

AI summaries in social media improve dialogue but reduce engagement

Michael Heseltine^{*1}, Christopher Bail³, Petter Törnberg⁴, Michelle Schimmel⁴, and Christopher Barrie^{2,1}

¹Department of Sociology, University of Oxford, Oxford, UK

²Department of Sociology, New York University, NY 10012, USA

³Department of Sociology, Duke University, Durham, NC 27516, USA

⁴Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Netherlands

March 26, 2026

Abstract

Generative artificial intelligence agents are becoming increasingly active participants in conversations on social media platforms, yet little is known about how they shape public discussion of social problems. We present two preregistered online experiments testing AI-generated summaries in simulated, interactive social media environments. AI summaries increased the quality of user comments, without systematically increasing toxicity or negative affect. At the same time, AI exposure reduced engagement with conversation threads. AI summaries also increased the semantic similarity between user comments and the AI-generated summaries, suggesting that these systems function as informational anchors that shape discussion. Together, the findings reveal a tradeoff: AI-generated summaries can improve conversation quality while narrowing conversational engagement and channeling how users articulate political arguments. These results speak to growing concerns about how embedded AI systems fundamentally alter platform dynamics and shape public discourse.

Preprint draft — please do not redistribute without permission

Keywords: generative AI | social media | political deliberation | engagement | framing

Generative artificial intelligence (AI) systems are rapidly being integrated into digital information environments. Large language models now generate summaries, contextual explanations, and conversational responses directly within social media platforms and search interfaces [1, 2]. For example, on X (formerly Twitter), the Grok assistant provides real-time AI commentary embedded within discussion threads [3, 4]. Similar systems are being deployed across Meta platforms, YouTube, and search engines, where AI-generated explanations appear alongside user-generated and journalistic content. These tools promise to help users navigate complex topics by producing concise, authoritative summaries.

This development has generated controversy. Disputes surrounding systems such as Grok have raised concerns about the risks of embedding AI agents within the infrastructures that shape political debate [3]. Emerging research shows that AI-generated explanations can influence beliefs and political judgments [5–7], chatbot interventions can improve the quality of online political discussions [8], chatbot advice can shape evaluations [9], and perceptions of AI systems can affect political legitimacy and engagement [10, 11]. Generative systems may also reproduce biases embedded in training data or reflect platform design choices [2]. Despite these concerns, we lack controlled experimental evidence on how AI-generated sum-

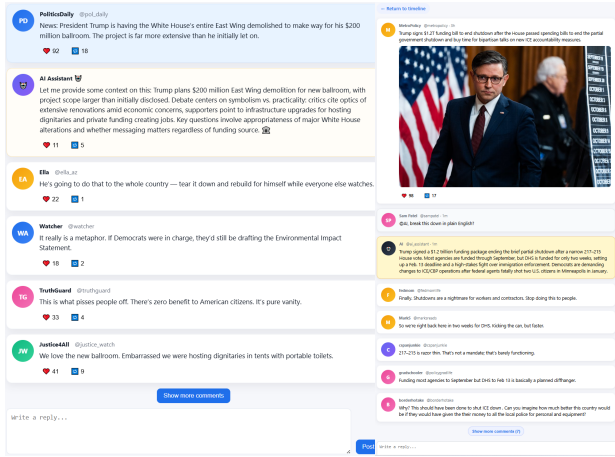
maries embedded within social media conversations shape the dynamics of political discussion.

We examine how AI-generated summaries influence political engagement and discussion quality within social media environments, drawing on longstanding theories of democratic competence and discussion [12–14]. Specifically, we test whether embedding AI summaries affects (i) engagement with conversation threads, and (ii) the quality of user contributions, and (iii) the framing of arguments users articulate.

To answer these questions, we conduct two preregistered online experiments using simulated social media interfaces.

Study 1 employs a controlled single-thread design within an online survey experiment. Participants view an interactive political conversation embedded in the survey styled to resemble a typical social media thread. Respondents are randomly assigned to one of four conditions: no AI summary, a neutral AI summary, a left-leaning AI summary, or a right-leaning AI summary. The summary appears directly beneath the main post and is labeled as AI-generated. Participants can scroll through the comments, expand replies, and post their own response. Behavioral telemetry records interaction patterns, time spent, and comment text. The summary is labeled simply as an “AI Summary,” without identifying the specific system.

Study 2 extends this design to a scrolling feed that more



(a) Study 1 interface

(b) Study 2 interface

Figure 1: Screenshots of the experimental platform used in Studies 1 and 2.

closely resembles naturalistic social media use. Participants encounter multiple political posts and choose which discussions to open and engage with. In this study, there is no randomization into content with differing political lean. AI summaries appear within conversation views but not on the main timeline, mirroring current platform implementations.

Across both studies, we observe a consistent pattern. Exposure to AI-generated summaries increases the quality of user comments: Participants write longer, more constructive, and more reasoned responses without systematic increases in toxicity, disrespect, or negative sentiment (including when exposed to summaries with outgroup lean). At the same time, AI summaries reduce engagement with conversation threads. Participants exposed to summaries are less likely to click to read additional comments in a thread and, in the feed environment, post fewer total comments. Finally, comments in the AI condition exhibit greater semantic similarity to the AI-generated summaries, suggesting that users partially adopt the language and framing introduced by the system, consistent with classic accounts of framing effects [15].

Together, these findings reveal a central tradeoff: AI-generated summaries can improve the structural quality of debate while simultaneously narrowing conversational engagement and channeling how users articulate political arguments. These results speak to growing concerns about how embedded AI systems fundamentally alter platform dynamics and shape public discourse.

Results

AI Summaries Increase Comment Quality

We first test whether exposure to AI-generated summaries improves the overall quality of user comments. Across both experiments, AI exposure increased multiple text-based quality indicators.

In Study 1 (single-thread environment), participants exposed to any AI summary wrote longer and more substantively engaged comments than those in the control condition. Across measures, including word count, constructiveness, reasoning, nuance, and readability-adjusted complexity, AI coefficients were uniformly positive. Most effects reached conventional levels of statistical significance ($P < 0.05$), and all pointed in the same direction. We also find no systematic differences between neutral, left-leaning, and right-leaning AI summaries across outcome measures (SI Appendix).

Study 2 (multi-post feed environment) replicates this overall pattern. Participants in the AI condition produced longer comments and scored higher on reasoning, and nuance measures. Effect sizes were comparable in magnitude to Study 1, highlighting that general increases in comment quality were robust in a more realistic feed-based environment.

AI Summaries do not Increase Negativity or Toxicity

Across both studies, we find little evidence that AI summaries systematically altered the tone or sentiment of comments. In Study 1, coefficients across multiple sentiment and toxicity measures were small and statistically indistinguishable from zero. Study 2 mirrors these findings, with no consistent effects on sentiment polarity, toxicity, respect, or polarizing language across NLP classifiers and API-based measures. Additionally, we find no evidence of changes in affective opinion of the party from AI exposure, with no significant changes to opinions about the parties or key political actors found in either study (S.M. 6).

