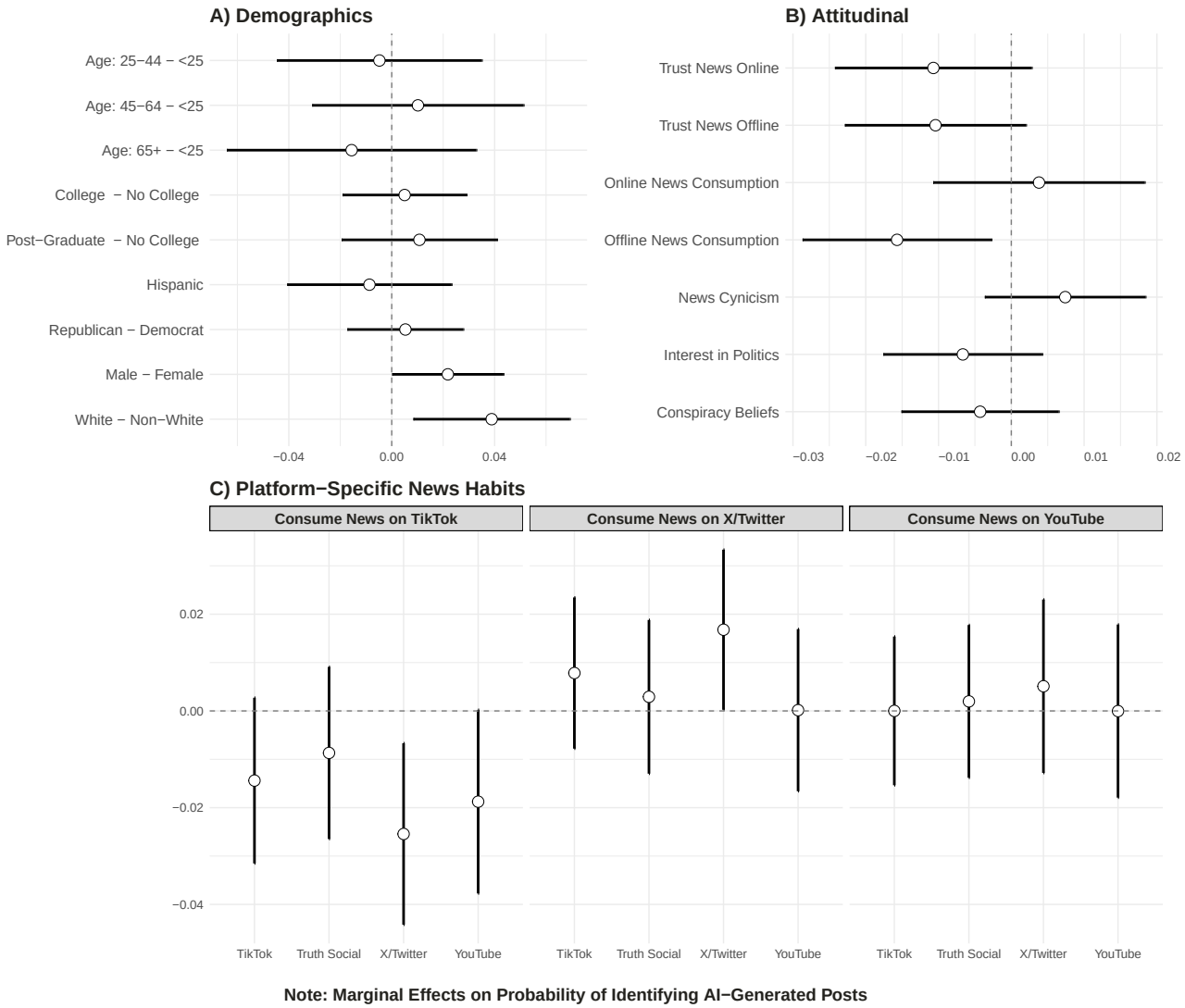
Political personification of LLMs and participants' partisan alignment. This section explores whether partisan congruence between respondents and the political persona of AI-generated posts shapes humanness judgments. As in the previous section, we model the user's decision using a mixed-effects logistic regression, with the outcome being the post selected by the respondent. We restrict this analysis only to AI-AI pairs to parse out accuracy motivations from partisan bias in the AI attribution.[6]

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 5.7 percentage points (*AME*=0.057, 95% CI [0.028, 0.085]) more likely to identify a progressive post as AI-generated compared

[5] The individual-traits are exclusively used for the human-AI pairings because this is the only pair type in which a ground-truth correct classification can be defined. In the human/human and AI-AI pairs, both posts share the same authorship category and there is no objectively correct answer, so accuracy cannot be operationalized.

[6] In human-AI pairs, a respondent's selection of an AI post as more AI-like could be either due to two distinct cognitive processes: accuracy goals (identifying the correct response) or a consequence of directional (partisan) motivations (34). In these pairs, the two cognitive mechanisms cannot be separated because an accurate response might be congruent to the respondent's partisan identity. Filtering the data only to AI-AI pairs allows us to remove accuracy as a competing explanation. With both posts being AI-generated, any systematic responses conditional on partisan identity must reflect partisan congruence instead of respondents' acting to correctly answer the survey tasks.

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



Note: Marginal Effects on Probability of Identifying AI−Generated Posts

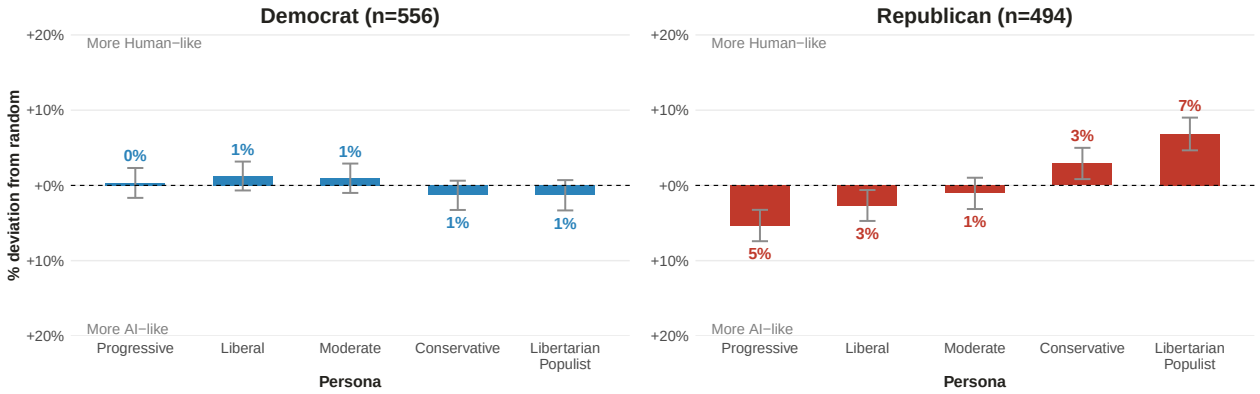
to Democrats, and 8.1 percentage points (AME=-0.081, 95% CI [-0.111, -0.052]) and 4.2 percentage points (AME=-0.042, 95% CI [-0.071, -0.014]) less likely to identify a post from a Libertarian Populist and Conservative personas, respectively, as AI-generated. We did not observe strong partisan effects among Democrats. Figure 4 presents these partisan differences as deviation from random chance.

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 (35), our study reveals a more complex reality. By examining *perceived humanness* of posts about the 2024 Harris-Trump presidential debate, we find that humans retain some capacity to detect AI-generated political content. However, there are important nuances to this finding that we revisit in this section.

Using the *PHI* scores, we show that participants in our sample are, on average, able to judge human-authored social media content as more human-like than AI-generated content. This difference holds across most platforms and models and is statistically significant but substantively small. When considering the detection rate for only human-AI pairs, participants correctly identified an AI-generated post for 58.6% of the pairs, with some substantial variation across platforms. We benchmark these results with two additional data sources. First, we consider a recent meta-analysis looking at accuracy rates across a large sample of misinformation studies. In this task, which is similar to ours, studies report a higher accuracy rate (67.8%) for false-news headline detection (36). Second, we benchmark our results with a zero-shot

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



prompting [7] analysis using GPT-5 (SI Appendix, Figure S13). It outperforms humans with an accuracy rate of 66.9%. These comparisons provide some interesting insights. Detecting AI-generated content specifically designed to mimic human social media posts appears more challenging than identifying online misinformation for the general population, which justifies broader concerns about AI use in politics. Additionally, employing a general-purpose LLM can indeed help humans identify this content at scale, with an average detection rate more than eight percentage points better than human detection. Furthermore, the zero-shot prompt analysis performs best on X/Twitter, precisely where human detection rate is lowest in our experiments, suggesting AI-detection can indeed complement human judgment. Lastly, we show that human cognition serves only as a partial safeguard against a flood of AI content on social media, statistically distinguishing posts based on their perceived humanness, but with a small gap, and performing only slightly above chance on direct detection measures. Therefore, to protect the integrity of the informational environment, human cognition should be complemented with other detection mechanisms. We encourage more work on AI-based detection models fine-tuned using short text from social media platforms. Our results from zero-shot GPT-5 show their potential. At the same time, well designed fine-tuned models could be used to more effectively manipulate humans to impact future elections, highlighting the tension between detection tools and deception ones.

Our findings also show that 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 changes can rapidly shift perceptions of humanness. One possible explanation, although untestable with our data, is that Meta's incorporation of Facebook and Instagram data into Llama's training corpus (37) improved the model's generation of human-like social media content. Similar patterns in UK election contexts show newer models like Llama 3 and Gemini achieving above-human levels of perceived humanness when generating election disinformation, with participants unable to distinguish AI-generated from human-written content over 50% of the time (22). Future work in this area should examine how changes in training data and model architecture affect the humanness of AIGC, particularly in politics and social media spaces.

Beyond LLM variations, our results also reveal contextual variation across social media platforms. Participants were 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 (7). One post-hoc interpretation of X/Twitter's results relates to the presence of well-known stylistic markers in its posts, such as hashtags, user tags, RTs (retweets), and others. Because our stimuli were constrained to roughly similar lengths across platforms (SI Appendix S3), we believe the presence of these marks in our synthetic AI posts hides an opportunity for respondents to identify more predictable word choices and sentence structures well-known to appear in AI-generated content (38). In the supplemental materials (SI Appendix S3), we provide exploratory evidence for this argument. We show participants judge longer posts as more likely AI-generated (SI Appendix, Figure S2). However, this effect primarily comes from substantive text after we remove characters that could be interpreted as platform markers. Furthermore, this effect comes mostly from AI writing more text (SI Appendix, Figure S2, *Standardized Effect of Substantive Text Length* β=-0.263, 95% CI [-0.321, -0.206]). In other words, the more an AI model writes substantive text, the easier it is for humans to detect. Twitter platform-specific markers, some of them native to the platform, crowd out this effect, and detection becomes harder.

We also show interesting variation in the emotive cues participants use to judge humanness. Our post-level emotive features results reveal that audiences distinguished between specific, nuanced emotional expressions and generic sentiment

[7] At the time of the experiments, GPT-5 had not been released; therefore, this model was not used in the generation of AI posts. Zero-shot prompting means we use the model off the shelf as a general purpose tool. We do not fine-tune the model with training examples and we do not add additional examples of correct responses in the prompt itself.

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 (39).

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 SI Appendix S8, we present suggestive evidence on the mechanisms behind these demographic differences. We fielded a follow-up survey (N = 723; a 64% recontact rate)[8] to enrich our data with questions absent from the original instrument. We regress AI usage, timing of AI adoption, breadth of AI use, digital literacy, and the attitudinal and media-consumption measures from Figure 3B on a range of participants' demographic information (SI Appendix, Table S17). Our results indicate men report adopting AI earlier (β = 0.218, 95% CI [0.074, 0.362]), use it more often (β = 0.342, 95% CI [0.088, 0.597]), and score higher on digital literacy compared to women (β = 0.467, 95% CI [0.326, 0.608]). White respondents go the opposite direction on reported AI usage, with statistically significantly lower reported AI usage than Non-White respondents (β = -0.441, 95% CI [-0.817, -0.065]), and show no digital literacy advantage (β = -0.128, 95% CI [-0.336, 0.081]). However, their media diets differ, with White respondents relying significantly less on offline news than non-white respondents (β = -0.499, 95% CI [-0.660,-0.338]), a significant predictor of AI detection (see Fig. 3B). The two advantages associated with these groups appear to run through different channels; higher AI detection for men seems to emerge from stronger AI adoption. As expected, stronger AI usage in the follow-up survey consistently improves AI detection rate (β = 0.138, 95% CI [0.061, 0.214], SI Appendix Table S18). However, for White respondents, the effects emerge from their news diet and lower reliance on offline news media, given that offline news consumption is the only predictor with a negative statistically significant effect in AI detection rate (Figure 3B). We emphasize these results as suggestive evidence, given that our sample is not well-powered to precisely estimate subgroup differences, and part of the evidence comes from a (significantly time delayed) recontacted follow-up. We encourage future replications to further unpack demographic differences in perceptions of humanness and the accuracy of detecting AI-generated content.

Lastly, 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. These findings align with 2024 public opinion data; concern about AIGC misinformation tracks partisanship and demographics more than actual exposure to synthetic content (40), and people's judgments of algorithmic systems tend to follow trust heuristics rather than technical knowledge (41).

Collectively, these results suggest that users may utilize heuristics (42, 43) in their decisions about whether content is generated by humans or by AI, offering a broader theory to interpret otherwise disparate results. User decisions operated through multiple distinct heuristics (content, platform, and identity) rather than a single discriminative judgment, each drawing on a different source of information and each failing based on individual factors. Content-based heuristics led participants to associate informal language, specific emotional expression, and norm-violating content with human authorship, consistent with heuristic-systematic processing accounts of how people evaluate online content (44). Users' focus on platform stylistic features like hashtags, retweets, and user tags on X/Twitter crowded out the stylistic signals participants relied on elsewhere, reducing detection to near-chance levels and suggesting that platform affordances can suppress the very cues people use to identify AI content. Finally, identity-based heuristics driven by partisan motivated reasoning (34, 45) operated independently of content and platform, leading participants to evaluate ideologically-congruent AI personas as more human regardless of their stylistic features. Each heuristic comes from a different type of information, vulnerable to a distinct set of conditions. The boundary between successful and failed AI detection shifts accordingly depending on the content itself, the platform context, and the partisan identity of the human evaluator.

We note 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 platforms, we could not verify authorship with complete certainty. Our social media data collection took place 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. Furthermore, despite a large and diverse sample, limitations of online survey experiments, such as the face validity of pairwise comparisons evaluated through a survey platform, should temper strong claims about how judgments operate in fully naturalistic environments. The AI-generated posts in this study were designed to closely emulate human-authored content in format, length, and style. In-the-wild, AI-generated content is evolving rapidly and may vary substantially in sophistication and likely varies in intent. Therefore, our findings may overestimate detection difficulty for lower quality AI outputs while underestimating it for future models trained more extensively on short-form platform-specific data. Future research could expand beyond single-event contexts to examine how *perceived humanness* evolves across issue domains and 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

[8] Our original instrument was short on attitudinal and behavioral items, since most of the survey was devoted to the pairwise classification tasks. The follow-up adds distinct batteries we left out from the original instrument, mainly focused on AI usage and digital literacy. In SI Appendix S8 Table S16, we present an extensive analysis of the internal validity of the follow-up survey.

persuasion or trust. Lastly, our results rely upon an online convenience sample. Our design uses quotas to match the U.S. general population on sex, party, and race (See SI Appendix Table S19), but the sample is slightly younger and more college-educated than the general population. A large body of research shows that convenience samples can be used to estimate population quantities in behavioral experiments (46, 47), and we think this sample appropriately approximates our target population of online social media users. We encourage future replications of our design with distinct samples to assess the generalizability of our findings.

Policy Implications. Because human detection has clear limits, policy interventions should be context-specific rather than one-size-fits-all. Where detection is weakest, in content from newer models like Llama 3.3 and on X/Twitter, platforms and developers should prioritize technical safeguards such as authenticity-verification tools and techniques. Targeted media literacy for groups with lower AI discernment will likely matter more than generic AI awareness campaigns. Regulatory approaches must stay adaptive as model capabilities and platform features evolve.

Human judgment currently offers only a partial defense against AI-generated political content, and as models train further on platform-specific data, today's detection cues will likely erode. New cues may emerge, but accurate identification will increasingly call for humans to draw on both content-level and source-level signals. Future research could benchmark human detection against automated AI classifiers, providing a deeper understanding of systematic differences from human heuristics. The health 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. Whether individuals can recognize AI-generated political content as such remains central to the informed, independent judgment that democratic legitimacy requires.

Materials and Methods

Survey design. To evaluate how individuals assess the *perceived humanness* of human-authored vs. AI-generated political social media content, we conducted two within-subjects online behavioral experiments (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 July 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 and participant recruitment and distribution occurred on Connect (48). Details on participants' demographic breakdown are included in *SI Appendix Table S19*. The study protocol was approved by Georgetown University's Institutional Review Board (STUDY00008554). Informed consent was obtained online from all survey participants. The study was described and the use of their data was explained within the consent form. Participants acknowledged agreement before proceeding with the online study.

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 respondents' engagement. Only respondents passing all checks were retained. Questionnaire and attention check wordings, and details on pairwise construction are available in *SI Appendix, Section S5*.

Pairwise comparisons provide robust scaling of latent perceptions. In our case, perceptions of humanness of social media posts were 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 shown that pairwise comparisons reduce bias and measurement error compared to traditional annotation methods (49), and lead to more reliable labels by reducing individual biases in scoring (50).

Data collection. We analyze posts from X/Twitter and Truth Social, and comments from YouTube and TikTok video discussions; for consistency, we refer to all content as posts throughout. 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 S3 in the SI Appendix 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 focused 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 were not AI-generated or purely human-authored, we took steps to verify that the posts originated from human-operated accounts rather than bots. For human-authored posts on X/Twitter, we screened accounts for bot activity using the Botometer API. Only two accounts exceeded the classification threshold; both were manually reviewed and determined to reflect human-operated behavior, and all posts were retained. No equivalent bot-detection tool existed for YouTube, TikTok, or Truth Social. We also cannot fully verify that human-authored posts are free of AI-generated content, as automated detection of AI-generated text in short-form social media remains unreliable

for our corpus (25). We note that our data collection took place in September 2024, when AI-generated text on political social media remained limited in scale; approximately 1.4% of X/Twitter posts during the 2024 presidential election were classified as AI-generated (51). This suggests our corpus reflects primarily organic human content. We acknowledge that assessing the authorship of collected posts with complete certainty remains a limitation.

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 , GPT-4o, GPT-4o mini), and the Llama family of open models (Llama 3.3, Llama 3.2, and Llama 2). 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, Section S2.2*. 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 punctuation, emojis, and hashtags related to the provided debate transcript. YouTube and TikTok prompts included video transcripts that we manually collected from these platforms (more details in *SI Appendix, Section S2.1*) and LLMs were prompted to generate video comment-like posts between 5-50 words related to the given transcripts.

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 (52). 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, Table S6*.

The Perceived Humanness Indicator (PHI) Score. We converted our survey's pairwise responses into continuous humanness scores using the Bradley-Terry (BT) model (53) using `statsmodels` (54), which estimates relative item strength from pairwise comparison outcomes. For posts i and j, the model defines the probability that the survey respondent perceives 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. Parameters were estimated via maximum likelihood estimation (MLE), and scores were normalized to $[-1, 1]$ using a rank-based transformation; full implementation details are provided in SI Appendix, Section S6. We term these normalized scores the *Perceived Humanness Indicator (PHI)*: the closer a post's *PHI* score was to -1, the more AI-like it appeared to survey respondents, while the closer it was 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.