AI Summaries Reduce Exploratory Engagement

In Study 1, exposure to AI summaries reduced exploratory interaction with the conversation: participants were less likely to click “show more comments” ($P < 0.05$). In Study 2, AI exposure reduced total comment production, and participant scroll depth (how far down the conversation thread they scrolled) was significantly reduced. Across both environments, AI summaries, in some way, decreased conversational engagement, consistent with substitution effects.

AI Summaries Anchor Semantic Framing

To further investigate how user comment behavior changes, Study 2 was designed with clear semantic frames of discussion in the AI summary presented for each post. We then compute the cosine similarity between each participant’s comment and the AI-generated summary associated with the corresponding post using sentence-level embeddings. AI exposure increased semantic similarity ($\beta = 0.018$, $P = 0.029$). Although modest in absolute terms, this represents roughly a 7–8% increase relative to the baseline mean similarity (0.25), indicating a consistent shift toward closer alignment with the AI-generated framing. Additionally, participants exposed to AI summaries

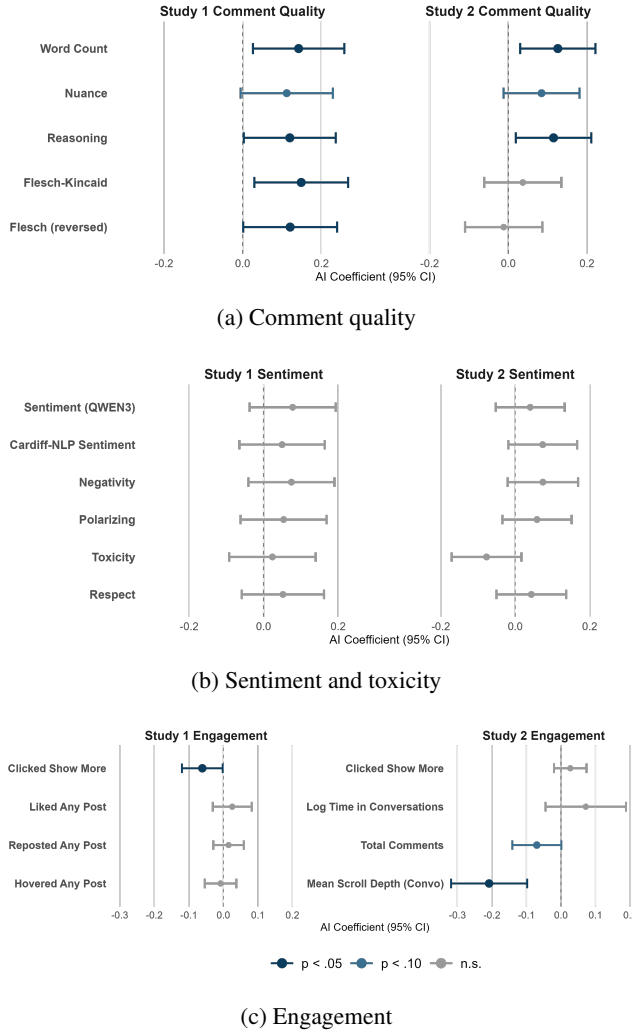


Figure 2: Estimated AI treatment effects (95% confidence intervals) across outcome families in Studies 1 and 2.

were significantly more likely to adopt distinctive terms appearing in the summary itself ($\beta = 0.047$, $P = 0.029$), providing direct evidence of semantic anchoring beyond general similarity measures. This indicates that AI summaries can function as informational anchors that shape how users frame their contributions.

Discussion

Across two preregistered experiments, we find a consistent tradeoff in the effects of AI-generated summaries on political discussion. AI summaries increase the structural quality of user comments, making them longer, more reasoned, and more constructive, without increasing toxicity. At the same time, they reduce exploration of the broader conversation and increase semantic alignment between user comments and the AI-generated framing.

These results suggest that AI summaries function both as deliberative scaffolds and informational anchors. By lowering

cognitive costs, they encourage more structured participation. By foregrounding particular interpretations, they also subtly shape how users frame their contributions. Improvements in comment quality therefore coexist with reduced conversational exploration and increased anchoring to AI-provided frames.

As generative AI systems become embedded within social media feeds, search engines, and news environments, they are likely to influence not only how much users engage, but how public issues are interpreted. Because AI summaries are presented as neutral and authoritative, they may shift agenda-setting and framing power to the systems that generate them. The central challenge is therefore not simply whether AI improves or degrades discourse, but how to design systems that enhance deliberation without narrowing the diversity of frames through which democratic debate unfolds.

Several limitations should be noted. Although our experimental interfaces approximate social media environments, they cannot fully capture real-world platform dynamics. Our measures focus on short-term changes in comment content and engagement, and do not assess longer-term effects on beliefs or participation. Results may also vary across topics, platforms, and populations not studied here. Future work should extend these findings in more naturalistic settings and over longer time horizons.

Materials and Methods

Participants and preregistration

Both studies were preregistered prior to data collection. Study protocols were approved by the Institutional Review Board at the University of Oxford (Protocol #2437050). All participants provided informed consent. Participants were recruited from Prolific Academic and restricted to adult residents of the United States using desktop or tablet. Full sample demographics are found in SM3. In Study 1, a total of $N = [1583]$ participants were recruited; in Study 2, $N = [1281]$ participants were recruited. Participants who failed attention checks, did not complete the study, or submitted empty required comment responses were excluded prior to analysis. Final analytic samples were $N = [1548]$ (Study 1) and $N = [1266]$ (Study 2).

Study 1 design

Participants were randomly assigned to view a simulated social media conversation under one of four conditions: (1) no AI summary (control), (2) neutral AI summary, (3) left-leaning AI summary, or (4) right-leaning AI summary. The underlying post and replies were identical across conditions. When present, the AI summary, generated by Claude Sonnet 4.5, appeared immediately below the main post and was clearly labeled as AI-generated (without specifying the precise model name). Participants could like, repost, expand additional comments, and post a required free-text response. The interface recorded behavioral telemetry including likes and repost clicks,

comment expansion, scroll depth, time spent on the page, and the full text of participants' comments.

Study 2 design

Study 2 embedded the manipulation within a multi-post, scrollable social media feed to enhance external validity. Participants encountered several political posts and could choose which conversation(s) to open and engage with. AI summaries, generated by GPT 5.2, appeared only within the conversation view (not on the main timeline) in the treatment condition. Multiple comments were permitted on one or more posts. Interactions, comment expansion, scroll depth and time spent in each conversation view was recorded.