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

Marginal Effects of Pooled- and Platform-Level Assessments of PHI Score. Using our dataset of posts and their *PHI* scores, we assessed differences in 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.

Emotive Expression Feature Analysis. Emotion analysis has been widely applied to automatically categorize text via natural language processing techniques by implementing methodologies from foundational psychology, psycholinguistics, cognitive psychology, and behavioral science human subjects research (19, 55). This interdisciplinary work has led to the development of extensive lexical resources and standardized taxonomies designed to identify and categorize human emotional expression in language (56, 57). Prior research has investigated the difference in emotive language use between human and AI authors, finding that human-authored text often contains richer and more varied emotive language than AI-generated text, making emotional expression a potentially useful indicator of AI authorship (30, 58).

To investigate how emotive expression may influence *perceived humanness*, we selected five (emotion recognition, sentiment analysis, irony detection, hate speech detection, and offensive language identification) of the seven classification tasks provided by TweetEval, a unified framework for evaluating emotional language, with each task derived from gold-standard benchmark datasets (31), accessed through the TweetNLP library(32). Each post is classified with a RoBERTa-base model for each of the five tasks. We exclude the emoji prediction and stance detection tasks because they were not sufficiently generalizable to our analysis.

Emotion recognition contains three emotions (Joy, Sadness, Optimism), where a given tweet is assigned one of the three. These emotions were selected from a set of 11 established emotion categories (57), as they are the most common emotions expressed individually, as opposed to being part of a combination of emotions expressed. Sentiment analysis provides the standard labels (Positive and Negative) derived from the SemEval-17 Task 4 (Sentiment Analysis in Twitter) dataset (59). We group irony, hate speech, and offensive language detection into a broader category of *Civility*, as these tasks produce binary indicators that directly correspond to the presence or absence of those features. The irony and hate speech datasets were collected via related hashtags (e.g., #irony and #sarcasm for irony and #migrant and #women for hate speech) and then manually annotated and refined

to create ground-truth classification datasets (60). Offensive language classification is trained on the benchmark dataset Offensive Language Identification Dataset (OLID) (61).

The emotive expression analysis leverages the pairwise design, as 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 analyses of how specific emotional expressions contribute to perceptions of humanness in social media content – regardless of their authorship. The models also control for other descriptive textual features, such as the number of characters in the post, the number of words in the post, the number of sentences, the average word length, the number of exclamation points, and the proportion of characters in all caps.

Individual-Level Predictors of Correctly Identifying GenAI Posts. 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. All continuous predictors were standardized to facilitate the interpretation of effect sizes. [9]

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 mixed-effects logistic regression conditional on participants' party identification (Democrat vs. Republican), the political persona assigned to the post, and their interaction. We also employ random intercepts at the participant-level and at the order in which the participant saw a given pair in the survey. For each pair, we recorded the survey respondent's party identification and which AI persona's post was selected as more AI-like. We treat party leaners as part of their respective parties and exclude real independents from this analysis.

ACKNOWLEDGMENTS. This research was supported by the Tech Public Policy (TPP) Grant at the McCourt School of Public Policy and Project Liberty. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript. 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. We are grateful to our team's research manager, Rebecca Vanarsdall, who manages the research project and the MDI technical team for their technical support.

[9] The variance component associated with the order random intercept is small relative to the participant-level component (e.g., *SD* = 0.23 vs. 0.51 on the logit scale in the demographic model), indicating that order effects contribute modestly to results compared to between-participant heterogeneity.

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. Christopher Summerfield, Lisa P Argyle, Michiel Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian K Hadfield, et al. The impact of advanced ai systems on democracy. *Nature Human Behaviour*, 9(12):2420–2430, 2025.
3. 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.
4. W. Lance Bennett and Shanto Iyengar. A new era of minimal effects? The changing foundations of political communication. *Journal of Communication*, 58(4):707–731, 2008. .
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. 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.
9. 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.
10. 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.
11. 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/.
12. 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.
13. 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.
14. 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.
15. 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/.
16. 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.
17. 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.
18. 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.
19. 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.
20. Susan Bickford. Emotion talk and political judgment. *The Journal of Politics*, 73(4): 1025–1037, October 2011. . URL https://doi.org/10.1017/S0022381611000740.
21. 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. .
22. Andrew R. Williams, Leila Burke-Moore, Rachel S.-Y. Chan, Faith E. Enock, Federico Nanni, Tara 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. .
23. 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.
24. 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/.
25. Hillary Dawkins, Kathleen C. Fraser, and Svetlana Kiritchenko. When detection fails: The power of fine-tuned models to generate human-like social media text. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13494–13527, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. . URL https://aclanthology.org/2025.findings-acl.695/.
26. Hauke Licht, Rupak Sarkar, Patrick Y. Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle. Measuring scalar constructs in social science with LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32144–32171, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. . URL https://aclanthology.org/2025.emnlp-main.1635/.
27. 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.
28. Kenneth Benoit, Kevin Munger, and Arthur Spirling. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2): 491–508, 2019.
29. Patrick Y Wu, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. Large language models can be used to estimate the latent positions of politicians. *arXiv preprint arXiv:2303.12057*, 2023.
30. Autumn Toney, Leticia Bode, Tiago Ventura, Ethan Wilcox, and Lisa Singh. Comparing the humanness of machine-generated and human-authored text. *ACM Computing Surveys*, 2026. .
31. Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, 2020.
32. Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, U.A.E., November 2022. Association for Computational Linguistics.
33. Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets, 2020. URL https://arxiv.org/abs/2005.10200.
34. Ziva Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498, 1990. .
35. 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.".
36. Mubashir Sultan, Alan N Tump, Nina Ehmann, Philipp Lorenz-Spreen, Ralph Hertwig, Anton Gollwitzer, and Ralf HJM Kurvers. Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences*, 121(47):e2409329121, 2024.
37. 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.
38. Adam Cheng, Yiqun Lin, Gabriel Reedy, Christine Joseph, Samantha Wirkowski, Viviane Mallette, Vikhashni Nagesh, David Krieser, and Aaron Calhoun. Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content. *Advances in Simulation*, 10:66, 2025. .
39. 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.
40. 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.
41. 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.
42. S. Chaiken. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39: 752–766, 1980.
43. S. Shyam Sundar and Jinyoung Kim. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, Conference on Human Factors in Computing Systems – Proceedings, Glasgow, United Kingdom, May 2019. Association for Computing Machinery. .
44. Miriam J. Metzger and Andrew J. Flanagin. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59:210–220, 2013. . URL https://www.sciencedirect.com/science/article/pii/S0378216613001768.
45. M. Lodge and C. S. Taber. *The Rationalizing Voter*. Cambridge University Press, 2013.
46. Kevin J Mullinix, Thomas J Leeper, James N Druckman, and Jeremy Freese. The generalizability of survey experiments. *Journal of Experimental Political Science*, 2(2): 109–138, 2015.
47. Alexander Coppock, Thomas J Leeper, and Kevin J Mullinix. Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49):12441–12446, 2018.
48. Connect Research Platform. Connect: Rapid Data Collection for Behavioral Science Research, 2024. URL https://connect.georgetown.edu. Accessed April 2, 2025.
49. 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.
50. 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.
51. Zhiyi Chen, Jinyi Ye, Emilio Ferrara, and Luca Luceri. Prevalence, sharing patterns, and spreaders of multimodal ai-generated content on X during the 2024 U.S. presidential election. *arXiv preprint arXiv:2502.11248*, 2025.
52. Nakatani Shuyo. Language detection library for java. https://github.com/shuyo/language-detection, 2010. Accessed March 2025.
53. 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.
54. Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.
55. Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. *The measurement of meaning*. Number 47. University of Illinois press, 1957.
56. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).
57. Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.
58. David M Markowitz, Jeffrey T Hancock, and Jeremy N Bailenson. Linguistic markers of inherently false ai communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1):63–82, 2023. .
59. Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. . URL https://aclanthology.org/S17-2088/.
60. Cynthia Van Hee, Els Lefever, and Véronique Hoste. SemEval-2018 task 3: Irony detection in English tweets. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. . URL https://aclanthology.org/S18-1005/.
61. Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. . URL https://aclanthology.org/N19-1144/.

Supporting Information for

Human-authored or AI-generated? Humans show limited ability to tell the difference for political content on social media

Sejin Paik, Tiago Ventura, Rebecca Ansell, Leticia Bode, Autumn Toney-Wails, and Lisa Singh

Corresponding author: Sejin Paik.
E-mail: sp1822@georgetown.edu

This PDF file includes:

Supporting Information Text

S1. Overview of the Analysis Pipeline. This GitHub repository **(to be provided)** provides a modular and reproducible workflow for the collection, processing, evaluation, and analysis of social media text (human-authored and AI-generated) in the context of perceived humanness. Our analytical workflow is designed to ensure transparency and reproducibility, ensuring consistent terminology and modular organization throughout all components of the study. We provide details on each component as follows:

Workflow: Data collection → Data processing → Human evaluation → Analysis.

- **Data collection** (Sections S2-S3): Acquisition of real-world posts from X/Twitter, YouTube, Truth Social, and TikTok within fixed temporal windows surrounding the 2024 U.S. presidential debate, along with AI-generated content created via large language models (LLMs) using persona-specific prompts and debate transcripts.
- **Data processing** (Section S4): Standardized cleaning, sampling, and pair construction procedures applied consistently across both real-world and AI-generated data. Steps included normalization, de-duplication, English-language filtering, and balanced pairing for annotation.
- **Human evaluation** (Section S5): Deployment of pairwise comparison surveys to non-expert U.S.-based annotators via the Connect Cloud Research platform. Participants completed approximately 30 judgments each, with attention checks and demographic balance targets.
- **Analysis** (Sections S6–S7): Bradley–Terry scaling of pairwise results to generate continuous *Perceived Humanness Indicator (PHI)* scores, followed by descriptive, stylistic, and emotional analyses across platforms and model families.

S2. Data Collection. The dataset is composed of two distinct parts: human-authored content and AI-generated content.

S2.1 Human-Authored Content. Collected from social media platforms during defined time windows surrounding the 2024 U.S. presidential debate:

- **X (formerly Twitter)**: Posts containing the hashtags #debatenight and #debate2024 from *September 9, 2024, 11:57 PM ET* to *September 11, 2024, 5:26 AM ET*.
- **Truth Social**: Posts containing #debate2024 or #debatenight from *September 10, 2024, 7:00 PM–11:59 PM ET*.
- **YouTube**: Top-level comments (i.e., original comments on the video, not replies) from 10 debate recap videos (Table S1).
- **TikTok**: Top-level comments from 10 debate recap videos (Table S1).

Table S1. TikTok and YouTube video sources

Source Type	Source Name	Handle	Followers	Stance	Leaning	TikTok ID	YouTube ID
News	Boston Globe	@bostonglobe	88.7K	Neutral	Moderate	7413232974096878891	OQgE0ETV81s
News	ABC News	@abcnews	17.6M	Neutral	Moderate	7413218924390632735	4aw5FziSc14
Public Figure	Tom Bilyeu	@tombilyeu	4.39M	Neutral	Moderate	7432437724176583968	FxgOR1KHCpQ
Public Figure	Earn Your Leisure	@earnyourleisure	1.56M	Neutral	Moderate	7413925988951657770	tv_xWNfp0Bc
News	MSNBC	@msnbc	1.6M	Pro-Kamala	Left	7413199911044517166	ceVV5rwV4L4
News	CNN	@cnn	17.1M	Pro-Kamala	Left	7413197374048341294	kADujQsWO68
Public Figure	The Daily Show	@thedailyshow	11.7M	Pro-Kamala	Left	7413239838541106478	KtHn59wqdBc
News	Fox News	@foxnews	12M	Pro-Trump	Right	7413405886116318495	fgN3DZNHtFI
Public Figure	Megyn Kelly	@megynkellyshow	2.61M	Pro-Trump	Right	7413572606147939630	5cft5c6mE8k
Public Figure	Valuetainment	@valuetainment	6.06M	Pro-Trump	Right	7414599502465600798	Xp_xIfyDIG4

Note. Video IDs can be used to construct full URLs: TikTok videos at `tiktok.com/@[handle]/video/[ID]`, YouTube videos at `youtube.com/watch?v=[ID]`. All sources provided commentary on the 2024 U.S. presidential debate.

Bot Detection. To confirm that we were collecting content from organic accounts rather than automated sources, we ran all X/Twitter authors through the Botometer API (3) accessed via RapidAPI.*

Across 150 posts, we had 147 unique authors, 79 of whom were successfully located and evaluated by Botometer. Botometer assigns each account a score between 0 (least bot-like) and 1 (most bot-like), estimating the likelihood that an account exhibits automated behavior. As shown in Figure S1, most accounts (77/79) scored below the 0.5 threshold (red dotted line), suggesting that the majority of X/Twitter authors in our dataset are likely human-operated. We note, however, that Botometer has documented false positive rates and its reliability varies by account type and language (16); we therefore treated screening as a verification step rather than a strict filter, retaining all posts including the two accounts above the threshold. Comparable bot-detection frameworks are not available for YouTube, TikTok, or Truth Social, as these platforms do not provide the necessary account-level metadata or APIs required for automated assessment.

Automated detection of AI-Generated Content in Human-Authored Posts. A potential concern is whether the posts collected as human-authored themselves contain AI-generated content. We considered using automated AI-generated text detectors to verify our human corpus as an additional screening step. However, automated detection of AI-generated content in short-form text remains an open and largely unsolved problem in the field. Dawkins et al (15) demonstrate that while AI-generated social media posts can be detected under idealized conditions with knowledge of the generating model, detection largely fails under realistic conditions where the generating model is unknown, which is the problem that we have. Furthermore, one of the more popular commercial text detectors, GPTZero, requires a minimum of 250 characters to produce reliable classifications, making it unusable for most of our corpus, as not only are our texts short but they contain hyperlinks, user mentions and hashtags.

S2.2 AI-Generated Content. Produced by LLMs using debate transcripts (for X/Twitter and Truth Social) or video transcripts (for YouTube and TikTok) with five political persona-specific prompts adapted to each platform's stylistic norms. All personas were equally represented across platforms in the generation process. The exact prompts used for each platform are included in the repository.

Large Language Models (LLMs) To generate AI posts, we sampled two dominant LLM families that jointly span (a) closed, frontier-grade systems widely used in production (OpenAI's GPT-4 line) and (b) state-of-the-art, open-weight models used in academia and industry research (Meta's Llama line). Within each OpenAI GPT and Llama families, we included both a flagship and a lighter-weight tier to probe the size and capability effects (GPT-4 and GPT 4o vs. GPT-4o-mini; Llama 3.3 vs. Llama 3.2) and also incorporated Llama 2 to capture a prior-generation, open model. This design balances ecological relevance (i.e., publicly popular models that people are using), representativeness across licensing approaches (i.e., closed vs. open-weight), and variation in parameter scale and release dates. Model documentation and specifications can be found in Table S2. The information was obtained from official sources: GPT-4 (4), GPT-4o (5), GPT-4o-mini (6), Llama 2 (7), Llama 3.2 (8), and Llama 3.3 (9).

*https://rapidapi.com/OSoMe/api/botometer-pro

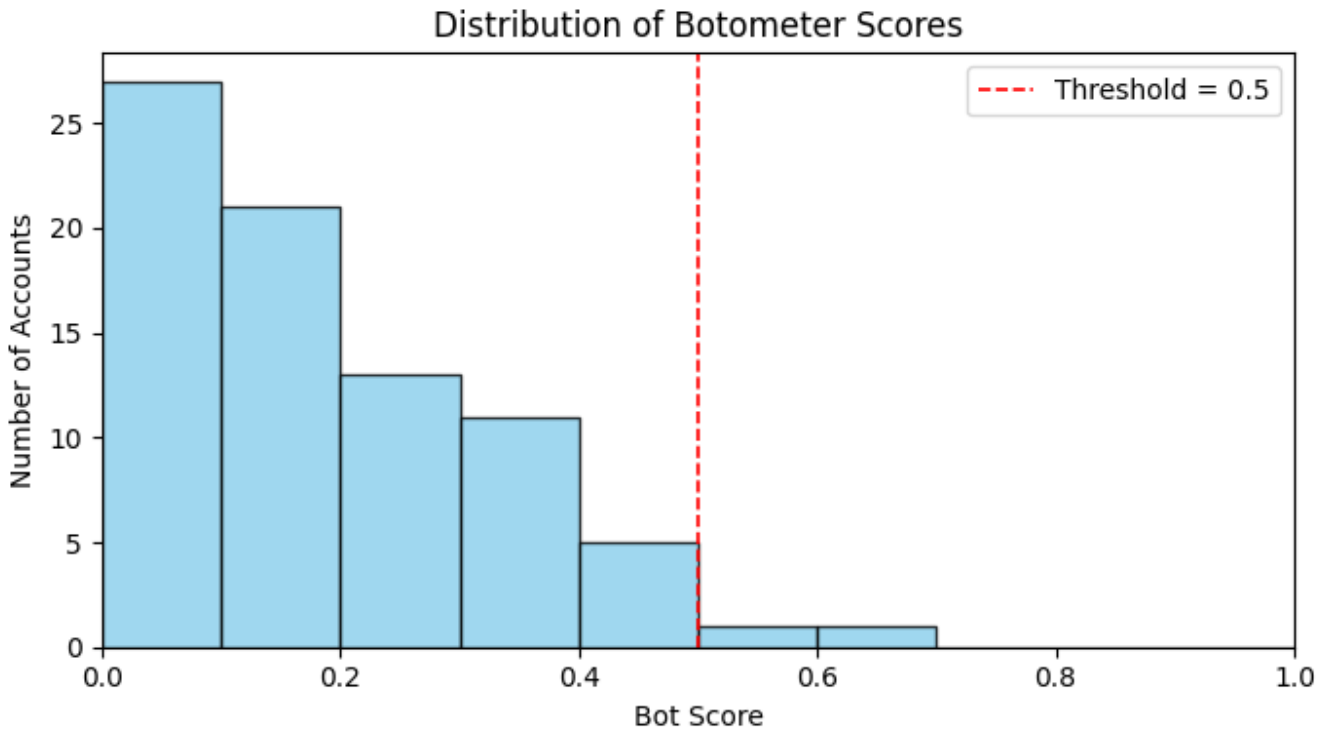


Fig. S1. Distribution of Botometer scores for X authors. Scores range from 0 (least bot-like) to 1 (most bot-like), with a dotted red threshold line at 0.5 indicating the midpoint between likely human and likely bot accounts.

Table S2. LLM details

Model	Release year	Knowledge cutoff	Parameters
GPT-4	March 14, 2023	Dec. 1, 2023	Undisclosed
GPT-4o	May 13, 2024	Oct. 1, 2023	Undisclosed
GPT-4o-mini	July 7, 2024	Oct. 1, 2023	Undisclosed
Llama 2	July 18, 2023	July 2023	70B
Llama 3.2	Sept 25, 2024	Dec 2023	3B
Llama 3.3	Dec 6, 2024	Dec 2023	70B

Political Persona Definitions (used equally across all AI-generation tasks): The persona descriptions were developed in two steps. First, we defined five ideological profiles corresponding to the AllSides (22) spectrum (far-left, left, center, right, far-right), with each persona assigned a distinct political stance, candidate preference, and disposition toward the 2024 presidential candidates. Each persona was assigned a distinct political stance, candidate preference, and disposition toward the 2024 presidential candidates. Second, the substantive content of each persona was grounded in three policy domains known to vary systematically across the ideological spectrum in the American electorate: views on government size and service provision, foreign policy orientation, and cultural values such as immigration, race, abortion, gun ownership, and criminal justice. The specific policy positions assigned to each persona were drawn from established survey-based evidence on how Americans at each ideological position tend to view these issues (18, 19, 21). To verify that generated posts reflected their intended personas, two members of the research team independently reviewed a random sample of 10 posts per persona and confirmed that the content was ideologically consistent with its assigned political profile. We acknowledge that persona-based prompting simplifies real ideological diversity, and that LLMs may not always faithfully reproduce the political perspectives they are prompted to adopt (20). However, the use of an established, validated ideological typology and domain-specific policy grounding was used to reduce the risk of producing ideologically incoherent or unrecognizable personas.