Measures

Across both studies, we collected behavioral engagement measures (e.g., comment expansion, interaction actions, time spent) and required participants to post at least one free-text comment. Comment quality was assessed using established readability indices (e.g., Flesch, Flesch–Kincaid) and reasoning and nuance from the Perspective API. Sentiment was measured using pretrained BERT-based classifiers of sentiment, negativity, and polarizing language, and Perspective API measures of toxicity and respect; Study 2 additionally measured semantic similarity between participant comments and the AI-generated summary. Continuous outcomes were standardized prior to analysis.

Statistical analysis

We estimate average treatment effects using linear regression models with topic/post fixed effects and pre-specified demographic and political covariates for precision. Binary outcomes are estimated using linear probability models with robust standard errors. Study 2 models cluster standard errors at the participant level to account for repeated observations.

Acknowledgments

Funding: Funding: ESRC ESZ0005481, NSF SMA-2436762.

Author contributions

M.H., C.Barrie, C.Bail, P.T., and M.S. designed research; M.H. performed research; M.H. analyzed data; M.H., C.Barrie, C.Bail, P.T., and M.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Data availability

Data, materials, and analysis code will be made publicly available at <https://osf.io/gd793/files/e76tk> upon publication.

References

- [1] Jean Kaddour et al. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023. URL <http://arxiv.org/abs/2307.10169>.
- [2] H. Naveed et al. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2024. URL <http://arxiv.org/abs/2307.06435>.
- [3] Katelyn Xiaoying Mei, Robert Wolfe, Nicholas Weber, and Martin Saveski. Grok in the wild: Characterizing the roles and uses of large language models on social media, 2026. URL <https://arxiv.org/abs/2602.11286>.
- [4] Thomas Renault, Mohsen Mosleh, and David G. Rand. @Grok Is This True? LLM-Powered Fact-Checking on Social Media. https://osf.io/preprints/psycharxiv/85quw_v2, January 2026. PsyArXiv preprint.
- [5] Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024.
- [6] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024. doi: 10.1126/science.adq2852. URL <https://www.science.org/doi/abs/10.1126/science.adq2852>.
- [7] Lisa P. Argyle, Ethan C. Busby, Joshua R. Gubler, Alex Lyman, Justin Olcott, Jackson Pond, and David Wingate. Testing theories of political persuasion using ai. *Proceedings of the National Academy of Sciences*, 122(18):e2412815122, 2025.
- [8] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120, 2023. doi: 10.1073/pnas.2311627120.
- [9] Yamil R. Velez, Donald P. Green, and Semra Sevi. Chatbot voting advice applications inform but seldom sway young unaligned voters. *Proceedings of the National Academy of Sciences*, 122(50):e2515516122, 2025. doi: 10.1073/pnas.2515516122.
- [10] Armin Granulo, Andreas Raff, and Christoph Fuchs. Perceiving ai as labor-replacing reduces democratic legitimacy and political engagement. *Proceedings of the National Academy of Sciences*, 123(4):e2523508123, 2026. doi: 10.1073/pnas.2523508123.
- [11] C. Y. Cyrus Chu, Juin-Jen Chang, and Chang-Ching Lin. Why does ai hinder democratization? *Proceedings of the National Academy of Sciences*, 122(19):e2423266122, 2025. doi: 10.1073/pnas.2423266122.
- [12] Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, Cambridge, MA, 1996.
- [13] Philip E. Converse. The nature of belief systems in mass publics. *Ideology and Discontent*, pages 206–261, 1964.
- [14] Diana C. Mutz. *Hearing the Other Side: Deliberative versus Participatory Democracy*. Cambridge University Press, Cambridge, 2006.
- [15] Dennis Chong and James N. Druckman. Framing theory. *Annual Review of Political Science*, 10(1):103–126, 2007. doi: 10.1146/annurev.polisci.10.072805.103054.

Supplementary Materials for: AI summaries in social media improve dialogue but reduce engagement Full Materials w/ Analysis

Michael Heseltine Christopher Bail Petter Törnberg
Michelle Schimmel Christopher Barrie

March 26, 2026

1 Preregistration and Deviations

Both studies were preregistered prior to data collection at https://osf.io/3xfp2/overview?view_only=19d2374fe968458abbb7ec4cc7c84aea and https://osf.io/gd793/overview?view_only=3b1bdc961b2148279421c8ccc8afe81d. The preregistration specified primary hypotheses concerning:

1. The effect of any AI summary on deliberative quality,
2. The effect of AI summaries on comment sentiment,
3. The effect of partisan alignment on sentiment and evaluation.

Deviations

Multiple-comparison adjustments (Study 1) The Study 1 preregistration indicated that Bonferroni corrections would be applied to hypothesis tests. Upon further consideration, we determined that this adjustment was inappropriate for the primary analyses reported here. The confirmatory models estimate a single treatment effect (AI vs. control) on pre-specified outcome indices, rather than conducting families of independent hypothesis tests across distinct theoretical constructs. Applying Bonferroni corrections in this context would therefore constitute over-adjustment, substantially inflating Type II error without corresponding control of a meaningful familywise error rate. Consistent with current methodological guidance for experimental designs with clearly defined primary outcomes, we report unadjusted p -values for the preregistered treatment effects. This decision does not alter model specification, outcome construction, or the direction of results; it concerns only the correction procedure.

Semantic similarity hypothesis (Study 2) The Study 2 preregistration explicitly stated that semantic similarity between participant comments and AI summaries would be measured and analyzed as an outcome. However, the final preregistration document did not include a separately enumerated hypothesis statement corresponding to this outcome. This omission was inadvertent. The analysis of semantic similarity follows directly from the preregistered outcome description and is consistent with the study’s stated objective of assessing whether AI summaries shape the substantive content of participant responses. No post hoc outcome measures were introduced; rather, we test the preregistered semantic similarity outcome as originally described.

2 Sample Recruitment and Exclusions

Participants were recruited via Prolific Academic between November 25 2025 and November 30 2025 for Study 1 and February 8 2026 and February 11 2026 for Study 2. Eligibility criteria included:

- Age 18 or older,
- U.S. residency,
- Participants from Study 1 excluded from taking Study 2

Study 1

A total of 1583 participants completed the study (those who started but did not finish were not recorded in the data). Of these respondents, 1548 successfully passed the attention check question.

Study 2

A total of 1281 participants completed the study (those who started but did not finish were not recorded in the data). Of these respondents, 1266 successfully passed the attention check question.

Attention Check

In Study 1, participants were asked: “Please select ‘Somewhat disagree’ for this question.” In Study 2, participants were asked: “Please select ‘Somewhat agree’ for this question.” Participants who failed were excluded prior to analysis.

3 Randomization Balance

To assess successful random assignment, we test for differences in pre-treatment covariates across conditions.

Covariates include:

- Age,
- Gender,
- Education,
- Party identification,
- Ideological extremity,
- Political interest.

Across both studies, we observe no statistically significant imbalances after correction for multiple comparisons (all $p > 0.05$). Full balance tables are reported in Table S1.

Table 1: Randomization balance (Study 1). Entries are mean (SD) for continuous covariates and count (%) for categorical covariates. P-values from bivariate tests of AI assignment.

Covariate	Group		p-value
	Control (0)	AI (1)	
Age	45.76 (15.45)	45.36 (15.60)	0.656
Ideological extremity	1.58 (1.06)	1.59 (1.08)	0.842
Political interest	3.63 (0.96)	3.72 (0.94)	0.079
<i>Gender</i>			0.904
Gender: Female	214 (51.0%)	576 (49.9%)	
Gender: Male	203 (48.3%)	568 (49.2%)	
Gender: Other/Prefer not to say	3 (0.7%)	10 (0.9%)	
<i>Education</i>			0.479
Education: Associate degree in college (2-year)	42 (10.0%)	129 (11.2%)	
Education: Bachelor's degree in college (4-year)	157 (37.4%)	438 (38.0%)	
Education: Doctoral degree	7 (1.7%)	27 (2.3%)	
Education: High school graduate (high school diploma or equivalent including GED)	51 (12.1%)	138 (12.0%)	
Education: Less than high school degree	2 (0.5%)	14 (1.2%)	
Education: Master's degree	59 (14.0%)	156 (13.5%)	
Education: Professional degree (JD, MD)	4 (1.0%)	23 (2.0%)	
Education: Some college but no degree	98 (23.3%)	229 (19.8%)	
<i>Party identification</i>			0.578
Party identification: Democrat	167 (39.8%)	513 (44.5%)	
Party identification: Independent	128 (30.5%)	323 (28.0%)	
Party identification: No preference	11 (2.6%)	25 (2.2%)	
Party identification: Other	6 (1.4%)	16 (1.4%)	
Party identification: Republican	108 (25.7%)	277 (24.0%)	

Table 2: Randomization balance (Study 2). Entries are mean (SD) for continuous covariates and count (%) for categorical covariates. P-values from bivariate tests of AI assignment.

Covariate	Group		p-value
	Control (0)	AI (1)	
Age	44.51 (15.63)	44.17 (15.82)	0.703
Ideological extremity	1.65 (1.05)	1.54 (1.05)	0.082
Political interest	3.68 (0.91)	3.66 (0.94)	0.776
<i>Gender</i>			0.959
Gender: Female	310 (49.3%)	317 (48.6%)	
Gender: Male	312 (49.6%)	327 (50.2%)	
Gender: Other/Prefer not to say	7 (1.1%)	8 (1.2%)	
<i>Education</i>			0.357
Education: Associate degree in college (2-year)	66 (10.5%)	74 (11.3%)	
Education: Bachelor’s degree in college (4-year)	251 (39.9%)	247 (37.9%)	
Education: Doctoral degree	18 (2.9%)	11 (1.7%)	
Education: High school graduate (high school diploma or equivalent including GED)	78 (12.4%)	71 (10.9%)	
Education: Less than high school degree	7 (1.1%)	4 (0.6%)	
Education: Master’s degree	81 (12.9%)	90 (13.8%)	
Education: Professional degree (JD, MD)	16 (2.5%)	12 (1.8%)	
Education: Some college but no degree	112 (17.8%)	143 (21.9%)	
<i>Party identification</i>			0.168
Party identification: Democrat	282 (44.8%)	282 (43.3%)	
Party identification: Independent	156 (24.8%)	192 (29.4%)	
Party identification: No preference	19 (3.0%)	15 (2.3%)	
Party identification: Other	16 (2.5%)	8 (1.2%)	
Party identification: Republican	156 (24.8%)	155 (23.8%)	

4 Experimental Stimuli

All conversation threads were adapted from real contemporary news posts and discussions on the X platform. Usernames were anonymized and standardized.

AI summaries were generated using **Claude Sonnet 4.5** for Study 1 and **ChatGPT 5.2** for Study 2 with the following prompt structure:

In Study 1, the following instructions were used. Some minor post-processing was then induced to standardize phrasing and lengths across treatments.

“Can you provide a tweet length summary of the following social media conversation in the style of a social media embedded AI like Grok. Focus primarily on the content of the first post, with the remaining comments as context. Do so from a ideologically neutral/liberal/conservative standpoint.”

In Study 2, the following instructions were used, and the text of an entire contemporary news article was also appended to the prompt for additional context.

“Can you provide a tweet length summary of the following social media conversation in the style of a social media embedded AI like Grok. For your context, the full text of the article is as follows: [article text].”

Final summaries were limited to approximately 280 characters.

We opted to switch from Claude to GPT for study 2 for purely pragmatic reasons, finding that GPT adhered better to the prompt, especially in the within-conversation AI call context that was utilized for Study 2, providing additional context from the article rather than merely repeating the content from the original post.

5 Text Classification and Measurement Details

Sentiment and Toxicity

We employ multiple independent classification systems:

- CardiffNLP Twitter-RoBERTa sentiment classifier (Barbieri et al., 2020),
- Qwen3 4B zero-shot sentiment classifier,
- Perspective API toxicity and respect metrics (Lees et al., 2022),
- Fine-tuned binary negativity and polarization classifiers (Ballard et al., 2023; Heseltine, 2024)

All classifiers were applied to raw comment text without manual editing. Continuous outputs were standardized prior to regression.

For the Qwen3 classifier, we classified raw comment text using the Qwen3 4b q8 model through Rollama, an R wrapper for Ollama. We used the following zero-shot prompt, based on existing LLM sentiment classification prompts from (Heseltine, 2025; Heseltine and von Hohenberg, 2024):

“classify the following social media comment as as negative (0), neutral (1), or positive (2). Negative is defined as a message which expresses negative sentiment directly towards an individual, group, or outcome and is not merely any message which contains negative valence. Positive is defined as a message which expresses positive sentiment directly towards an individual, group, or outcome and is not merely a message which mentions something positive. Do not add an introduction, qualification or any additional text. The output will be machine processed so return only the predicted class as a number.”

Deliberative Quality

Reasoning and nuance were measured using Perspective API experimental features.

Readability was calculated using Flesch–Kincaid grade level and reversed Flesch reading ease.

Word count was computed as the total number of tokenized words per comment.

For full transparency, we also attempted to create an LLM-based measure of comment constructiveness, similar in process to the Qwen3-based sentiment classification outlined above. However, across three different models used for testing, accuracy was deemed to be low and the correlation between model outputs was not sufficiently high enough to be deemed a reliable measure.

Semantic Similarity

To assess whether participants adopted the framing introduced by AI-generated summaries, we compute the semantic similarity between each participant comment and the corresponding AI summary shown for that post.