- **Progressive Commentator** (Left): You have a deep dislike for Trump, seeing him as a symbol of everything that stands in the way of social progress. You strongly oppose his leadership style and policies, and you're enthusiastic about Harris. While she may not push far enough on certain issues, you believe she's the best hope for moving the country toward a more just and inclusive future. Views of government: You favor larger government that provides more services. Foreign policy: As a left-leaning, Kamala Harris supporter, you believe Harris is more likely than Trump to say that the United States should take into account the interests of its allies, and that is at least very important for the U.S. to

have an active role in world affairs. Cultural values: 1. Immigration and America's openness to people from all over the world is essential to the U.S. 2. Legacy of slavery greatly affects the position of black people in american society today 3. Pro-LGBTQ+ and pro-abortion

- **Liberal Commentator** (Lean left): You are a liberal commentator on the social media platform, X, with a left leaning political stance in the U.S. You are critical of Trump, viewing his tenure as chaotic and damaging to national unity. You like Harris, though you're not entirely aligned with her more progressive stances. You support her primarily because she represents stability, competence, and a path forward, while Trump represents a direction you firmly oppose.Views of government: You favor larger government that provides more services. Foreign policy: As a left-leaning, Kamala Harris supporter, you believe Harris is more likely than Trump to say that the United States should take into account the interests of its allies, and that is at least very important for the U.S. to have an active role in world affairs. Cultural values: 1. Immigration and America's openness to people from all over the world is essential to the U.S. 2. Legacy of slavery greatly affects the position of black people in american society today 3. Pro-LGBTQ+ and pro-abortion

- **Moderate Commentator** (Center): You are a moderate commentator on the social media platform, X, with a center leaning political stance in the U.S. You are conflicted about both Trump and Harris. You dislike Trump's confrontational, divisive style, but you can acknowledge some of his past economic accomplishments. Harris feels like a safer, more unifying option, but you have reservations about her leaning too far left. Ultimately, you have not made up your mind yet.

- **Conservative Commentator** (Lean right): You are a conservative commentator on the social media platform, X, with a right leaning political stance in the U.S. You dislike Harris, believing she represents policies that will steer the country in the wrong direction. You are more inclined to support Trump, even if you don't always agree with his behavior. You appreciate his defense of conservative values, particularly regarding the economy and national security, and see him as the better choice to protect the country's interests.Views of government: You favor smaller government, providing fewer services. Foreign policy: As a right-leaning Donald Trump supporter, you are more likely to support policies aimed at maintaining America's role as the world's lone military superpower. Cultural values: 1. Gun ownership does more to increase safety by allowing law-abiding citizens to protect themselves. 2. Criminal justice system in the US is generally not tough enough on criminals 3. Anti-abortion

- **Libertarian Populist Commentator** (Right): You are a libertarian/populist commentator on the social media platform, X, with a right political stance in the U.S. You strongly dislike Harris, viewing her as a symbol of government overreach and political correctness. You have a deep distrust of the establishment she represents. You fully support Trump, not necessarily because you agree with everything he says, but because you see him as the only one willing to challenge the elites and disrupt the system in the ways you think are necessary. Views of government: You favor smaller government, providing fewer services. Foreign policy: As a right-leaning Donald Trump supporter, you are more likely to support policies aimed at maintaining America's role as the world's lone military superpower. Cultural values: 1. Gun ownership does more to increase safety by allowing law-abiding citizens to protect themselves. 2. Criminal justice system in the US is generally not tough enough on criminals 3. Anti-abortion

Generation details

- **X/Twitter**: Generated in batches of approximately 20 per persona prompt using the debate transcript.
- **Truth Social**: Generated in batches of approximately 20 per persona prompt using the debate transcript.
- **YouTube**: Generated per video per persona using the video transcript as context.
- **TikTok**: Generated per video per persona using the transcript as context.

S2.3 Total counts of posts. Following the procedure described in the prior sections, Table S3 presents the final count of posts collected for each platform and LLM combination used in our surveys.

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

Table S3. Final counts of posts by platform and model

S3. Descriptive Information of Social Media Posts. In this section, we provide several descriptive statisics of our stimuli, both human and AI-generated. Tables S4 reports the character-length distribution of posts by author group for each platform (Truth Social, X/Twitter, YouTube, and TikTok respectively). Table S5 reports the mean number of stylistic markers per post (hashtags, retweets (RTs), mentions, and URLs) for human and AI-generated posts across all four platforms, along with the percentage of posts containing each marker.

Figure S2 presents results of OLS models regressing the *PHI* scores on several of these textual descriptive statistics; *substantive text length* represents the total number of characters on a post minus characters dedicated to platform, *platform's marker count* indicates the counts of platform-style markers in a post (hashtags, tags, URLs, Retweets), *setence counts* indicates the number of sentences, *average word length* the mean number of characters in a word for each posts, *exclamation count* indicates counts of exclamation points, and *% caps* the share of characters of a post that are upper case. We estimate models pooling all posts together, and separating the effects conditional on the post ground truth, either human or AI-generated. Results indicate that longer posts reduce on average the *PHI* score. However, the effect is mostly driven by longer substantive length AI writing, which parses out characters spent on platform markers. This finding provides explanatory explanation for the low AI detection rates of X/Twitter, a platform in which these markers are more popular, and native to the platform. Therefore, less substantive AI writing goes in the posts shown to participants.

Table S4. Post length in characters by platform: Human vs. AI Pooled (all AI models combined).

Platform	Author	N	Min	Max	Mean	Median	SD
Truth Social	Human	161	41	289	116.7	99.0	60.1
	AI Pooled	432	44	356	118.9	117.0	36.6
X/Twitter	Human	203	38	313	136.3	123.0	60.9
	AI Pooled	390	42	205	103.8	104.0	30.7
YouTube	Human	153	40	2134	160.0	100.0	201.0
	AI Pooled	447	41	314	89.7	83.0	41.7
TikTok	Human	179	41	259	102.5	83.0	51.0
	AI Pooled	421	41	241	83.7	76.0	32.4
All Platforms	Human	696	38	2134	128.3	106.0	108.9
	AI Pooled	1690	41	356	98.9	96.0	38.3

Stylistic Markers across Platforms. Table S5 reports the mean number of stylistic markers per post (hashtags, retweets (RTs), mentions, and URLs) for human and AI-generated posts across all four platforms, along with the percentage of posts containing each marker.

Table S5. Stylistic markers by platform: human vs. AI Average (all AI models pooled). Values are mean per post (percentage of posts containing the marker).

Platform	Author	N	Hashtags	RTs	Mentions	URLs
	Human	161	2.17 (98%)	0.09 (9%)	0.34 (24%)	0.34 (34%)
	AI Average	432	1.09 (88%)	0.14 (14%)	0.22 (20%)	0.16 (16%)
	Human	203	2.07 (97%)	0.22 (22%)	0.43 (36%)	0.61 (56%)
	AI Average	390	1.26 (98%)	0.14 (14%)	0.26 (24%)	0.01 (1%)
	Human	153	0.05 (1%)	0.00 (0%)	0.02 (2%)	0.00 (0%)
	AI Average	447	0.40 (22%)	0.00 (0%)	0.01 (1%)	0.00 (0%)
	Human	179	0.31 (14%)	0.00 (0%)	0.16 (16%)	0.00 (0%)
	AI Average	421	0.50 (37%)	0.00 (0%)	0.00 (0%)	0.00 (0%)

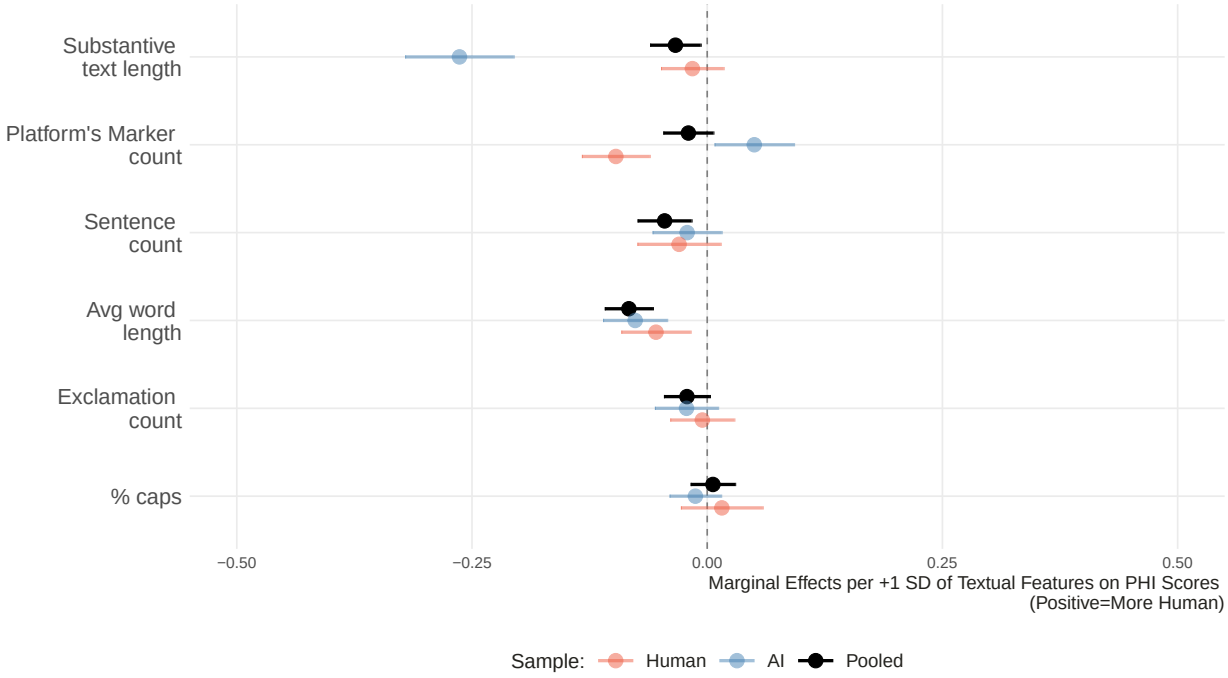


Fig. S2. Standardized Marginal Effects of Textual Features on Phi Scores (Pooled and Conditional on Post Type)

S4. Data Processing. Table S6 summarizes the preprocessing steps applied to human-authored and AI-generated posts for each platform.

Table S6. Platform-specific data processing steps applied to real-world and AI-generated datasets.

Platform	Human-Authored Data Processing	AI-Generated Data Processing
X/Twitter	• Replace URLs with broken `t.co`-style tokens • Censor user mentions to `@USER` • English-only filtering • Require `#debate2024` or `#debatenight` • Remove `"RT @USER"` patterns • Deduplicate across full dataset	• Replace URLs with broken `t.co` tokens • Censor user mentions to `@USER` • Include `RT @USER` patterns in subset of posts (prompt instruction) • Enforce first-5-words uniqueness across generated posts
YouTube	• Replace URLs with `HTTPURL` token • Censor user mentions to `@USER` • Remove comments with fewer than 5 words • English-only filtering • Sample exactly 50 unique top-level comments per video • Total of 500 comments across 10 videos	• Enforce first-5-words uniqueness across generated comments • No external links expected in final subset
Truth Social	• Strip HTML from post content • Replace URLs with broken `tinyurl` tokens • Censor user mentions to `@USER` • Filter to debate evening window (19:00–23:59 ET, 2024-09-10) • Require `#debate2024` or `#debatenight`	• Replace URLs with broken `tinyurl` tokens • Censor user mentions to `@USER` • Include `RT @USER` patterns in subset of posts (prompt instruction) • Remove placeholder emoji strings (e.g., `:happy`)
TikTok	• Replace URLs with broken `tinyurl` tokens • Censor user mentions to `@USER` • English-only filtering	• Remove placeholder emoji strings (e.g., `:happy`, `:cry`)

Post Pair Construction for Survey Pairs were constructed by randomly sampling from the human-authored posts collected from each of the four platforms as well as from the AI-generated posts produced from each of the six different LLMs. Each selected item was paired multiple times to ensure sufficient comparisons for robust Bradley-Terry ranking estimations. For our first survey wave on X/Twitter and YouTube, we had 10 pairings per post, and for our second wave on Truth Social and TikTok, each post was paired with 20 other posts from the same platform. This structure yielded approximately 12,000 unique pairs in the second wave, rather than the initial target of 10 comparisons per post. The design did not include duplicate pair labels; instead, a larger number of unique pairings increased coverage and diversity of comparisons. Random pairs would include one of the following conditions: a) human-authored:AI-generated, b) human-authored:human-authored , c) AI-generated:AI-generated.

S5. Human Evaluation and Quality Control. Non-expert, U.S.-based respondents (recruited via the Connect Cloud Research online participant platform) completed pairwise judgments with embedded attention checks. Each participant evaluated approximately 30 pairs and answered additional survey questions. Demographic characteristics of participants are reported in Table S19. Survey response descriptive statistics for political interest, news trust, news cynicism, and conspiracy belief variables are provided in Tables S20, S21, and S22. Survey 1 refers to our X/Twitter and YouTube Survey (n=601) and Survey 2 refers to our Truth Social and TikTok Survey (n=606).

Attention Checks. Three distinct attention checks were embedded in each survey to ensure data quality and participant comprehension. Participants who failed any of the three checks were excluded from the final dataset. Only participants who passed all attention checks were retained for analysis.

1. **Instructional Check**
   Question: "Paying attention and reading the instructions carefully is critical. President Obama is the youngest politician listed in the options below. If you are paying attention, please choose the youngest politician from the list."
   Correct Answer: "President Obama"

2. **Comprehension Check**
   Question: "Please consider the text below. An AI model has generated this social media post. Please select the option below 'Strongly AI'.
   Now, indicate how likely you think this social media post was human-written or AI-generated."
   Correct Answer: "Strongly AI"

3. **Pairwise Comprehension Check**
   Question: "Please, read the social media posts below carefully. An AI model has generated this social media post. Please select the option POST A below.
   'Which of the posts (Post A or Post B) is more likely to have been generated by an AI chatbot?"
   Correct Answer: "Post A"

In Table S7, we present the results of the three attention checks for each survey.

Table S7. Attention check performance across survey platforms. Values indicate the number and percentage of participants who passed or failed each check.

Check	Survey 1 (n = 601)	Survey 2 (n = 606)
Check 1	592 (97.9%)	589 (96.9%)
Check 2	579 (95.7%)	591 (97.2%)
Check 3	584 (96.5%)	582 (95.7%)
Pass all	559 (92.4%)	563 (92.6%)
Fail all	1 (0.2%)	4 (0.7%)

S6. Bradley–Terry Scaling of Perceived Humanness (Implementation Details). Pairwise comparison data from the annotation task were converted into continuous humanness scores following the Bradley–Terry framework described in the main text. Each post's latent strength parameter (γ_i) was estimated via maximum likelihood estimation (MLE), implemented in `Python` (v3.12) with `statsmodels` (v0.14.5) (`statsmodels.api.GLM`, Binomial family) (14). The model was estimated separately for each platform (TikTok, Truth Social, X/Twitter, YouTube).

Because the Bradley–Terry model is only identified up to an additive constant, we fixed one item's strength to zero as a reference point; all other parameters are interpreted relative to it. Raw strength parameters are therefore not directly comparable across platforms.

To place scores on a common scale, we applied a rank-based transformation: each item's raw strength was ranked among all N items estimated within that platform, and ranks were rescaled to $[-1, 1]$ via:

$$\mathrm{PHI}_i \;=\; -1 + \frac{2\,r_i}{N-1} \qquad [1]$$

where $r_i \in \{0, \ldots, N-1\}$ is the item's rank (0 = lowest, $N-1$ = highest). Scores were then inverted so that higher values indicate greater perceived humanness.

The resulting ***Perceived Humanness Indicator (PHI)* score** provides a continuous, bounded measure of how human-like each post was perceived to be. All estimation and processing scripts are publicly available in the project's GitHub repository.

Example Posts per persona across platforms. Tables S8–S11 present representative posts sampled from each platform across the full range of political personas and model types, along with their corresponding *PHI* scores.

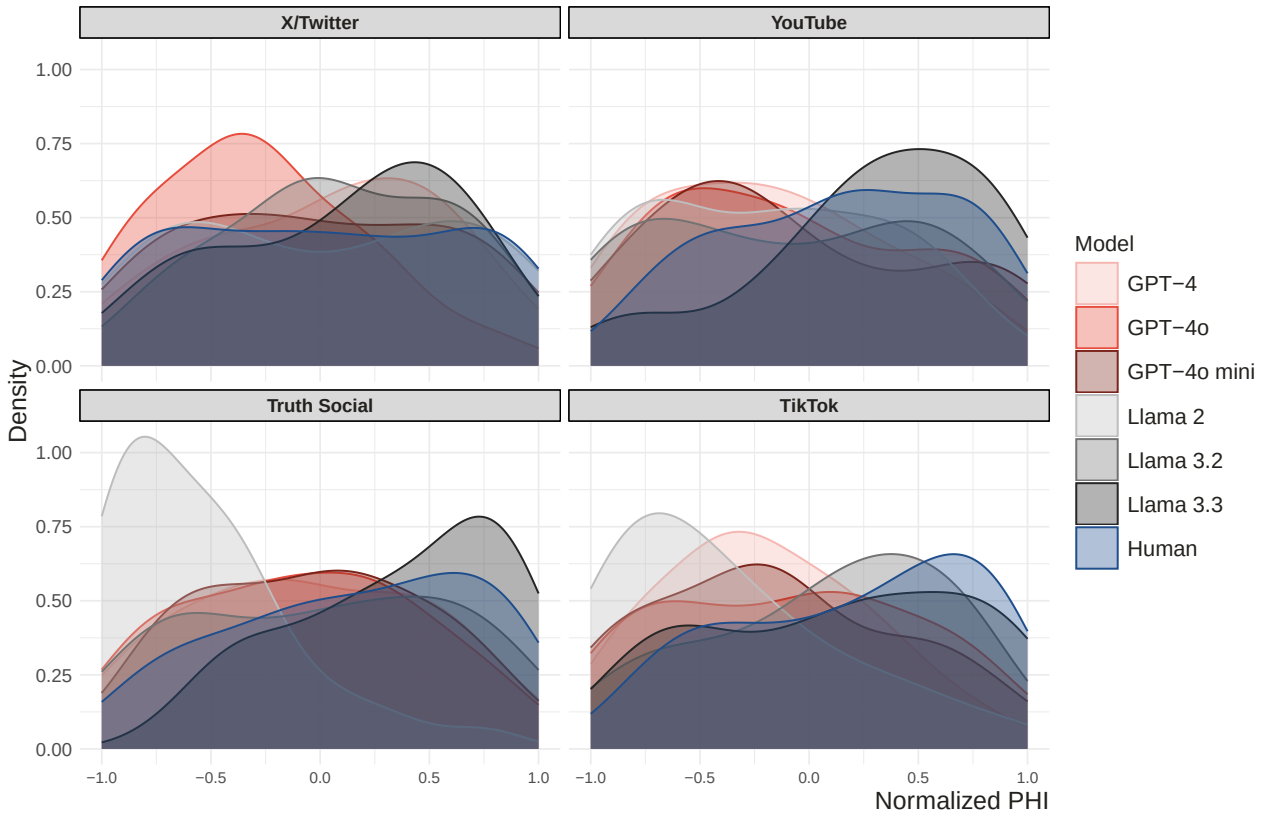


Fig. S3. Kernel density estimates of normalized PHI scores across four social media platforms. Each curve is independently normalized (area = 1) and comparable across panels.

S7. Additional Analysis.

Cross-Platform Distributions of Perceived Humanness. Figure S3 presents kernel density estimates (KDEs) of normalized *PHI* scores for posts from four social media platforms: **X/Twitter**, **YouTube**, **Truth Social**, and **TikTok**. Each subplot shows the empirical distribution of perceived humanness for one platform, with one Gaussian kernel density curve per model estimated separately using `ggplot2` library in R. All KDEs were normalized such that each curve integrates to unity, and axis limits were fixed to $[-1, 1]$ on the PHI axis and $[0, 1.2]$ on the density axis for cross-panel comparability. The distributions reveal that human-authored posts are consistently skewed toward higher *PHI* values, whereas AI-generated outputs vary in both central tendency and spread across model families. Table S23 summarize the central tendency and dispersion of normalized *PHI* scores for each model family across the four platforms.

Qualitative Assessment of Perceptions of Humanness. In addition to examining the pairwise data, we further analyzed participant responses to the question of which cues they used to decide what text was more AI-like. Open-ended responses shed light on the specific cues participants relied on when judging whether a post was human- or AI-authored. The most frequently cited primary cues were Grammar (n = 337) and Emojis/hashtags/symbols (n = 263). When disaggregated by classification choice (whether the participant labeled a post as AI, human, or expressed uncertainty), Grammar was mentioned 105 times in AI classifications but only 38 times in human classifications, and Emojis/hashtags/symbols was mentioned 100 times in AI classifications but only 14 times in human classifications. This skew suggests that linguistic polish and conspicuous formatting were often taken as signs of synthetic authorship, which helps explain why GPT-family outputs, known for consistent syntax, scored lower on perceived humanness. Emotional language was more balanced in its distribution but still leaned toward human classifications (17 mentions for human vs. 42 for AI), reinforcing the emotive features analysis finding that affective tone increases perceived humanness. However, its appearance in AI classifications shows that when AI content successfully incorporates emotional expression, participants could still misattribute it as human. These patterns substantiate the quantitative finding from above that affective expressiveness, especially emotional tone and informality, is a strong predictor of perceived humanness. They also revealed a surprising asymmetry in which features like grammar consistency and conspicuous formatting, which might be taken as neutral quality indicators, were instead widely interpreted as markers of AI authorship.

Fig. S4. Qualitative assessment of top linguistic and stylistic indicators in AI, human or unclear authorship judgments

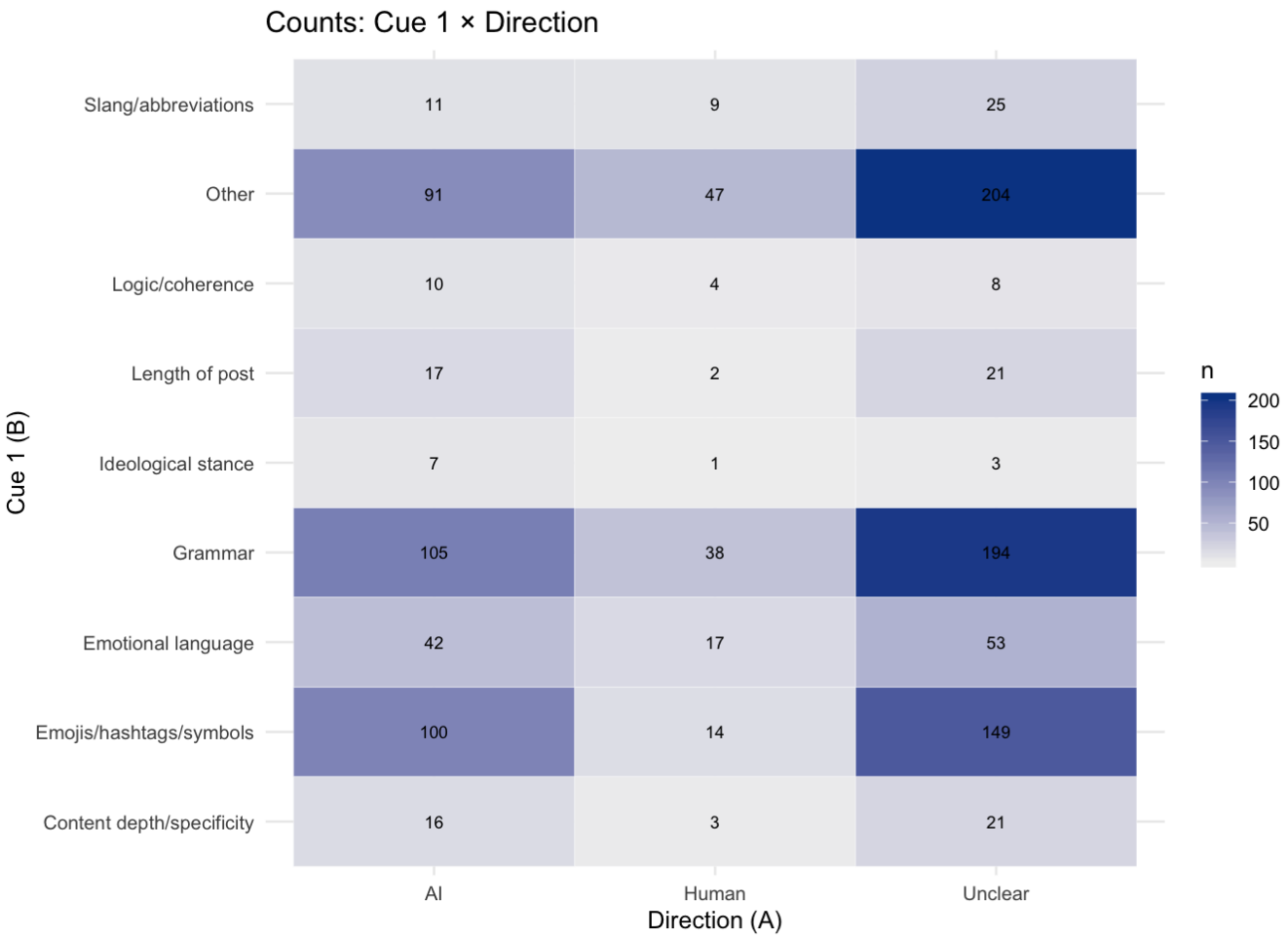


Table S8. Example posts from the X/Twitter dataset across persona and model conditions, with corresponding PHI scores.

Persona	Model	Post	PHI Score
Progressive	GPT-4o	Harris's plan to support small businesses is a game-changer! #Debate2024	-0.983
Progressive	Llama 3.2 3B	RT @USER - kamala harris just said she'd work with republicans to find common ground... meanwhile, trump is still trying to divide us #Debate2024 #UniteOrDie	-0.034
Progressive	Llama 2 70B	Why is it that @USER can dish it out but can't take it? #GrowUp #DebateNight	0.983
Liberal	GPT-4o	RT @USER: Harris for president! A real leader who cares about ALL Americans. Not just the wealthy few. #Debate2024	-0.919
Liberal	GPT-4	Supporting small businesses, check. Healthcare for all, check. Addressing economic disparity, check. This, folks, is the Harris plan! #Debate2024	0.014
Liberal	Llama 2 70B	Did Donald Trump really say we should "bomb the shit out of 'em"? #WarMonger # DebateNight	0.966
Moderate	Llama 3.3	abortion access is on the ballot this november #Debate2024	-0.976
Moderate	GPT-4	Harris speaks of united values. Hope she can back it up with policies that bring us together. #Debate2024	0.007
Moderate	Llama 2 70B	Trump is such a child, can't even answer questions properly #KamalaHarrisForPresident #DebateNight	0.889
Conservative	GPT-4	Trump's directness is REFRESHING after years of Harris's HALF-TRUTHS! #Debate2024	-0.997
Conservative	GPT-4o mini	Kamala doesn’t represent the vast majority of Americans who stand for LIFE! Don’t be fooled! #Debate2024	-0.118
Conservative	Llama 3.2 3B	abortion debate is SO heated... how can u expect to lead the country when u cant even lead on reproductive rights? #ProChoice #ReproductiveRights	0.956
Libertarian Populist	GPT-4o	RT @USER: Trump is the only one who can protect our economy. Harris will just destroy it. #Debatenight	-0.936
Libertarian Populist	Llama 3.2 3B	what's the deal with TRUMP'S TAX RETURNS? Like we haven't heard that one before #Debate2024 #TaxReturns	0.057
Libertarian Populist	GPT-4	There she goes again, Kamala blasting America. Sad really... #Debate2024	0.99
Real-world	Human	Catch up on the important takeaways from the Presidential debate. #Debate2024 : https://t.co/iqtmjczr https://t.co/zurypbnm	-0.986
Real-world	Human	Why would Kamala bring up Trump inviting the Taliban when HE ACTUALLY WAS NEGOTIATING PEACE? Which actually worked… and then Kamala violated the terms of the agreements and violence returned… thank you Kamala for giving points to Trump. #Debate2024	0.098
Real-world	Human	RT @USER: When I send you this, I know you f*cking lying #Debate2024 https://t.co/iqtmjczr	1

Table S9. Example posts from the Truth Social dataset across persona and model conditions, with corresponding PHI scores.

Persona	Model	Post	PHI Score
Progressive	GPT-4o	Kamala Harris is the leader we need to tackle America's health care crisis. Trump's done nothing but make it worse. #HealthCareForAll #Debate2024	-0.943
Progressive	GPT-4	Harris' stance on abortion rights is why I'm voting for her. Women's rights are human rights. #Debatenight	-0.145
Progressive	Llama 3.3	just watched @USER rip trump a new one on character and fitness #Debatenight	0.99
Liberal	Llama 2 70B	The debate was a great opportunity to see the candidates' visions for America's future. Who do you think has the best plan? #DebateNight	-0.983
Liberal	GPT-4o	trump's foreign policy was a mess. Kamala will bring us back to the world stage with dignity. #GlobalLeadership #Debate2024	-0.186
Liberal	Llama 3.3	what a total disaster for trump on economy questions... he has no clue #Debatenight	0.912
Moderate		The candidates discuss the role of government in society, with Harris emphasizing the need for a more active role in addressing social issues & Trump advocating for limited government intervention. Which approach do you support? #Debate2024	-0.966
Moderate	Llama 3.3	economy, schmeconomy... trump only cares about his own POCKET https://tinyurl.com/qh8t1u2n #Debate2024	-0.166
Moderate	Llama 3.2 3B	just watched @USER say she's been a 'good girl' for 30 yrs lol what does that even mean?? #Debatenight	0.983
Conservative	Llama 3.2 3B	Can we talk about health care for a sec? Kamala's plan is SO much better than Trump's #HealthcareForAll #Debate2024	-0.993
Conservative	GPT-4o mini	Kamala's views on criminal justice are too lenient. We need a system that prioritizes victims! Trump understands this! #Debate2024	0.034
Conservative	Llama 3.3	harris just owned trump on healthcare... period #Debatenight	0.986
Libertarian Populist	Llama 2 70B	RT @USER: The #DebateNight was intense! Harris and Trump went head-to-head on so many issues. What were your takeaways? #Debate2024	-0.953
Libertarian Populist	GPT-4	RT @USER: Trump's the only one on that stage who understands what it means to be a business owner. That's why he gets my vote! #Debate2024	0.041
Libertarian Populist	GPT-4o mini	RT @USER: Kamala Harris just proved she's ALL TALK and NO ACTION. Where's the plan for REAL Americans, huh? #Debatenight #WeakLeadership	1
Real-world	Human	WOW — Trump brings up debunked stories about migrants eating pets in Ohio, which prompts David Muir to fact-check him. opinions #Debate2024 #debate #Debatenight #DebateNight2024 #TrumpHarrisDebate #Trump #HarrisTrumpDebate2024 _____ https://tinyurl.com/tw2y5j9k	-0.939
Real-world	Human	Trump cryin about how dems "weaponized the justice department" which is false first of all. He routinely says how he will use the justice department to go after his political enemies. I feel like some of yall who don't know this yet fr live in the upside down. #DebateNight	0.341
Real-world	Human	Can't believe Trump thinks he's still in office #AlternativeFacts #Debatenight	0.98

Table S10. Example posts from the YouTube dataset across persona and model conditions, with corresponding PHI scores.

Persona	Model	Video	Post	PHI Score
Progressive	Llama 3.2 3B	V1	Economy update: Trump's plan to cut taxes is a RECIPE FOR ECONOMIC DISASTER!	-1
Progressive	GPT-4o	V7	Kamala Harris just wiped the floor with Trump. Her calm, collected demeanor vs. his chaotic ramblings. #Harris2024	0.058
Progressive	GPT-4o mini	V9	did trump really say immigrants are eating pets?? like wtf?! this guy is delusional!	0.97
Liberal	GPT-4o mini	V1	Kamala is making it clear that we need to protect Roe v. Wade. It's a fundamental right!	-0.973
Liberal	GPT-4	V17	Kamala's side-eye might've been off-putting to some, but I'm here for that level of scrutiny. #KamalaSideEye	0.035
Liberal	GPT-4o	V9	Trump's talking about dogs being eaten by immigrants... what the actual fck?!	0.977
Moderate	GPT-4o mini	V13	Kamala really needs to define her stance on healthcare! It's a crucial issue.	-0.967
Moderate	Llama 3.3	V6	i'm so done with these debates... can we just vote already??	-0.132
Moderate	GPT-4	V2	He admires dictators" - Harris on Trump. If true, that's pretty fing scary.	0.997
Conservative	Llama 2 70B	V13	Harris just talked about her plan for education reform. She really cares about our future generations #EducationMatters #InvestInOurFuture	-0.95
Conservative	GPT-4o	V9	When Harris talks about healthcare, all I hear is more government control. Nope. Not happening.	-0.065
Conservative	Llama 3.3	V7	i'm literally screaming at the tv right now...	1
Libertarian Populist	Llama 2 70B	V7	I can't believe how Kamala Harris just shut down Trump's foreign policy nonsense. She knows her stuff. #ForeignPolicy #KamalaHarrisSaysNo	-0.997
Libertarian Populist	GPT-4	V1	Crime through the roof? Harris is ignoring the reality. Trump's got the stats.	-0.025
Libertarian Populist	Llama 3.3	V17	character and fitness? trump has neither	0.99
Real-world	Human	V1	So… the question is remaining.. Kamala Harris or Donald Trump??	-0.923
Real-world	Human	V9	Thank you, Jon. I really needed some sanity to stabilize me during this.. I don't even know what to call it anymore.	0.185
Real-world	Human	V4	What a bumbling dumb @$$ this guy is..a total mccluster fk..	0.987

Table S11. Example posts from the TikTok dataset across persona and model conditions, with corresponding PHI scores.

Persona	Model	Video	Post	PHI Score
Progressive	GPT-4o	V7	Immigrants are essential to America! Trump's xenophobia ain't it! #OpenBorders	-0.997
Progressive	GPT-4	V1	A 13-year-old survivor forced to carry a pregnancy to term? That's the real horror story here.	0.005
Progressive	Llama 3.3	V5	can't believe he's trying to control our bodies	1
Liberal	Llama 3.2 3B	V4	Kamala Harris SLAYED that immigration debate! Trump's gotta go #KamalaHarris2024	-0.987
Liberal	GPT-4o	V5	Trump's "project 2025" sounds like a dystopian nightmare.	-0.219
Liberal	GPT-4o mini	V1	Worst president"??? LMAO, as if we forgot the chaos he left behind. Harris is the stability we need! #VoteBlue	0.95
Moderate	Llama 2 70B	V3	I dislike Trump's confrontational, divisive style, but I can acknowledge some of his past economic accomplishments. Can Harris offer a better unity-in-action plan? #Harris2024 #UnityInAction #EconomicAccomplishments	-1
Moderate	Llama 3.2 3B	V6	Can we pls just talk about something else? #DebateProbs	-0.005
Moderate	GPT-4o	V6	Trump's 'I'm an open book'... yeah, a book full of surprises!	0.953
Conservative	GPT-4o	V7	Trump is right, our country needs stronger borders! #SecureTheBorder	-0.96
Conservative	Llama 3.3	V5	ROE V WADE IS NOT THE ISSUE focus on the real problems, kamala!	-0.122
Conservative	Llama 3.2 3B	V9	Kamala Harris is a total joke! Can't believe those mods didn't ask her about her record on crime	0.957
Libertarian Populist	Llama 3.2 3B	V9	This debate was like watching a 3-on-1 fight poor Trump had no chance	-0.957
Libertarian Populist	GPT-4	V6	Harris: 'I'll be a president for all Americans.' Sure, as long as they agree with you.	-0.222
Libertarian Populist	Llama 3.3	V3	whys kamala talking like shes an economist lol	0.967
Real-world	Human		@conservative.commentator42: "Harris wants to open our borders? Trump's got the right idea, build the wall and keep us safe #BuildTheWall #ImmigrationReform	-0.913
Real-world	Human	V5	This is America our Freedom to chose what we do for with our own bodies!	0.292
Real-world	Human	V7	trump got his ass kicked on stage LIVE. and fact checked LIVE. It was glorious.	0.987

Regression Results for PHI Scores Conditional on Post Source. Figure 1 in the main manuscript presents the marginal means of *PHI* scores conditional on post source human and six distinct AI Models) in the left plot, and broken down across the four distinct platforms included in the study in the right plot. In this section, we present OLS regression models for the same results. The regression models allow us to more directly make pairwise inferential comparisons between the human and AI-generated content. Table S12 presents five distinct models: a pooled model across all platforms and four platform specific models. For all five models, the human posts are the baseline, for each all other AI-generated posts are compared to. In addition, Table S13 presents pooled regression results comparing human posts with all other AI-generated posts.