Sentence embeddings were generated using the `all-MiniLM-L6-v2` model from the `SentenceTransformers` library (Reimers and Gurevych, 2019). This model produces 384-dimensional sentence-level embeddings optimized for semantic similarity tasks. All embeddings were L2-normalized prior to similarity computation.

For each post, the AI-generated summary was embedded once. Participant comments were embedded individually. Cosine similarity was then calculated between each comment embedding and the embedding of the AI summary associated with the same post. Because embeddings were normalized, cosine similarity is equivalent to the dot product between vectors.

In Study 2, where participants could comment on multiple posts, similarity was computed at the comment level and then aggregated to the participant–post level by averaging across comments. Regression models estimate the effect of AI exposure on mean similarity, controlling for post fixed effects and clustering standard errors at the participant level to account for repeated observations.

To provide a more interpretable test of semantic anchoring, we additionally examine whether participants explicitly reused distinctive lexical terms from the AI-generated summaries in their own comments.

For each post, we constructed a post-specific anchor dictionary derived from the corresponding AI summary. Summary text was tokenized using the `tidytext` package, converted to lowercase, and stripped of standard English stopwords. Tokens shorter than four characters were excluded to remove function words and generic terms. For each post, we selected the 15 most frequent remaining terms in the AI summary to define a post-level anchor dictionary.

We then created a binary indicator for each participant comment equal to 1 if the comment contained at least one anchor term associated with the relevant post, and 0 otherwise. Matching was performed using case-insensitive word-boundary regular expressions to ensure that only exact lexical matches were counted.

We estimate linear probability models of the form:

$$Anchor_{ip} = \beta_0 + \beta_1 AI_i + \gamma_p + \epsilon_{ip},$$

where $Anchor_{ip}$ indicates whether participant i used at least one anchor term in a comment on post p , AI_i indicates exposure to an AI-generated summary, and γ_p are post fixed effects. Standard errors are clustered at the participant level.

This keyword-based approach complements the embedding-based similarity measure by directly capturing the adoption of salient lexical framing introduced by the AI system.

6 Treatment Effects on Respondent Opinions

As a secondary outcome, we also tested whether exposure to AI summaries affected participants’ political attitudes. After interacting with the conversation content in the main study, participants were asked the following questions.

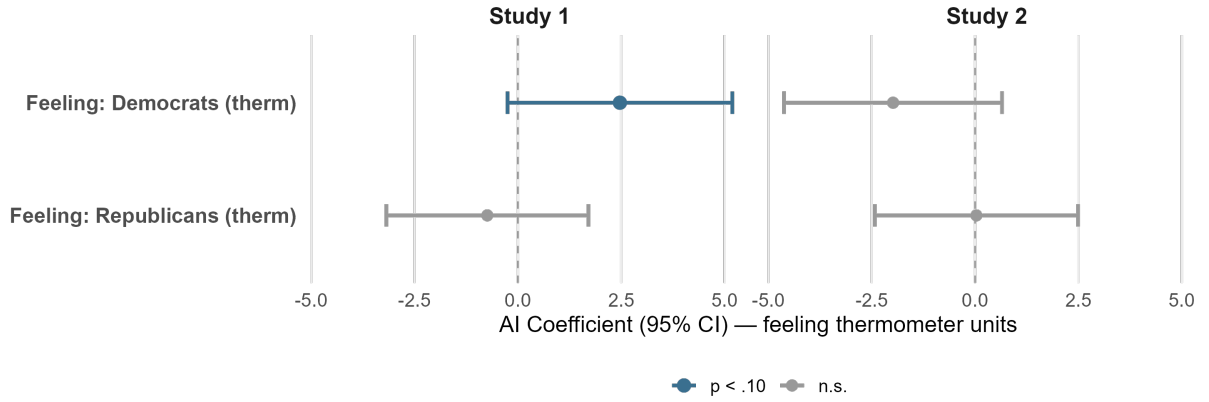
Participants were asked to report their opinions using the following measures in both studies:

- **Feeling thermometers:** “Please rate your feelings toward the Democratic and Republican political parties on a scale from 0 to 100, where 0 means very cold or unfavorable and 100 means very warm or favorable.”
 - The Democratic Party
 - The Republican Party
- **Political approval:** “On a scale of 1-5, 1 being “strongly disapprove” and 5 being “strongly approve”, please indicate the extent to which you approve of the job performance of the following groups and individuals.”
 - Democrats in Congress
 - Republicans in Congress
 - President Donald Trump
 - The Supreme Court

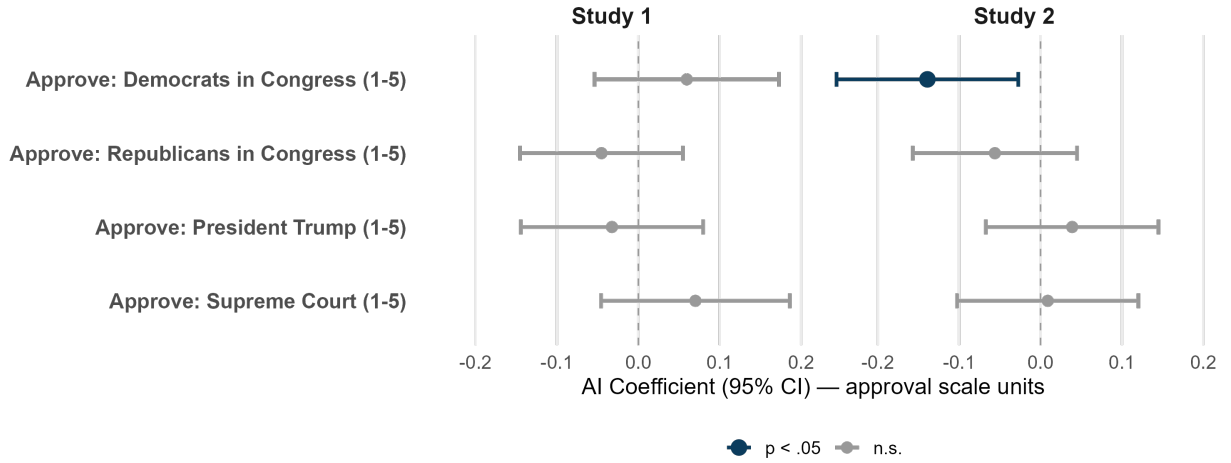
We then estimate the effect of AI summary exposure on each of these outcomes using linear regression models of the form:

$$Y_i = \beta_0 + \beta_1 AI_i + \gamma PartyID_i + \epsilon_i$$

where AI_i indicates exposure to an AI-generated summary and $PartyID_i$ controls for respondent party identification.



(a) Party Feeling



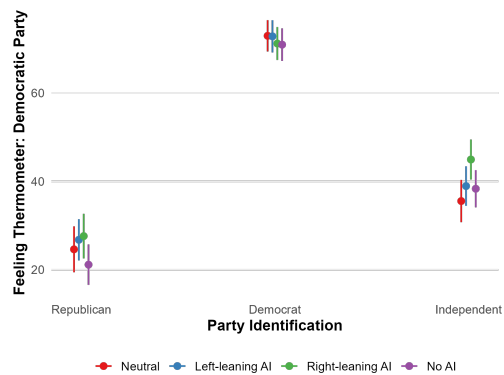
(b) Political Entity Approval

Figure 1: Estimated AI treatment effects (95% confidence intervals) on political opinions in Studies 1 and 2.

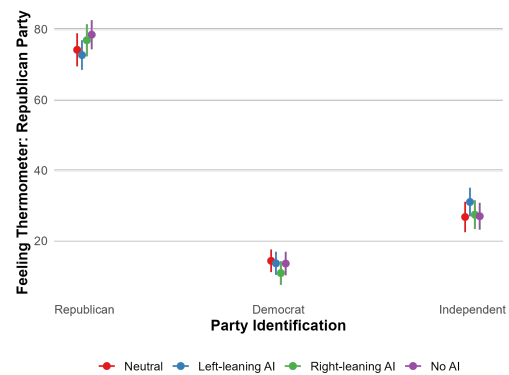
We find generally null results. In Study 2, approval of Democrats in Congress was significantly lower and in Study 1 feelings towards the Democratic Party were significantly higher. However, we have no real theoretical explanation for why these particular differences may occur.

6.1 Heterogeneous Opinion Treatment Effects

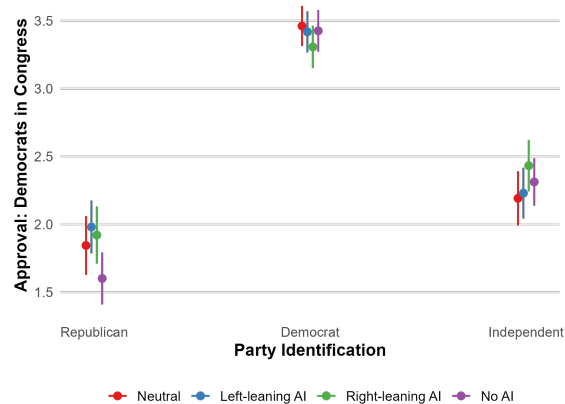
As Study 1 contained three different AI treatment arms with differing partisan leans, we break out those results here. As can be seen, there is little difference across partisan identification and treatment arm combinations. There is some suggestive evidence that the left-leaning AI reduced approval for copartisan figures among Republicans (President Trump and the Republican Party), but these differences remain small in size.



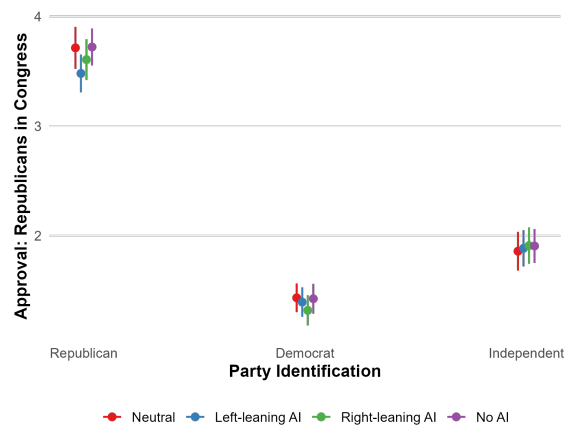
(a) Feeling thermometer: Democratic Party



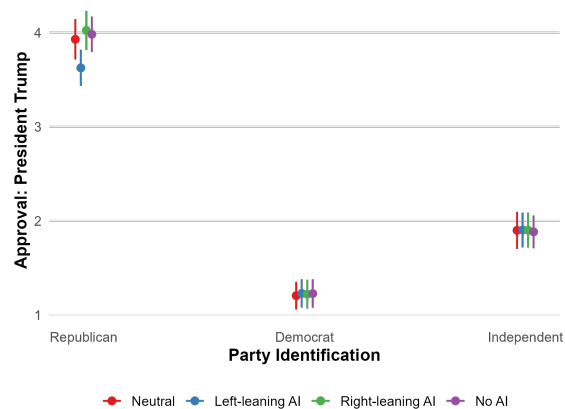
(b) Feeling thermometer: Republican Party