Table S12. Pairwise Differences in PHI Scores Against Human Posts

	Pooled	X/Twitter	YouTube	TikTok	Truth Social
GPT-4	-0.220***	0.012	-0.310***	-0.389***	-0.199**
	(0.039)	(0.081)	(0.079)	(0.076)	(0.074)
GPT-4o	-0.245***	-0.289***	-0.214**	-0.279***	-0.230**
	(0.039)	(0.081)	(0.079)	(0.076)	(0.075)
GPT-4o Mini	-0.186***	-0.029	-0.210**	-0.341***	-0.176*
	(0.039)	(0.081)	(0.079)	(0.076)	(0.074)
Llama 2	-0.395***	0.013	-0.313***	-0.572***	-0.693***
	(0.041)	(0.086)	(0.080)	(0.090)	(0.075)
Llama 3.2	-0.094*	0.071	-0.219**	-0.116	-0.133
	(0.039)	(0.078)	(0.079)	(0.076)	(0.077)
Llama 3.3	0.105**	0.082	0.155*	-0.064	0.207**
	(0.040)	(0.084)	(0.079)	(0.077)	(0.075)
Intercept (Human baseline)	0.120***	0.016	0.138**	0.193***	0.151***
	(0.021)	(0.040)	(0.045)	(0.041)	(0.042)
Num.Obs.	2386	593	600	600	593
R2	0.065	0.032	0.076	0.100	0.176
R2 Adj.	0.063	0.022	0.066	0.091	0.168

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: **Each column reports OLS estimates from a separate regression of PHI scores on post source, with human posts as the reference category. Coefficients represent the average difference in perceived humanness between posts from a given model and human posts. Column 1 pools all platforms; columns 2-5 estimate the same model separately for each platform. Standard errors in parentheses.**

Table S13. Regression Models: PHI Scores Differences Between Pooled AI-Generated Posts vs Human Posts

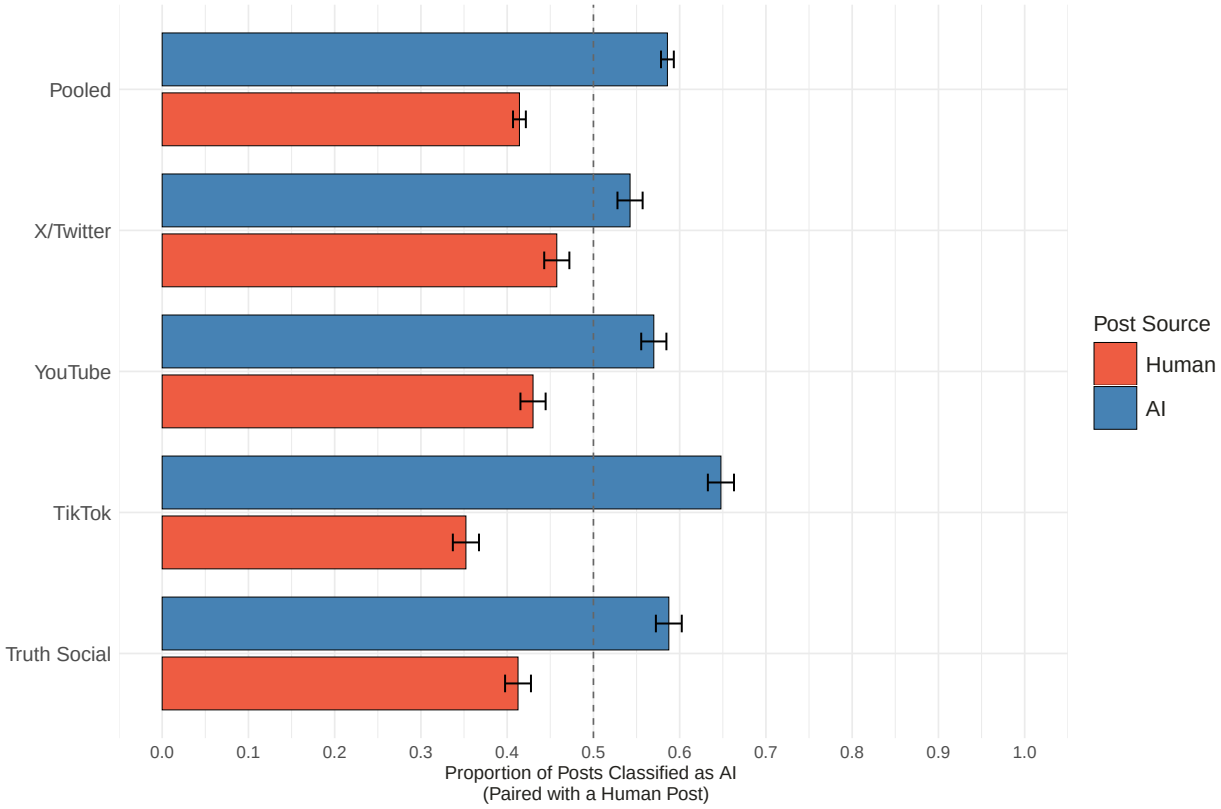
	Pooled	Twitter/X	YouTube	TikTok	Truth Social
Human (Intercept)	0.120***	0.016	0.138**	0.193***	0.151***
	(0.022)	(0.041)	(0.046)	(0.042)	(0.045)
AI-Generated	-0.169***	-0.024	-0.185***	-0.276***	-0.207***
	(0.026)	(0.050)	(0.054)	(0.050)	(0.053)
Num.Obs.	2386	593	600	600	593
R2	0.018	0.000	0.019	0.048	0.025
R2 Adj.	0.017	-0.001	0.018	0.046	0.024

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Each column reports OLS estimates from a separate regression of PHI scores between pooled AI-Generated post versus human posts as the reference category. Coefficients represent the average difference in perceived humanness between AI-Generated posts and human posts. Column 1 pools all platforms; columns 2-5 estimate the same model separately for each platform.
Standard errors in parentheses.

AI Detection Rate in Human-AI Pairings. In this section, we present an additional quantity of interest to understand participants' ability to detect AI content. Instead of using the *PHI* scores, which are calculated via Bradley-Terry model at the post level, we use the AI detection rate, i.e., when participants identify an AI post as AI-generated when paired with another human post. Figure S5 presents pooled results across all models, split by platform, and Figure S6 presents the results for the six AI models used in the data generation process. Lastly, Figure S7 validates these results by plotting the detection rate across bins of the *PHI* scores. Note that the detection rate sums to one within each platform, since the response was a forced-choice selection.

Fig. S5. AI Detection Rates in human-AI Pairings (Pooled Across Models)



Robustness Checks of Emotive Features Analysis Models. To classify emotive features across all posts, we used TweetEval (12) to generate predicted labels for nine features across three categories — Civility (Offensive, Irony, Hateful Speech), Emotion (Anger, Optimism, Joy, Sadness), and Sentiment (Positive, Negative) — and regressed PHI scores on these features using OLS regression. Our primary classifier, referred to as TweetEval (12) in the main text, is implemented using the TweetNLP library (10); results from this primary analysis are reported in Figure 2 of the main manuscript.

These classifiers are standard, validated tools with documented benchmark performance. On the TweetEval test sets, the Twitter-RoBERTa models report macro-F1 of 78.5 (emotion) and 80.5 (offensive), 72.6 macro-recall (sentiment), 61.7 ironic-class F1 (irony), and 52.3 macro-F1 (hate speech) (12).

To assess robustness to the choice of classifier, we replicated this analysis using a second Twitter-specialized model, BERTweet (11), generating labels from both binary predictions and continuous probability scores. The pattern of results is broadly consistent across both classifiers and both labeling approaches, providing confidence that the feature findings reported in the main text are not an artifact of the specific classifier or labeling scheme.

Figures S8 and S9 present the robustness checks across both classifiers using binary labels and continuous probability scores, respectively. Furthermore, Figure S10 presents the marginal effects split by the original source of the posts (Human vs. AI). The effects are homogeneous across these two types, indicating that humans use these semantic cues in the same way when assessing the humanness of social media content, regardless of post origin. We also run a joint hypothesis test for these eight interaction terms to assess if the differences between human and AI-generated posts are jointly distinct from zero. The joint test of the eight interaction terms is not statistically significant ($F(8, 2360) = 1.27$, p-value $= 0.26$). Lastly, Figure S11 presents the distribution of TweetEval probability scores by authorship family.

Fig. S6. AI Detection Rates in Human-AI Pairings per AI-Generated Post Source

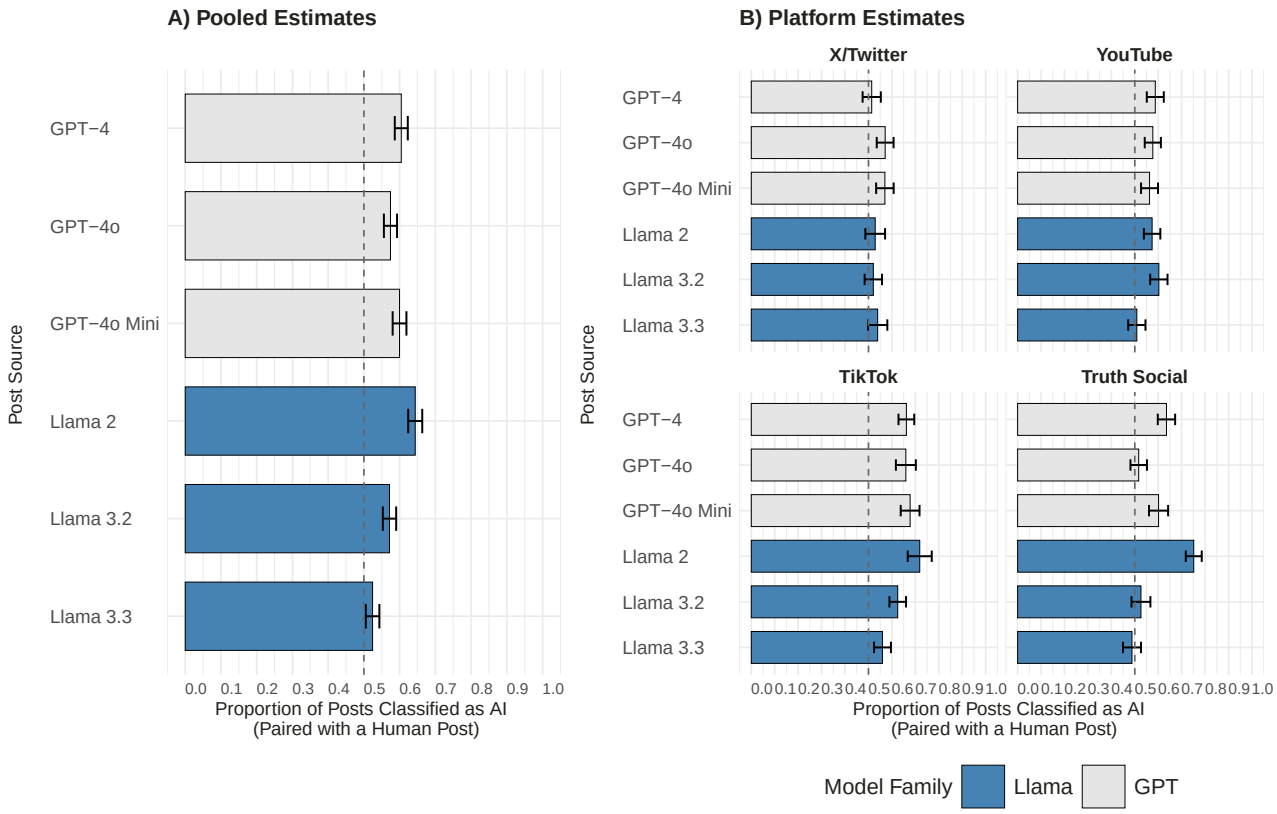


Fig. S7. Validation: AI Detection Rate Across PHI Score Bins

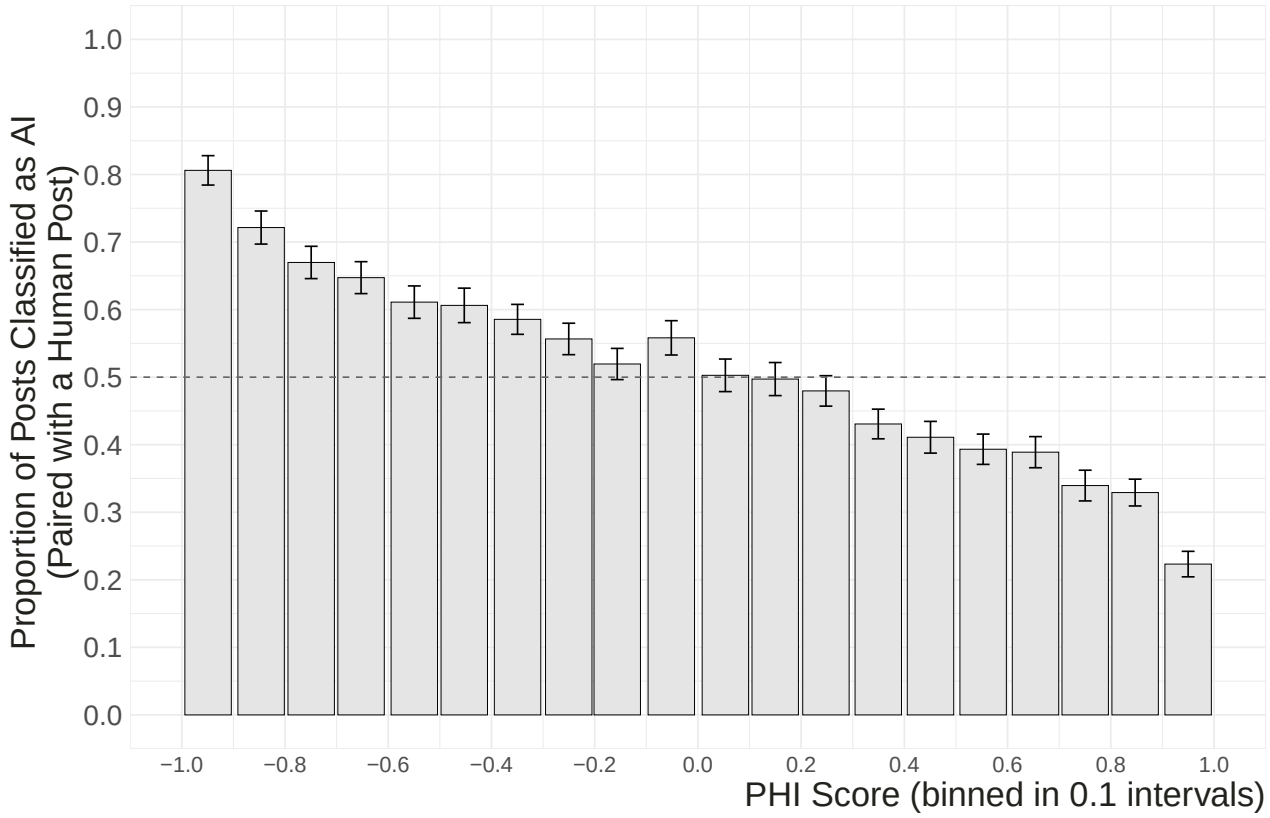


Fig. S8. Impact of emotive text features on perceived humanness using binary classifier labels. Points indicate OLS coefficient estimates and horizontal bars represent 95% confidence intervals. Results are shown separately for TweetEval (blue), and BERTweet (orange).

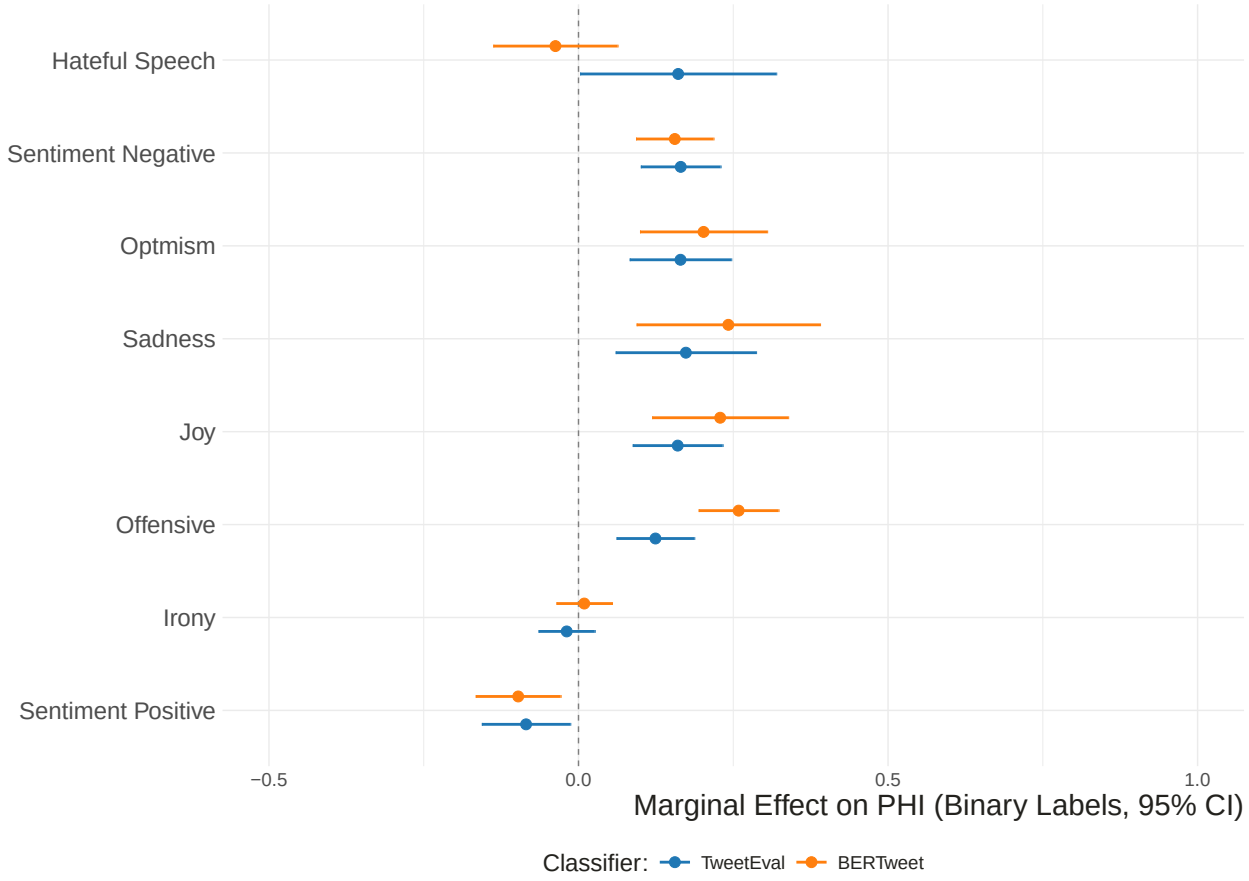


Fig. S9. Impact of emotive text features on perceived humanness using continuous probability scores from each classifier. Points indicate OLS coefficient estimates and horizontal bars represent 95% confidence intervals. Results are shown separately for TweetEval (blue), and BERTweet (orange).

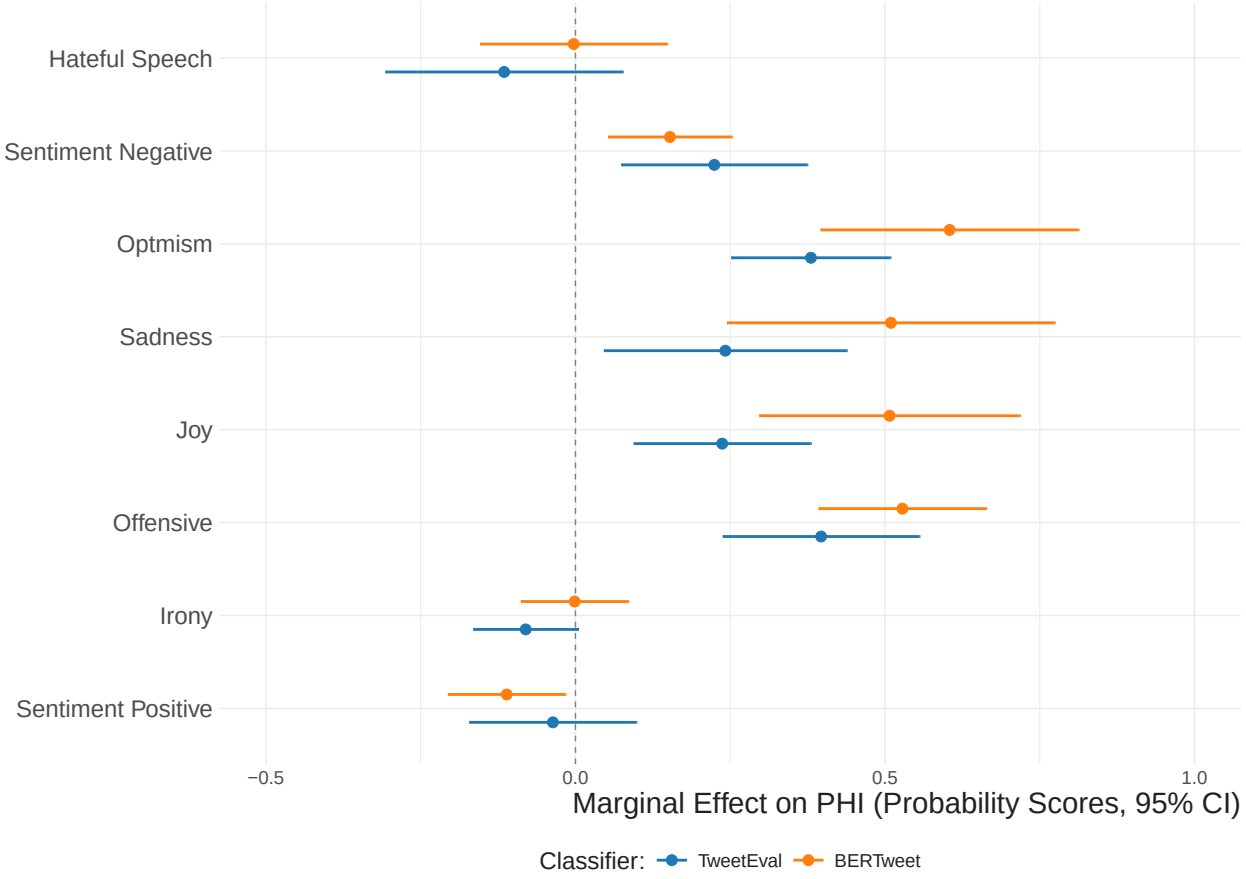


Fig. S10. Emotive Language Features Conditional on Post Source Type

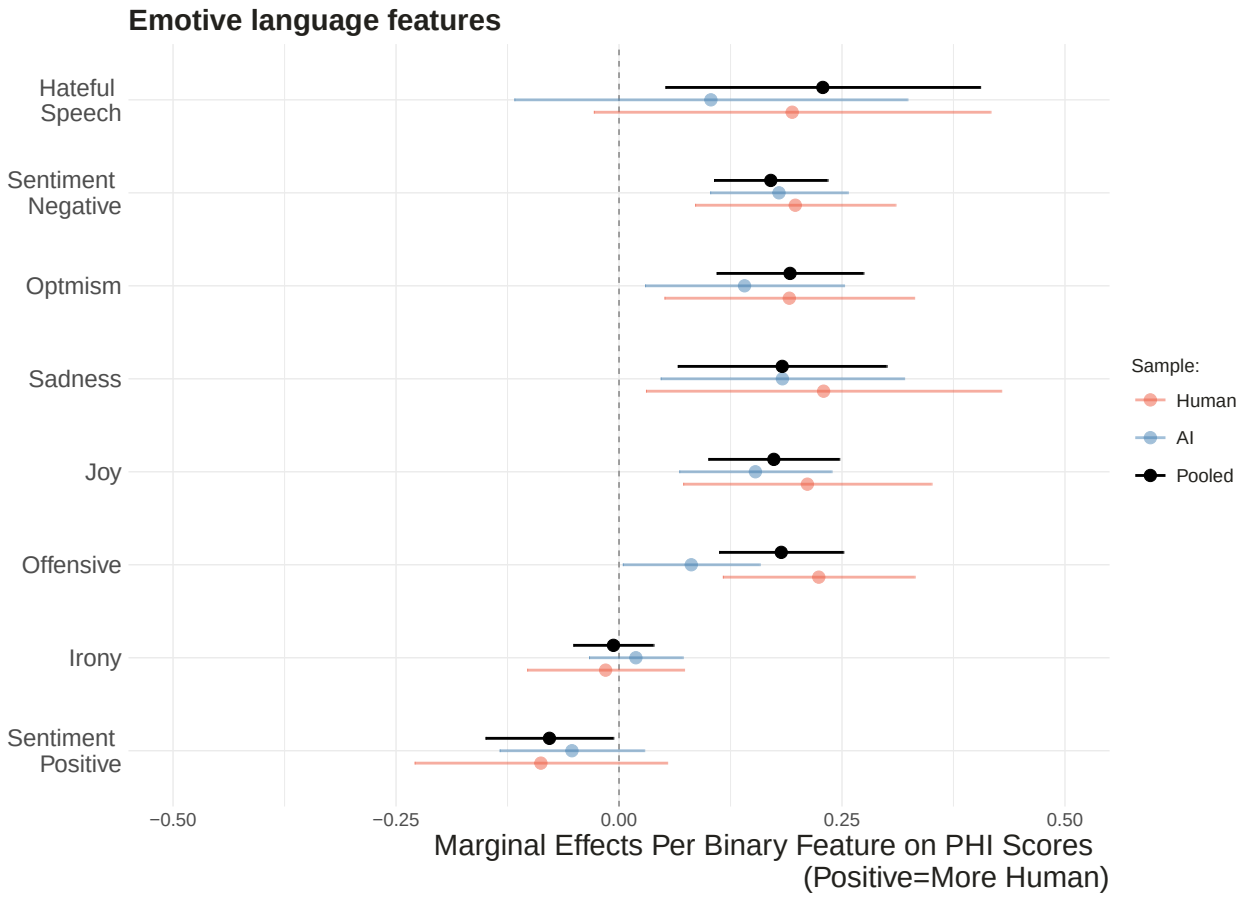
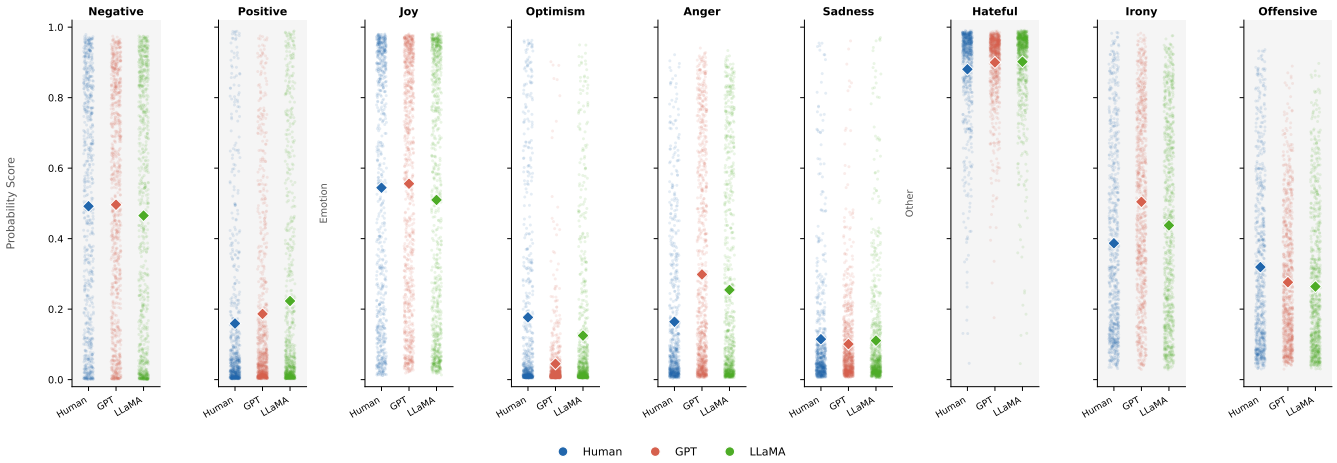


Fig. S11. Distribution of TweetEval probability scores for each emotive text feature by authorship family (Human, GPT, Llama). Each point represents an individual post; diamond markers indicate group means. Features are organized by category: Sentiment (Negative, Positive), Emotion (Joy, Optimism, Anger, Sadness), and Other (Hateful, Irony, Offensive).



Political Lean of Human-Authored Posts. Our AI data generation process used 5 distinct political personas to span a range of viewpoints. We addressed whether the political lean of human posts is comparably distributed using `politicalBiasBERT` (13), a BERT-based classifier that assigns each post to one of three categories: left, center, or right. All posts were classified independently based on their text, with no other metadata provided.

Distribution by platform Table S14 reports the distribution of political lean by platform. Right-leaning content is the most prevalent category across all platforms, ranging from 51.2% on X/Twitter to 69.3% on YouTube. X/Twitter shows the greatest left-leaning representation at 43.3%, while center-leaning posts remain a small minority on every platform (4.6–7.5%).

Table S14. Political lean of human-authored posts by platform, classified by `politicalBiasBERT`. Values are row proportions (%).

Platform	Center	Left	Right	Total
TikTok	7.3	27.4	65.4	179
Truth Social	7.5	32.9	59.6	161
X/Twitter	5.4	43.3	51.2	203
YouTube	4.6	26.1	69.3	153
Total	6.2	33.0	60.8	696

Political lean and humanness scores To test whether the unequal distribution of political perspectives biases our humanness rankings, we regressed our *PHI* scores on political-lean labels using OLS, with left-leaning posts as the reference group. If annotators systematically awarded higher humanness scores to posts of a particular lean, we would expect a significantly large coefficient.

Table S15 reports OLS estimates of the contrasts relative to Left, pooled and by platform. Pooled, Right-leaning posts score significantly higher than Left ($\hat{\beta} = 0.097$, $p = .040$), while Center-leaning posts do not significantly differ from Left ($\hat{\beta} = -0.085$, $p = .374$). The pattern is inconsistent across platforms: on TikTok, Center-leaning posts score significantly lower than Left ($\hat{\beta} = -0.416$, $p = .016$), while no significant differences between groups emerge on X/Twitter, YouTube, or Truth Social. These results indicate that political lean does not produce a systematic or directionally consistent effect on humanness scores.

Fig. S12. Semantic Features Conditional on Platform

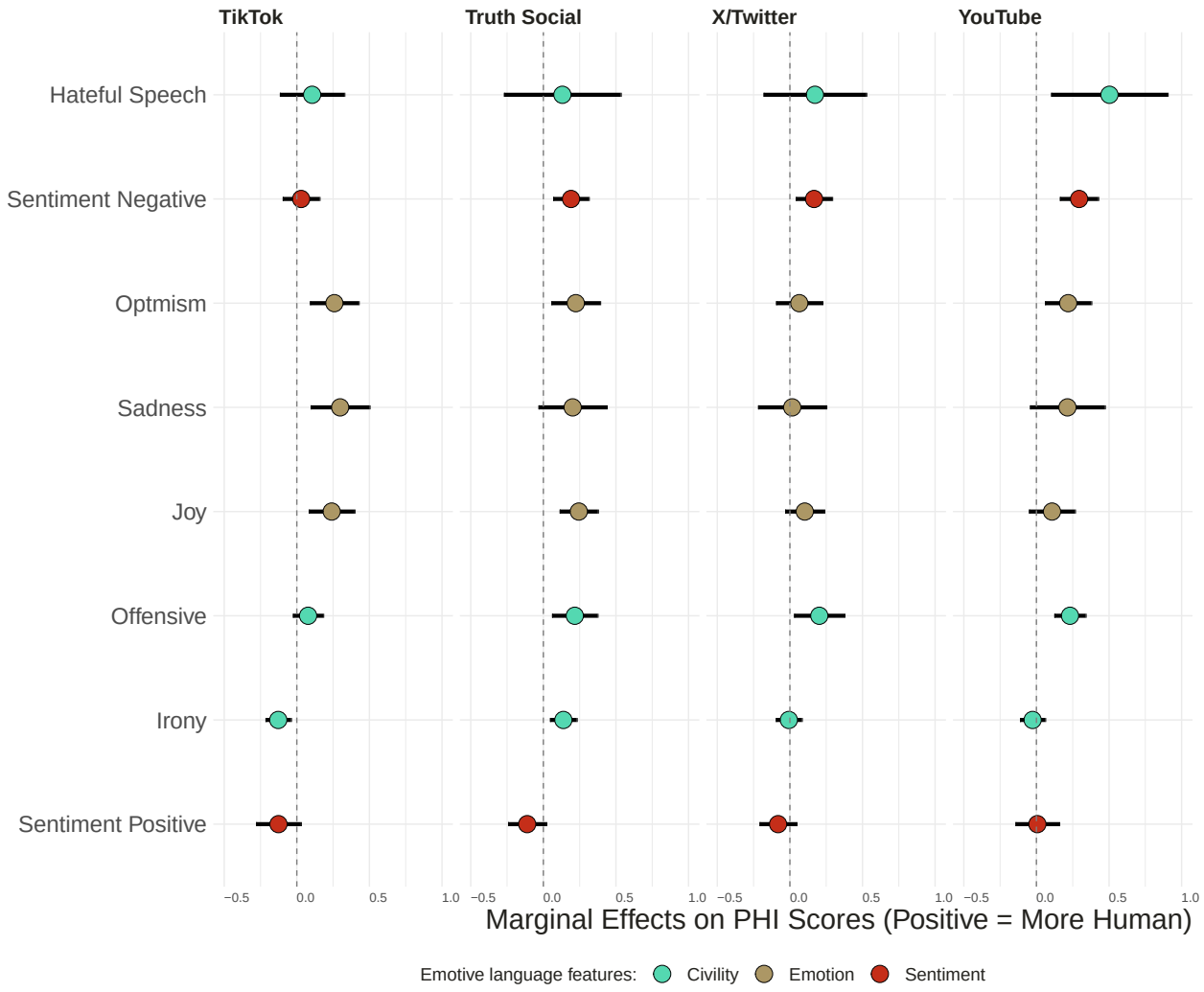


Table S15. OLS Regression: Humanness Score by Political Lean (ref. = Left)

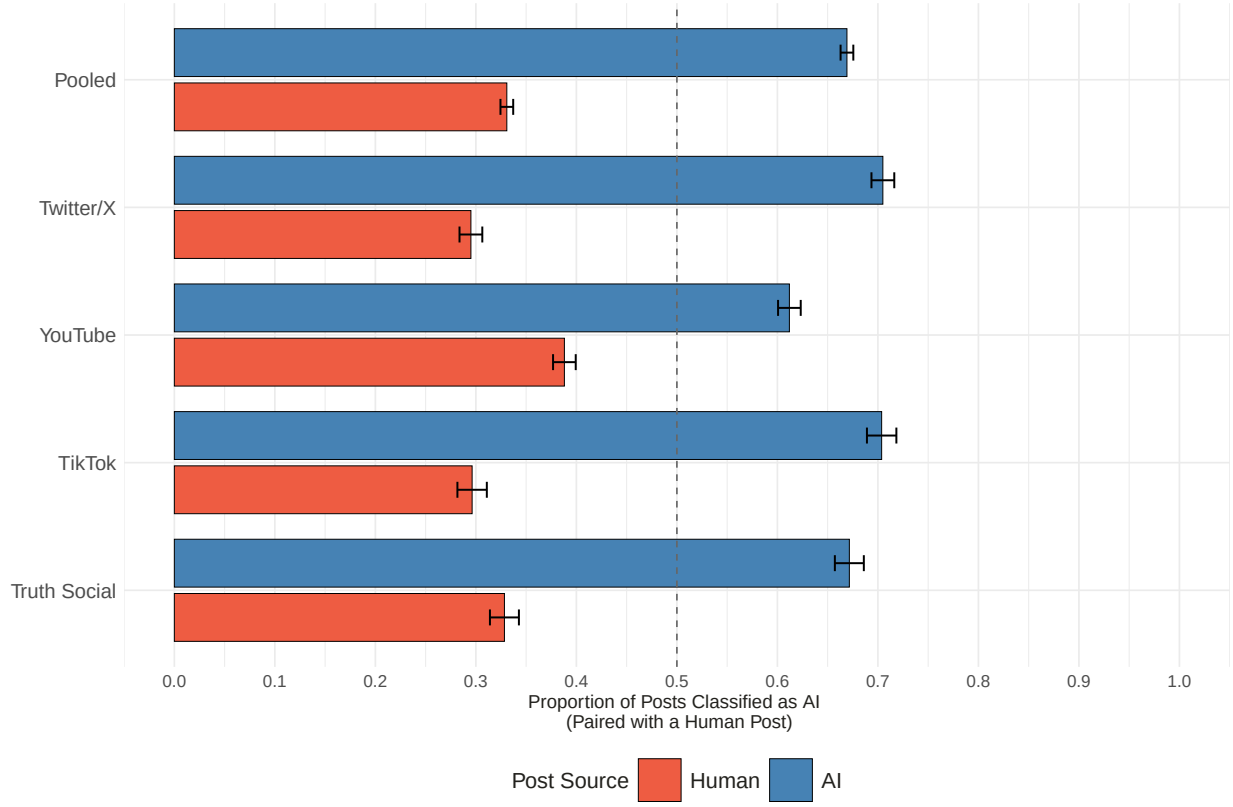
	Pooled	X/Twitter	YouTube	TikTok	Truth Social
Left (Intercept)	0.066	−0.058	0.032	0.241**	0.135
	(0.038)	(0.067)	(0.084)	(0.079)	(0.079)
Center	−0.085	0.061	−0.193	−0.416*	0.077
	(0.096)	(0.201)	(0.218)	(0.173)	(0.183)
Right	0.097*	0.138	0.164	−0.027	0.017
	(0.047)	(0.091)	(0.099)	(0.094)	(0.098)
Num. Obs.	696	203	153	179	161

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Each column reports OLS estimates from a regression of humanness scores on political lean, with Left as the reference group. The Left row reports the intercept (mean of Left posts); Center and Right rows report contrasts relative to Left. Column 1 pools all platforms; columns 2–5 estimate the same model separately for each platform. Standard errors in parentheses.

Benchmarking results with zero-shot GPT-5. In this section, we benchmark our results against a state-of-the-art LLM. We use OpenAI's GPT-5 to perform the same pairwise task as our respondents for all human-AI pairs, enabling us to directly measure the AI detection rate. To do so, we sent the exact same survey question our respondents saw ("Which of the posts (Post A or Post B) is more likely to have been generated by an AI chatbot?") as a prompt to GPT-5 via the OpenAI Chat Completions API (model = gpt-5, temperature = 1, reasoning effort = "minimal"), with an input containing the "Post A" and "Post B". Each pair was scored three times to estimate sampling variance (21,201 calls total), and we select the model response when disagreements emerged. The model gave the same answer across all three independent calls in 83.8% of pairs. Figure S13 presents the results. Results indicate that across all platforms, GPT-5 correctly identified the AI post 66.9% of the time (95% CI [66.3, 67.6]). Per platform, the model did best on Twitter/X (70.5, 95% CI [69.4, 71.6]) and TikTok (70.4%, 95% CI [68.9, 71.8]), slightly lower on Truth Social (67.2, 95% CI [65.7, 68.6]), and weakest on YouTube (61.2%, 95% CI [60.1, 62.3]). Note that the detection rate sums to one within each platform since the response was a forced-choice selection.

Fig. S13. Benchmarking Results: AI Detection Rates of GPT-5 Zero Shot Prompts (Same Pairs as in Human Labels)



S8. Results from Follow-up Survey: Mechanisms Behind the Demographic Differences in AI Detection. The main study shows correlational demographic differences in participants' ability to identify AI-generated posts (Figure 3A). To discuss further the mechanisms behind these demographic differences, we fielded a *follow-up survey* to recontact participants that answered the behavioral experiments described in the paper. This *follow-up survey* focuses on measuring digital media consumption patterns not included in our original instrument. Mainly, the *follow-up survey* asks about AI chatbot use, timing of adoption, breadth of AI tool use, digital literacy, and AI-related attitudes.

The *follow-up* survey was fielded between May 14 and May 25, 2026 through Connect Cloud Research platform. Of the 1,122 respondents who completed the original pairwise task, 723 (∼64%) returned to take the *follow-up*. The instrument was short by design, with a median completion time close to five minutes.

In this section, we present three results using the *follow-up survey*. Table S16 assesses the internal validity of the recontact sample. Table S17 reports the results of how demographics differ in their media habits, AI usage, and digital literacy. Lastly, Table S18 re-estimates the Figure-3B attitudinal model adding regressors for AI-Usage. We interpret these results as suggestive evidence, given the nature of the follow-up study.