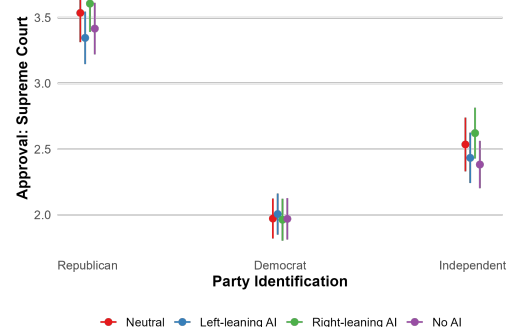
(c) Approval: Democrats in Congress



(d) Approval: Republicans in Congress



(e) Approval: President Trump

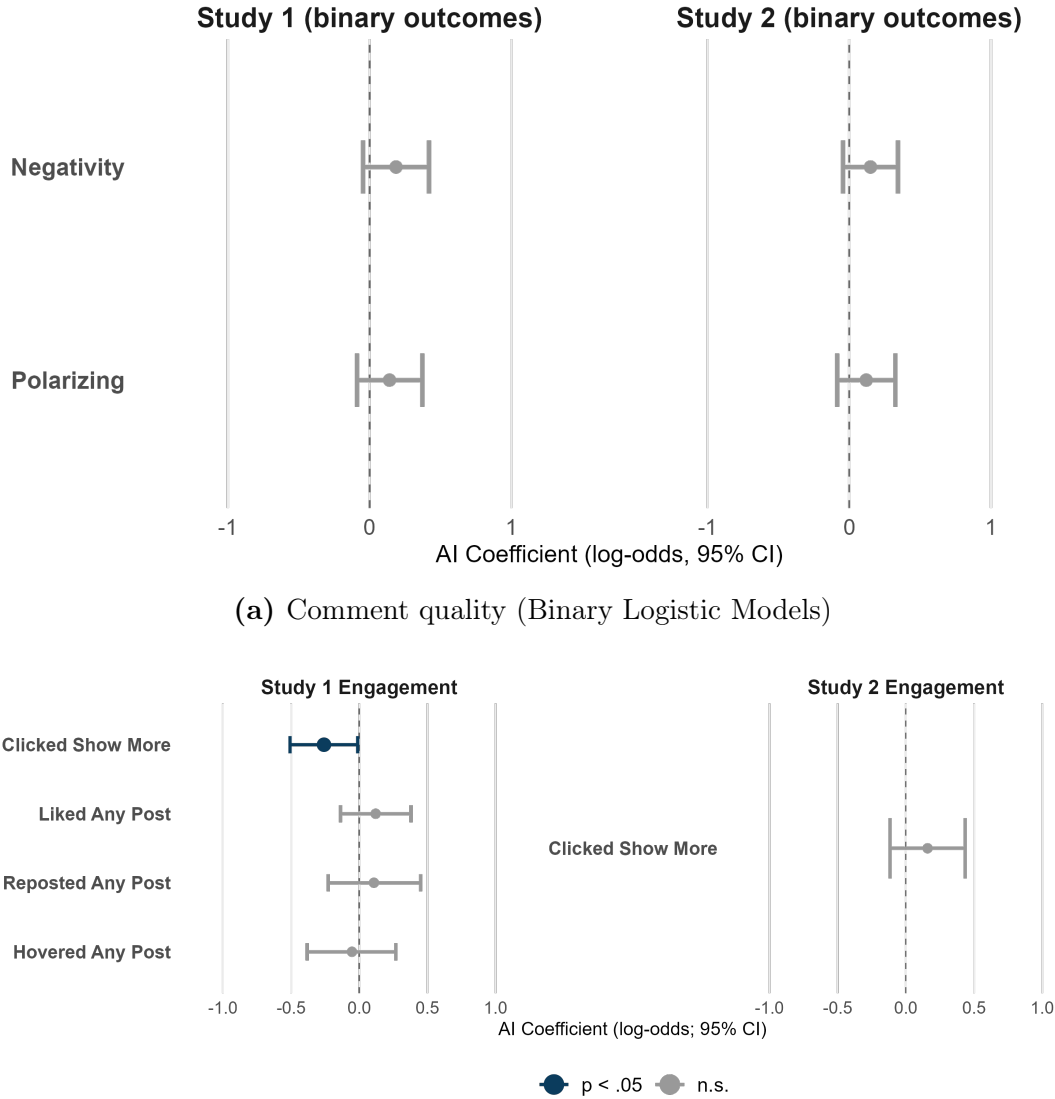


(f) Approval: Supreme Court

Figure 2: Predicted political evaluations by party identification and AI summary lean (Study 1). Panels show interaction model predictions with 95% confidence intervals for feeling thermometers toward the Democratic and Republican parties (top row) and approval of political actors and institutions (remaining rows).

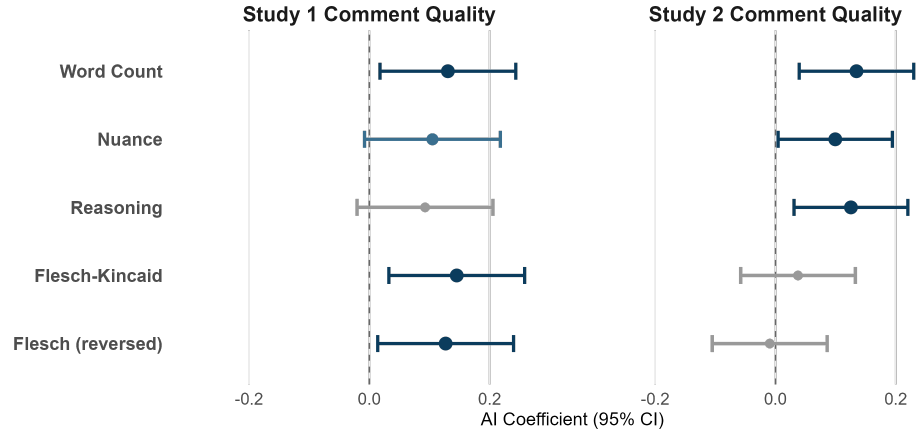
7 Robustness Checks

Results are robust to excluding demographic controls and using logistic regression instead of linear probability models. Across specifications, when excluding demographic and political controls, the direction and statistical significance of primary effects remain substantively unchanged, although a small number of specific coefficients shift over or below the threshold of significance. For the binary variables of interest, using logistic regression results in no substantive change in results.

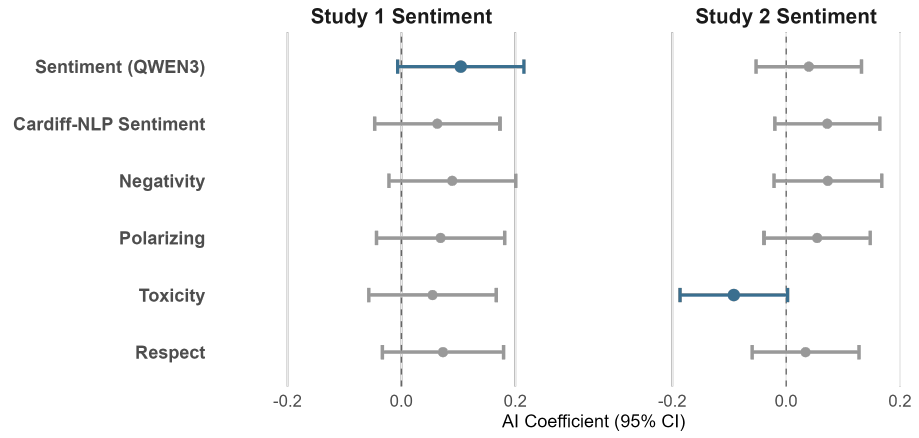


(b) Sentiment and toxicity (Binary Logistic Models)

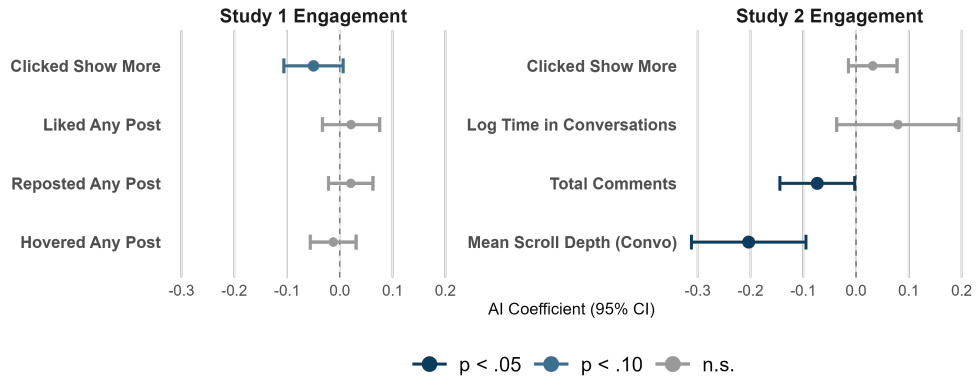
Figure 3: Estimated AI treatment effects (95% confidence intervals) across outcome families in Studies 1 and 2, with logistic regression for binary variables.