Selection Into the Follow-up. Table S16 shows that, compared to our full samples, the participants who answered the *follow-up* are somewhat older and slightly whiter, consume less news online and offline, and exhibit lower trust in online news. The samples do not differ in any of the other variables included in the analysis. Most importantly for internal validity, the two groups do not differ on AI detection ($\beta = 0.00$, $p = 0.75$): the people who returned are no better or worse at the pairwise task than those who did not, supporting the interpretation of the follow-up results as informative about the mechanisms behind this particular outcome.

Table S16. Selection into the follow-up survey: full main-study sample vs. follow-up respondents.

	Full sample (N=1137)		Follow-up (N=723)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p-value
Age: <25	0.09	0.29	0.07	0.25	-0.03	0.04
Age: 25-44	0.37	0.48	0.33	0.47	-0.05	0.03
Age: 45-64	0.40	0.49	0.43	0.50	0.04	0.10
Age: 65+	0.13	0.34	0.17	0.38	0.04	0.04
Male	0.49	0.50	0.49	0.50	-0.01	0.76
White	0.78	0.41	0.82	0.38	0.04	0.05
College degree	0.38	0.48	0.35	0.48	-0.02	0.32
No college degree	0.45	0.50	0.46	0.50	0.01	0.54
Post-graduate degree	0.18	0.38	0.18	0.39	0.01	0.61
Democrat	0.49	0.50	0.49	0.50	0.00	0.97
Republican	0.43	0.50	0.42	0.49	-0.01	0.61
Hispanic	0.16	0.36	0.14	0.34	-0.02	0.21
Online news consumption (z)	0.00	1.00	-0.16	0.84	-0.16	<0.01
Offline news consumption (z)	0.00	1.00	-0.13	0.86	-0.13	<0.01
Trust in online news (z)	-0.00	1.00	-0.10	0.90	-0.10	0.02
Trust in offline news (z)	0.00	1.00	-0.03	0.97	-0.03	0.56
Interest in politics (z)	0.00	1.00	0.00	1.02	0.00	0.94
Conspiracy beliefs (z)	-0.00	1.00	-0.09	1.03	-0.09	0.08
News cynicism (z)	0.00	1.00	0.08	0.99	0.08	0.08
AI detection (prop. correct)	0.59	0.17	0.59	0.17	0.00	0.75

Joint orthogonality test of follow-up participation regressed on all listed variables.
$F(17, 1028) = 5.64$, $p < 0.001$).
Categorical levels are shares of the full sample;
attitudinal/media measures are z-scored on the full sample.

Demographic Mediators of AI Detection. Table S17 presents results of a set of OLS models regressing each candidate mediator related to media and digital habits (AI use frequency, timing of AI adoption, breadth of AI tool use, digital literacy, and the attitudinal and media-consumption measures from Figure 3B) on the demographic battery discussed in the main results. The follow-up adds four measures. *AI frequency*, which we also refer in the manuscript as AI usage, records how often participants report to use AI chatbots on a seven-point scale ranging from never to almost every day. *AI tools* count how many of eight major chatbots (ChatGPT, Gemini, Copilot, Grok, Meta AI, Claude, DeepSeek, and Perplexity) respondents report using at least a few times a year; *AI Adoption timing* records when respondents first started using AI chatbots, from never (1) to early adopters who began in 2023 or before (5); lastly, *Digital literacy* is the mean self-rated familiarity (five-point scale) with five digital tools/concepts: recommendation algorithms, clickbait, phishing, shadowbanning, and content moderation.

We discuss these results in detail in the *Discussion* section of the main manuscript. In summary, we find that men are earlier, and heavier AI users and score higher on digital literacy than women. Meanwhile, white respondents move in the opposite direction on AI usage, and show no statistically significant difference in their digital usage, but rely significantly more on offline news. The two demographic advantages, therefore, appear to run through different mechanisms.

Table S17. Regression Models: Demographic differences in AI Detection Rate and Potential Mechanisms

	AI Detection (Figure 3a)	Interest in Politics	Conspiracy Believes	Cynicism	Trust online	Trust offline	News online	News offline	AI Frequency	AI tools	AI Adoption timing	Digital literacy
White (vs Non-White)	0.039*	-0.024	-0.279***	0.202*	-0.538***	-0.299***	-0.713***	-0.499***	-0.441*	-1.014***	-0.400***	-0.128
	(0.016)	(0.081)	(0.082)	(0.082)	(0.079)	(0.080)	(0.076)	(0.082)	(0.191)	(0.227)	(0.108)	(0.106)
Hispanic	-0.009	-0.125	0.155	-0.158	0.034	-0.238**	0.187*	-0.047	0.297	0.082	0.033	0.147
	(0.016)	(0.086)	(0.087)	(0.087)	(0.084)	(0.085)	(0.081)	(0.087)	(0.202)	(0.240)	(0.115)	(0.112)
Male (vs Female)	0.022*	-0.314***	0.034	0.091	0.006	0.062	0.097	0.126*	0.342**	0.862***	0.218**	0.467***
	(0.011)	(0.059)	(0.059)	(0.059)	(0.058)	(0.058)	(0.055)	(0.060)	(0.130)	(0.154)	(0.073)	(0.072)
Age 25-44 (vs <25)	-0.005	-0.097	0.016	0.055	-0.306**	0.240*	-0.288**	0.066	0.138	-0.151	-0.320*	-0.029
	(0.020)	(0.106)	(0.107)	(0.107)	(0.104)	(0.106)	(0.100)	(0.108)	(0.275)	(0.327)	(0.159)	(0.154)
Age 45-64 (vs <25)	0.010	-0.386***	-0.158	0.298**	-0.574***	0.329**	-0.614***	-0.041	0.352	-0.130	-0.301	-0.368*
	(0.021)	(0.109)	(0.110)	(0.111)	(0.108)	(0.109)	(0.103)	(0.111)	(0.283)	(0.336)	(0.163)	(0.158)
Age 65+ (vs <25)	-0.016	-0.545***	-0.336**	0.437***	-0.620***	0.477***	-0.797***	0.061	0.035	-0.281	-0.367*	-0.821***
	(0.025)	(0.129)	(0.130)	(0.131)	(0.127)	(0.129)	(0.122)	(0.132)	(0.310)	(0.368)	(0.178)	(0.173)
College (vs No College)	0.005	-0.238***	-0.077	-0.030	0.210**	0.260***	0.172**	0.276***	-0.074	0.255	0.096	0.066
	(0.012)	(0.065)	(0.066)	(0.066)	(0.064)	(0.065)	(0.062)	(0.066)	(0.142)	(0.169)	(0.080)	(0.079)
Post-Grad (vs No College)	0.011	-0.204*	-0.240**	0.150	0.108	0.122	0.135	0.278***	0.314	0.470*	0.137	0.069
	(0.015)	(0.082)	(0.083)	(0.083)	(0.081)	(0.082)	(0.078)	(0.084)	(0.174)	(0.207)	(0.098)	(0.096)
Republican (vs Democrat)	0.005	0.192**	0.356***	-0.374***	-0.177**	-0.399***	0.237***	0.208***	0.189	0.598***	0.018	-0.148*
	(0.011)	(0.061)	(0.061)	(0.061)	(0.060)	(0.061)	(0.057)	(0.062)	(0.135)	(0.160)	(0.076)	(0.075)
N	15457	1103	1103	1103	1103	1103	1103	1103	652	652	625	650

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

AI Detection = AME on P(correctly identifying AI) from a binomial GLMER (logit). For the first nine columns, we use OLS models in the full samples from the original instrument.

For the last four, results are based on OLS models with follow-up respondents. Reference Categories: Non-White / Female / under 25 / No College / Democrat.

The effects of AI Usage on AI Detection Rate. Finally, we ask whether AI usage, early adoption, breadth of AI tools used, and digital literacy itself predict AI detection rate, in addition to the attitudinal and media-consumption measures already in Figure 3 Panel B. To do so, in Table S18, we present the logit estimates as in Figure 3B attitudinal model on the full sample, and the same model re-estimated in the follow-up sample, plus four additional regressors: AI use frequency, breadth of AI tools, timing of AI adoption, and digital literacy.

Table S18 indicates that higher AI usage improves AI detection in our AI-Human pairwise tasks ($\beta_{\text{logit}} = 0.138$, $SE = 0.039$, $p < 0.001$. The other additional AI variables are not statistically significant, once AI use frequency is in the model. Most importantly, the attitudinal pattern from Figure 2B holds: offline news consumption continues to predict worse AI detection in this new model.

Table S18. Regression Models: Figure 3B + AI Usage and Digital Literacy (Follow-up)

	(1) Attitudinal (full)	(2) + AI usage & Digital literacy (follow-up)
Online news consumption	0.016	-0.011
	(0.032)	(0.046)
Offline news consumption	-0.067*	-0.092*
	(0.028)	(0.039)
Trust news online	-0.046	-0.028
	(0.030)	(0.039)
Trust news offline	-0.045	-0.055
	(0.027)	(0.035)
Interest in politics	-0.029	-0.005
	(0.024)	(0.031)
Conspiracy beliefs	-0.018	-0.013
	(0.024)	(0.030)
News cynicism	0.032	0.040
	(0.024)	(0.032)
AI use frequency		0.138***
		(0.039)
AI tools		0.012
		(0.036)
AI adoption timing		-0.011
		(0.032)
Digital literacy		0.053
		(0.032)
(Intercept)	0.442***	0.348***
	(0.086)	(0.100)
Num.Obs.	16851	9610

* p < 0.05, ** p < 0.01, *** p < 0.001
Logit models. Outcome: Correctly identifying the AI post in human-vs-AI pairs.
Continuous predictors z-scored. Random intercepts for participant and trial order.
Model 1 uses the full main-study sample;
Model 2 adds AI usage and digital literacy on the follow-up subsample.

S9. Reproducibility and Availability.

- **Data**: All the data will be available at an open source repository upon publication
- **Code**: All code to reproduce all the analysis will be available at an open source repository upon publication
- **IRB**: This study was approved by Georgetown IRB, under the project STUDY00008554.

Table S19. Demographic characteristics of Survey 1 and Survey 2 compared with the US adult general population.

	Survey 1		Survey 2		Total		Gen Pop
Value	Count	%	Count	%	Count	%	%
Age Range							
18-24	60	10.0	57	9.4	117	9.7	11.7
25-34	134	22.3	146	24.1	280	23.2	17.3
35-44	94	15.6	89	14.7	183	15.2	17.2
45-54	113	18.8	121	20.0	234	19.4	15.3
55-64	116	19.3	117	19.3	233	19.3	15.6
65-74	73	12.1	67	11.1	140	11.6	13.3
75-84	11	1.8	9	1.5	20	1.7	7.2
85+	0	0.0	0	0.0	0	0.0	2.4
Education							
Graduate or professional degree (Master's, MD, JD, PhD)	104	17.3	103	17.0	207	17.1	14.7
Bachelor's degree	209	34.8	249	41.1	458	37.9	22.1
Associate degree	77	12.8	60	9.9	137	11.4	8.8
Some college, no degree	120	20.0	114	18.8	234	19.4	18.5
High school graduate (GED)	84	14.0	76	12.5	160	13.3	25.7
Less than high school	5	0.8	2	0.3	7	0.6	10.1
No formal education	1	0.2	0	0.0	1	0.1	NA
Prefer not to say	1	0.2	2	0.3	3	0.2	NA
Sex							
Male	303	50.4	299	49.3	602	49.9	49.0
Female	297	49.4	307	50.7	604	50.0	51.0
Prefer not to say	1	0.2	0	0.0	1	0.1	NA
Political Party							
Republican (with leaners)	259	43.1	265	43.7	524	43.4	46.0
Democrat (with leaners)	293	48.8	293	48.3	586	48.6	45.0
Independent	49	8.2	48	7.9	97	8.0	9.0
Race							
White	469	78.0	469	77.4	938	77.7	75.3
Black or African American	84	14.0	89	14.7	173	14.3	13.7
Asian	27	4.5	33	5.4	60	5.0	6.4
Native Hawaiian or Pacific Islander	1	0.2	1	0.2	2	0.2	0.3
American Indian or Alaska Native	3	0.5	2	0.3	5	0.4	1.3
Other / Not listed	17	2.8	12	2.0	29	2.4	3.0
Ethnicity							
Hispanic or Latino	97	16.1	97	16.0	194	16.1	19.5
Non-Hispanic	504	83.9	509	84.0	1013	83.9	80.5

Notes: Sources: Survey 1 (X/YouTube) and Survey 2 (TikTok/Truth Social). General population reference values are from US Census Bureau (ACS 1-year 2024, Population Estimates Program 2024) for age, education, sex, race, ethnicity, and employment; Pew Research NPORS (fielded February–June 2025) for political party (with-leaner classification).

328 **S10. Survey Descriptives.**

Table S20. Survey responses for political interest and news trust. Interest in Politics: "How often do you keep up with politics, regardless of the source—whether through TV, radio, newspapers, or online platforms?" Trust in News: "How much, if at all, do you trust the information you get from..."

		Survey 1		Survey 2	
Measure	**Response**	**N**	**%**	**N**	**%**
Interest in Politics	Extremely closely	94	15.5	99	16.3
	Very closely	193	31.7	190	31.4
	Moderately closely	207	34.0	195	32.2
	Slightly closely	91	15.0	98	16.2
	Not at all	23	3.8	24	4.0
Trust - Newspapers	Not at all	36	5.9	45	7.4
	A little	131	21.5	137	22.6
	Some	260	42.8	254	41.9
	Quite a lot	153	25.2	139	22.9
	All the time	28	4.6	31	5.1
Trust - Social media	Not at all	192	31.6	192	31.7
	A little	265	43.6	253	41.7
	Some	121	19.9	127	21.0
	Quite a lot	26	4.3	29	4.8
	All the time	4	0.7	5	0.8
Trust - News websites	Not at all	34	5.6	42	6.9
	A little	130	21.4	134	22.1
	Some	268	44.1	254	41.9
	Quite a lot	153	25.2	145	23.9
	All the time	23	3.8	31	5.1
Trust - Radio news	Not at all	50	8.2	54	8.9
	A little	139	22.9	153	25.2
	Some	248	40.8	237	39.1
	Quite a lot	143	23.5	135	22.3
	All the time	28	4.6	27	4.5
Trust - Local news	Not at all	27	4.4	32	5.3
	A little	107	17.6	112	18.5
	Some	247	40.6	252	41.6
	Quite a lot	193	31.7	179	29.5
	All the time	34	5.6	31	5.1
Trust - National news	Not at all	60	9.9	77	12.7
	A little	145	23.8	154	25.4
	Some	226	37.2	216	35.6
	Quite a lot	147	24.2	131	21.6
	All the time	30	4.9	28	4.6

Table S21. News cynicism responses

	Survey 1		Survey 2	
Item / Response	**N**	**%**	**N**	**%**
You cannot trust news stories people share on social media				
Strongly Disagree	10	1.7	1	0.2
Disagree	13	2.2	13	2.1
Somewhat Disagree	34	5.7	43	7.1
Neutral	90	15.0	84	13.9
Somewhat Agree	193	32.1	211	34.8
Agree	150	25.0	160	26.4
Strongly Agree	111	18.5	94	15.5
Credible news is mixed with misinformation on social media				
Strongly Disagree	4	0.7	2	0.3
Disagree	5	0.8	3	0.5
Somewhat Disagree	14	2.3	11	1.8
Neutral	45	7.5	37	6.1
Somewhat Agree	142	23.6	154	25.4
Agree	213	35.4	221	36.5
Strongly Agree	178	29.6	178	29.4
You should not rely on social media for news information				
Strongly Disagree	13	2.2	11	1.8
Disagree	32	5.3	26	4.3
Somewhat Disagree	48	8.0	50	8.3
Neutral	72	12.0	63	10.4
Somewhat Agree	126	21.0	144	23.8
Agree	137	22.8	145	23.9
Strongly Agree	173	28.8	167	27.6
I feel confident I can find the truth about political issues				
Strongly Disagree	13	2.2	14	2.3
Disagree	23	3.8	35	5.8
Somewhat Disagree	59	9.8	54	8.9
Neutral	96	16.0	91	15.0
Somewhat Agree	167	27.8	202	33.3
Agree	188	31.3	166	27.4
Strongly Agree	55	9.2	44	7.3

Table S22. Conspiracy beliefs responses

Item / Response	Survey 1 N	Survey 1 %	Survey 2 N	Survey 2 %
Much of our lives are being controlled by plots hatched in secret places				
Strongly disagree	151	25.1	162	26.7
Somewhat disagree	120	20.0	125	20.6
Neither agree nor disagree	128	21.3	119	19.6
Somewhat agree	150	25.0	140	23.1
Strongly agree	52	8.7	60	9.9
Even though we live in a democracy a few people will always run things anyway				
Strongly disagree	62	10.3	51	8.4
Somewhat disagree	73	12.1	86	14.2
Neither agree nor disagree	89	14.8	79	13.0
Somewhat agree	245	40.8	270	44.6
Strongly agree	132	22.0	120	19.8
The people who really 'run' the country are not known to the voter				
Strongly disagree	104	17.3	101	16.7
Somewhat disagree	101	16.8	125	20.6
Neither agree nor disagree	108	18.0	101	16.7
Somewhat agree	187	31.1	200	33.0
Strongly agree	101	16.8	79	13.0
Big events like wars, recessions, and the outcomes of elections are controlled by small groups of people who are working in secret against the rest of us				
Strongly disagree	135	22.5	147	24.3
Somewhat disagree	129	21.5	130	21.5
Neither agree nor disagree	101	16.8	123	20.3
Somewhat agree	179	29.8	150	24.8
Strongly agree	57	9.5	56	9.2

Table S23. Descriptive statistics of normalized PHI scores across platforms. (a) X, (b) YouTube, (c) Truth Social, (d) TikTok.

*(a) X/Twitter

Model	Count	Mean	SD	Min	Max	Median
Human	203	0.02	0.63	-1.00	1.00	0.02
GPT-4	67	0.03	0.54	-1.00	0.99	0.11
GPT-4o	67	-0.27	0.46	-0.99	0.84	-0.32
GPT-4o Mini	66	-0.01	0.58	-0.97	0.93	-0.05
Llama 2	56	0.03	0.63	-0.95	0.98	0.06
Llama 3.2	74	0.09	0.51	-0.93	0.96	0.06
Llama 3.3	60	0.10	0.54	-0.98	0.94	0.23

*(b) YouTube

Model	Count	Mean	SD	Min	Max	Median
Human	153	0.14	0.53	-0.92	0.99	0.19
GPT-4	75	-0.17	0.53	-0.99	1.00	-0.23
GPT-4o	75	-0.08	0.57	-0.98	0.98	-0.14
GPT-4o Mini	75	-0.07	0.61	-0.97	0.99	-0.24
Human	153	0.14	0.53	-0.92	0.99	0.19
Llama 2	73	-0.18	0.54	-1.00	0.89	-0.17
Llama 3.2	74	-0.08	0.62	-1.00	0.98	-0.08
Llama 3.3	75	0.29	0.53	-0.99	1.00	0.39

*(c) Truth Social

Model	Count	Mean	SD	Min	Max	Median
Human	161	0.15	0.57	-0.99	0.98	0.19
GPT-4	75	-0.05	0.54	-0.98	0.99	-0.06
GPT-4o	72	-0.08	0.53	-0.96	0.92	-0.11
GPT-4o Mini	75	-0.03	0.51	-0.93	1.00	-0.02
Llama 2	73	-0.54	0.41	-1.00	0.77	-0.63
Llama 3.2	66	0.02	0.59	-0.99	1.00	0.05
Llama 3.3	71	0.36	0.48	-0.71	0.99	0.42

*(d) TikTok

Model	Count	Mean	SD	Min	Max	Median
Human	179	0.19	0.56	-0.97	0.99	0.27
GPT-4	75	-0.20	0.48	-0.97	0.99	-0.27
GPT-4o	75	-0.09	0.58	-1.00	0.98	-0.03
GPT-4o Mini	75	-0.15	0.56	-0.99	0.95	-0.19
Llama 2	47	-0.38	0.51	-1.00	0.91	-0.53
Llama 3.2	76	0.08	0.55	-0.99	0.96	0.18
Llama 3.3	73	0.13	0.60	-0.97	1.00	0.23

Table S24. Summary of marginal mean on pooled- and platform-level estimates on PHI scores, corresponding to Fig. 1

Model	Pooled	Twitter/X	YouTube	TikTok	Truth Social
Human	0.119 [0.078, 0.161]	0.016 [-0.060, 0.092]	0.138 [0.050, 0.225]	0.193 [0.112, 0.274]	0.151 [0.065, 0.236]
GPT-4	-0.100 [-0.165, -0.036]	0.029 [-0.104, 0.161]	-0.173 [-0.298, -0.048]	-0.195 [-0.321, -0.070]	-0.048 [-0.173, 0.077]
GPT-4o	-0.125 [-0.190, -0.061]	-0.273 [-0.406, -0.141]	-0.076 [-0.201, 0.049]	-0.086 [-0.211, 0.039]	-0.080 [-0.207, 0.048]
GPT-4o Mini	-0.066 [-0.131, -0.002]	-0.013 [-0.147, 0.120]	-0.073 [-0.198, 0.052]	-0.148 [-0.273, -0.022]	-0.025 [-0.150, 0.100]
Llama 2	-0.275 [-0.345, -0.206]	0.029 [-0.115, 0.174]	-0.176 [-0.303, -0.049]	-0.378 [-0.536, -0.220]	-0.543 [-0.670, -0.416]
Llama 3.2	0.026 [-0.039, 0.090]	0.087 [-0.039, 0.213]	-0.082 [-0.208, 0.044]	0.078 [-0.047, 0.202]	0.018 [-0.116, 0.151]
Llama 3.3	0.225 [0.159, 0.290]	0.098 [-0.042, 0.238]	0.293 [0.168, 0.418]	0.129 [0.002, 0.256]	0.358 [0.229, 0.486]

S11. Survey Instruments. This section reproduces the full wording and administration order of the three surveys used in the study. Survey 1 (X/Twitter and YouTube) and Survey 2 (Truth Social and TikTok) used the same instrument and differed only in the platform stimuli and a small number of items, listed at the end of Section S11.1. The sample sizes for both surveys are reported in Section S5. In the pairwise tasks, individual posts were piped into each comparison from the platform-specific post pools described in Table S3 and Tables S8–S11; we reproduce each repeated task once and indicate the number of repetitions. The third, follow-up survey (May 2026), described in Section S8, recontacted prior participants to measure AI familiarity, use, and perceptions. Bracketed italics denote piping, branching, or administration notes and were not shown to participants.

S11.1 Survey 1 and Survey 2: Pairwise Humanness Comparison. Participants first viewed an IRB-approved informed-consent form (Research Title: "Labeling Humanness on Social Media Data"; Georgetown IRB) summarizing the study purpose, expected duration, minimal risks, voluntary participation, and data handling. They then answered: *Do you consent to participating in this study?*

- Yes, I would like to participate in this study.
- No, I do not want to participate in this study.

[Participants then saw: "We want to start with some questions about you."]

Political attention. How often do you keep up with politics, regardless of the source—whether through TV, radio, newspapers, or online platforms?

- Extremely closely
- Very closely
- Moderately closely
- Slightly closely
- Not at all

News trust. How much, if at all, do you trust the information you get from… *[Matrix; each row rated on the scale below.]*

- Newspapers
- Social media
- News websites
- Radio news
- Local news organizations
- National news organizations (i.e., ABC, CBS, NBC)

Scale: Not at all; A little; Some; Quite a lot; All the time

[Survey 2 only – a Social media usage battery was inserted here. See the Survey 2 differences at the end of this subsection.]

Party identification. Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent or something else?

- Republican
- Democrat
- Independent
- Something else

Party lean. *[Shown only if Party identification = Independent or Something else.]* Do you think of yourself as closer to the Republican Party or to the Democratic Party?

- Closer to the Republican Party
- Closer to the Democratic Party
- Neither

Ideological self-placement. When it comes to politics, would you describe yourself as liberal, conservative, or neither liberal nor conservative?

- Very liberal
- Somewhat liberal
- Slightly liberal
- Moderate, middle of the road
- Slightly conservative
- Somewhat conservative
- Very conservative

Conspiracy beliefs. How strongly do you agree or disagree with the following statements: *[Matrix; each row rated on the scale below.]*

- Much of our lives are being controlled by plots hatched in secret places.
- Even though we live in a democracy a few people will always run things anyway.
- The people who really 'run' the country are not known to the voter.
- Big events like wars, recessions, and the outcomes of elections are controlled by small groups of people who are working in secret against the rest of us.

Scale: Strongly disagree; Somewhat disagree; Neither agree nor disagree; Somewhat agree; Strongly agree

Social media cynicism. Please indicate your agreement with the following statements: *[Matrix; each row rated on the scale below.]*

- You cannot trust the news stories people share on social media
- Too often credible news information is mixed up with misinformation on social media
- You should not rely on social media for news information
- I feel confident that I can find the truth about political issues

Scale: Strongly Disagree; Disagree; Somewhat Disagree; Neutral; Somewhat Agree; Agree; Strongly Agree

Task instructions. Social Media Task. For the next questions, you will be asked to compare two social media posts and label if you believe they were produced by an AI (artificial intelligence) chatbot or by a human. You will be shown a series of pairs of social media posts. For each pair, you must pick the post that seems more likely to be produced by an AI chatbot. Even if you are not certain, we ask you to decide which of the two posts from each pair is more likely to have been produced by an AI chatbot. Important: You will repeat this task 15 times, and sometimes, the same message might be repeated in a different pair.

Pairwise comparison item. *[First task block. Administered for 15 pairs. Posts piped from the first platform pool: YouTube in Survey 1, TikTok in Survey 2.]*

- *Prompt:* Please, read the social media posts below carefully.
- *Post A:* [post piped from the platform pool]
- *Post B:* [post piped from the platform pool]
- *Question:* Which of the posts (Post A or Post B) is more likely to have been generated by an AI chatbot?
- *Response options:* Post A; Post B

[Embedded attention check at pair 10: Post A was replaced with the text "An AI model has generated this social media post. Please select the option POST A below."]

[Interstitial: "Thank you for answering the first set of comparisons between social media posts. Before we ask you to label other examples for us, we have a few more questions for you."]

News consumption. How often do you use each of the following as a source of news? By news, we mean national, international, regional/local news, and other topical events. *[Matrix; each row rated on the scale below.]*

- Television news (including streaming)

- Radio or audio news (including podcasts)
- Printed newspapers
- Websites/apps of newspapers and news outlets
- From talking with friends, family and co-workers
- Facebook
- X/Twitter
- Tiktok
- Youtube

Scale: Never; Sometimes; About half the time; Most of the time; Always

Attention check. Paying attention and reading the instructions carefully is critical. President Obama is the youngest politician listed in the options below. If you are paying attention, please choose the youngest politician from the list.

- President Joe Biden
- President Trump
- President Obama
- Nancy Pelosi
- Mike Pence
- Don't know

Task instructions (repeated). *[The same Social Media Task instructions shown above were displayed again before the second platform block.]*

Pairwise comparison item. *[Second task block. Administered for 15 pairs. Posts piped from the second platform pool: X/Twitter in Survey 1, Truth Social in Survey 2. Format identical to the first block, including the embedded attention check at pair 10.]*

[Interstitial: "Thank you for completing the second set of comparisons of social media posts. Now we have some questions for you before you label the final set of posts."]

Misinformation literacy. How likely are you to do any of the following when navigating social media and news websites? *[Matrix; each row rated on the scale below. Survey 1 wording shown; Survey 2 rewordings are listed in the Survey 2 differences below.]*

- Not share a news story when I am unsure about its accuracy
- Check a number of different sources to see whether a news story is reported in the same way
- Rely on sources of news that are considered more reputable
- Stop using certain news sources when I am unsure about the accuracy of their reporting
- Discuss a news story with a person I trust because I am unsure about its accuracy
- Stop paying attention to news shared by someone because I am unsure whether I trust that person

Scale: Extremely unlikely; Somewhat unlikely; Neither likely nor unlikely; Somewhat likely; Extremely likely

Final task instructions. Final Social Media Task. For the next questions, you will be asked to look at 5 social media posts and label if you believe they were produced by an AI (artificial intelligence) chatbot or by a human. You will look to a single post, not a pair. You will repeat this task 5 times.

Single-post humanness rating. *[Administered for 5 single posts: 4 rating items plus the attention check below. Posts piped from the platform pools.]*

- *Prompt:* Please consider the text below.
- *Social Media Post:* [post piped from the platform pool]
- *Question:* Now, indicate how likely you think this social media post was human-written or AI-generated.

- *Response options:* Strongly Human; Likely Human; Human; I am not sure; AI; Likely AI; Strongly AI

[Single-post attention check: the post text read "An AI model has generated this social media post. Please select the option below 'Strongly AI'." Same 7-point response scale.]

Open-ended cues. In the previous tasks, you identified whether social media posts were produced by an AI chatbot or a human. Now, we want to understand more about how you decided. Please explain any cues, words, style, ideas, or anything else that led you to decide if the posts were produced by an AI (artificial intelligence) chatbot or by a human. *[Open response.]*

Cue importance. Please rate the importance of the following factors in determining whether a social media post was likely written by a human. *[Matrix; each row rated on the scale below.]*

- Length of the post
- Grammar (e.g. punctuation, spelling)
- Use of emojis, symbols, or hashtags
- Use of slang or abbreviations
- Emotional language
- Ideological stance (e.g. political, religious)

Scale: Not important; Slightly important; Moderately important; Very important; Extremely important

Feedback. Do you have any other feedback or comment to our research team about your experience taking this survey? *[Open response.]*

Differences between Survey 1 and Survey 2. Survey 2 used the same instrument as Survey 1 with the following changes:

1. The platform stimuli were Truth Social and TikTok. The first pairwise block used TikTok and the second used Truth Social.
2. A social media usage battery was added to the pre-task block, immediately after News trust: "How often do you use each of the following social media platforms?" Rows: YouTube; TikTok; X (Formerly Twitter); Truth Social; Instagram; Facebook; Reddit. Scale: Not at all; A little; Some; Quite a lot; All the time.

S11.2 Follow-up Survey (May 2026). The follow-up survey (described in Section S8) recontacted prior participants and measured AI familiarity, use, and perceptions. It is reproduced below in administration order.

Consent. Participants viewed the same IRB-approved consent form as Survey 2 (Research Title: "Perceived Humanness of Social Media Posts"; Georgetown IRB) and answered: *Do you consent to participating in this study?* (Yes, I would like to participate in this study; No, I do not want to participate in this study.)

[Participants then saw: "First, we want to know about your usage of and thoughts on Artificial Intelligence (AI) chatbots."]

AI use frequency. How often do you use AI chatbots (e.g. ChatGPT, Gemini, Claude, etc.)? *[If "Never," skip to AI detection confidence.]*

- Never
- About once a year
- Several times a year
- About once a month
- Several times a month
- Once or twice a week
- Almost every day

AI use by chatbot. How frequently do you use the following AI chatbots in your daily life? *[Matrix; each row rated on the scale below.]*

- ChatGPT (OpenAI)
- Gemini (Google)
- Copilot (Microsoft)

- Grok (xAI)
- Meta AI
- Claude (Anthropic)
- DeepSeek
- Perplexity

Scale: Never; About once a year; Several times a year; About once a month; Several times a month; Once or twice a week; Almost every day; Multiple times a day

AI adoption timing. When did you start using AI chatbots?

- I have never used an AI chatbot
- Recently, only this year (2026)
- Around last year (2025)
- About two years ago (2024)
- Early, when these tools were launched (2023 or earlier)

AI use purposes. For what purposes do you generally use AI chatbots? (select all that apply)

- Keep up with politics and news
- Get information or answers to questions (instead of or alongside a search engine)
- Help with work or education-related tasks
- Collaborate on creative tasks (e.g., writing, brainstorming)
- For entertainment purposes
- For companionship
- Conduct health-related information seeking
- For shopping or product research
- Use for travel planning
- Other [open response]

AI detection confidence. How confident are you that you would be able to tell if something was created by AI?

- Not confident at all
- A little confident
- Somewhat confident
- Very confident
- Extremely confident

AI detection importance. How important is it for you to be able to tell if digital media (e.g. social media posts) were made by AI or made by humans?

- Not at all important
- Slightly important
- Somewhat important
- Very important
- Extremely important

AI sentiment. Overall, would you say the increased use of AI in daily life makes you feel...

- More concerned than excited
- Equally concerned and excited
- More excited than concerned

AI awareness. When you are consuming content on social media, how often do you think about whether it might have been created by an AI?

- Never
- Sometimes
- About half the time
- Most of the time
- Always

Human-authorship preference. How much do you prefer that content on social media was created by a human?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

Digital literacy. How familiar, if at all, are you with the following computer and Internet-related items or actions? *[Matrix; each row rated on the scale below.]*

- Recommendation algorithm
- Clickbait
- Phishing
- Shadow banning
- Content moderation

Scale: Not familiar at all; Slightly familiar; Moderately familiar; Very familiar; Extremely familiar

Attention check. Paying attention and reading the instructions carefully is critical. President Obama is the youngest politician listed in the options below. If you are paying attention, please choose the youngest politician from the list.

- President Joe Biden
- President Trump
- President Obama
- Nancy Pelosi
- Mike Pence
- Don't know

[Participants then saw: "Now, we want to know more about your general social media consumption habits and political perspectives."]

Political attention. How often do you keep up with politics, regardless of the source – whether through TV, radio, newspapers, or online platforms?

- Extremely closely
- Very closely
- Moderately closely
- Slightly closely

- Not at all

News consumption. How often do you use each of the following as a source of news? (By news, we mean national, international, regional/local news, and other topical events.) *[Matrix; each row rated on the scale below.]*

- Television news (including streaming)
- Radio or audio news (including podcasts)
- Printed newspapers
- Websites/apps of newspapers and news outlets
- From talking with friends, family and co-workers
- X (Formerly Twitter)
- Tiktok
- Youtube
- Truth Social
- AI chatbots

Scale: Never; Sometimes; About half the time; Most of the time; Always

Social media usage. How often do you use each of the following social media platforms? *[Matrix; each row rated on the scale below.]*

- YouTube
- TikTok
- X (Formerly Twitter)
- Truth Social
- Instagram
- Facebook
- Reddit

Scale: Not at all; A little; Some; Quite a lot; All the time

Party identification. Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent or something else?

- Republican
- Democrat
- Independent
- Something else

Party lean. *[Shown only if Party identification = Independent or Something else.]* Do you think of yourself as closer to the Republican Party or to the Democratic Party?

- Closer to the Republican Party
- Closer to the Democratic Party
- Neither

Ideological self-placement. When it comes to politics, would you describe yourself as liberal, conservative, or neither liberal nor conservative?

- Very liberal
- Somewhat liberal
- Slightly liberal

- Moderate, middle of the road
- Slightly conservative
- Somewhat conservative
- Very conservative

Task instructions. Social Media Task. For the next question, you will be asked to compare two social media posts and label if you believe they were produced by an AI (artificial intelligence) chatbot or by a human. An AI chatbot is a computer program that can talk and message with people to answer questions, provide recommendations or help complete tasks. Examples of AI chatbots include ChatGPT, Claude, Microsoft Copilot, and Google's Gemini. You will be shown a pair of social media posts. For the pair, you must pick the post that seems more likely to be produced by an AI chatbot. Even if you are not certain, we ask you to decide which of the two posts from each pair is more likely to have been produced by an AI chatbot.
Example pairwise item. *[A single fixed example pair was shown.]*

- *Prompt:* Please, read the social media posts below carefully.
- *Post A:* trump accidentally said company instead of country [two laughing emoji] #Debate2024
- *Post B:* Watching Harris tonight makes me sick. She has NO character and NO fitness to lead. TRUMP is the only one with the guts to defend our values! [flexed-biceps emoji] #Debatenight
- *Question:* Which of the posts (Post A or Post B) is more likely to have been generated by an AI chatbot?
- *Response options:* Post A; Post B

S12. References for Supplement.

References

1. R. A. Bradley and M. E. Terry. *Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika*, 1952.
2. Barbieri, Francesco, Ballesteros, Miguel, and Camacho-Collados, Jose. *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
3. Davis, Clayton A., Varol, Onur, Ferrara, Emilio, Flammini, Alessandro, and Menczer, Filippo. *BotOrNot: A System to Evaluate Social Bots. Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 273–274, 2016.
4. OpenAI. GPT-4. *OpenAI API Model Documentation*, 2023. Available at: https://platform.openai.com/docs/models/gpt-4 (accessed October 20, 2025).
5. OpenAI. GPT-4o. *OpenAI API Model Documentation*, 2024. Available at: https://platform.openai.com/docs/models/gpt-4o (accessed October 20, 2025).
6. OpenAI. GPT-4o Mini. *OpenAI API Model Documentation*, 2024. Available at: https://platform.openai.com/docs/models/gpt-4o-mini (accessed October 20, 2025).
7. Meta AI. Llama 2 70B. *Model Card on Hugging Face*, 2023. Available at: https://huggingface.co/meta-llama/Llama-2-70b (accessed October 20, 2025).
8. Meta AI. Llama 3.2 3B. *Model Card on Hugging Face*, 2024. Available at: https://huggingface.co/meta-llama/Llama-3.2-3B (accessed October 20, 2025).
9. Meta AI. Llama 3.3 70B Instruct. *Model Card on Hugging Face*, 2024. Available at: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct (accessed October 20, 2025).
10. J. Camacho-Collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa-Anke, F. Liu, E. Martínez-Cámara, et al. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, U.A.E. Association for Computational Linguistics, 2022.
11. D. Q. Nguyen, T. Vu, and A. T. Nguyen. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, 2020.
12. F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, 2020.
13. Bucket Research. politicalBiasBERT: Political bias classification using finetuned BERT model, 2023. https://huggingface.co/bucketresearch/politicalBiasBERT. DOI: 10.57967/hf/0870.
14. S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*, 2010.

15. H. Dawkins, K. C. Fraser, and S. Kiritchenko. When detection fails: The power of fine-tuned models to generate human-like social media text. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13494–13527, Vienna, Austria, 2025.
16. A. Rauchfleisch and J. Kaiser. The false positive problem of automatic bot detection in social science research. *PLOS ONE*, 15(10):1–20, 2020.
17. Z. Chen, J. Ye, E. Ferrara, and L. Luceri. Prevalence, sharing patterns, and spreaders of multimodal AI-generated content on X during the 2024 U.S. Presidential Election. *arXiv preprint arXiv:2502.11248*, 2025.
18. American National Election Studies. ANES 2022 Pilot Study [Data set]. University of Michigan and Stanford University, 2022. https://electionstudies.org/data-center/2022-pilot-study/.
19. P. J. Conover and S. Feldman. The origins and meaning of liberal/conservative self-identifications. *American Journal of Political Science*, 25(4):617–645, 1981.
20. K. Hackenburg and H. Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121, 2024.
21. M. Levendusky. *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. University of Chicago Press, 2009.
22. AllSides. Media bias rating methods, 2026. https://www.allsides.com/about/media-bias-rating-methods.