(a) Comment quality (No Controls)



(b) Sentiment and toxicity (No Controls)

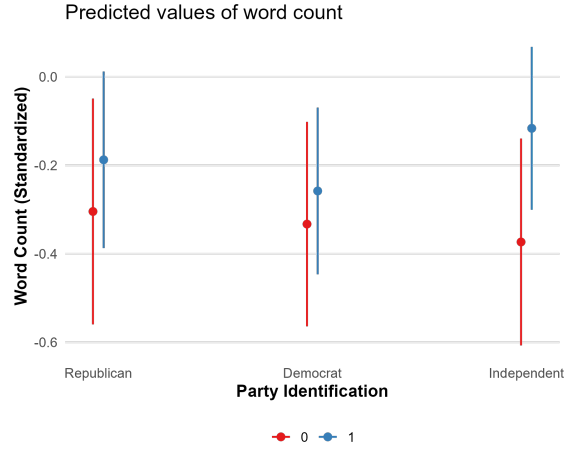


(c) Engagement (No Controls)

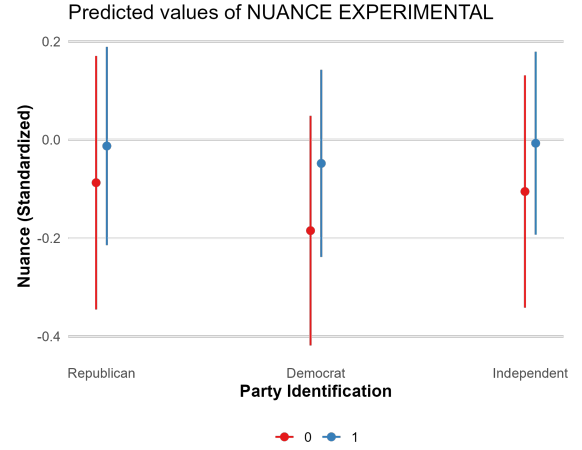
Figure 4: Estimated AI treatment effects (95% confidence intervals) across outcome families in Studies 1 and 2, with no control variables.

As our primary observed outcome is increased comment quality from exposure to AI

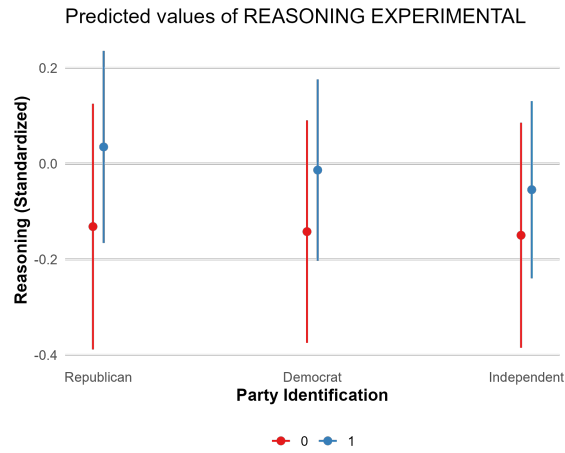
treatment, we also explore whether this increase is consistent or variable across respondent party identification. Across both studies, we find some suggestive evidence that Independent identifiers were most responsive to the treatment, significantly increasing comment quality across multiple measures while Democratic and Republican identifiers showed no difference between treatment status.



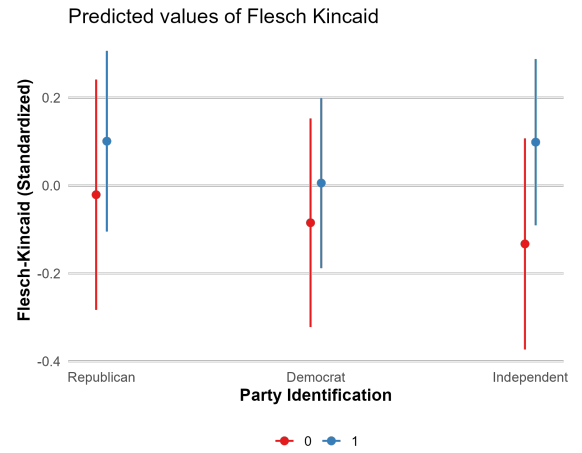
(a) Feeling thermometer: Democratic Party



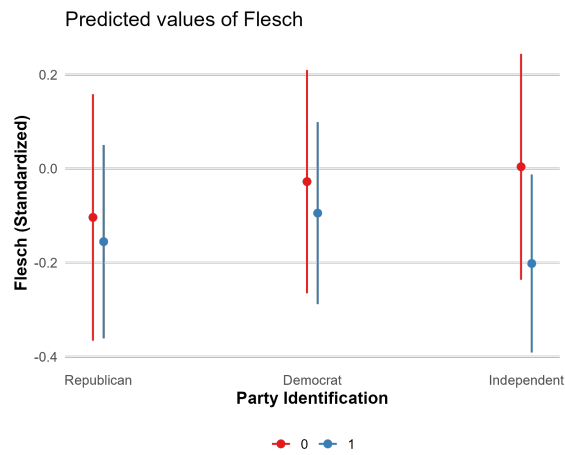
(b) Feeling thermometer: Republican Party



(c) Approval: Democrats in Congress

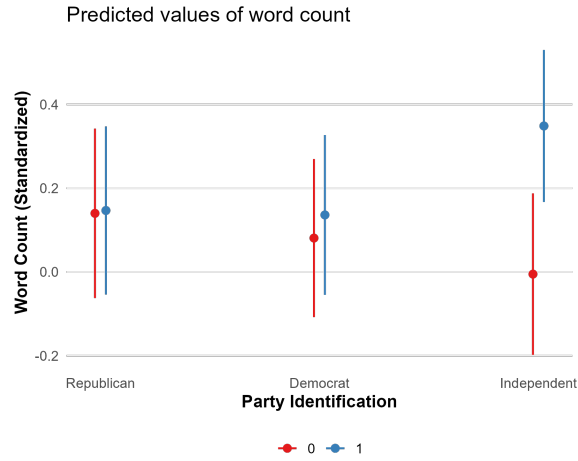


(d) Approval: Republicans in Congress

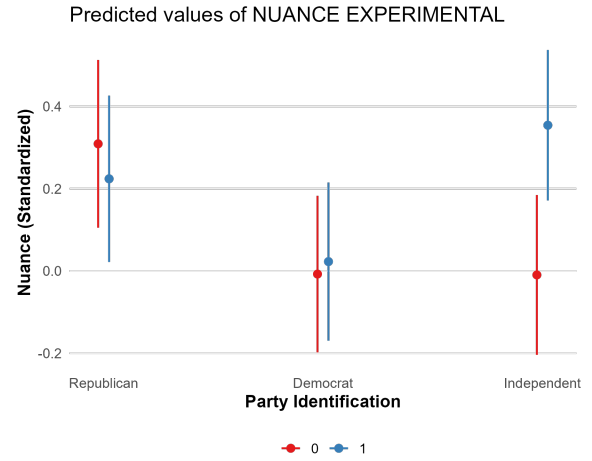


(e) Approval: President

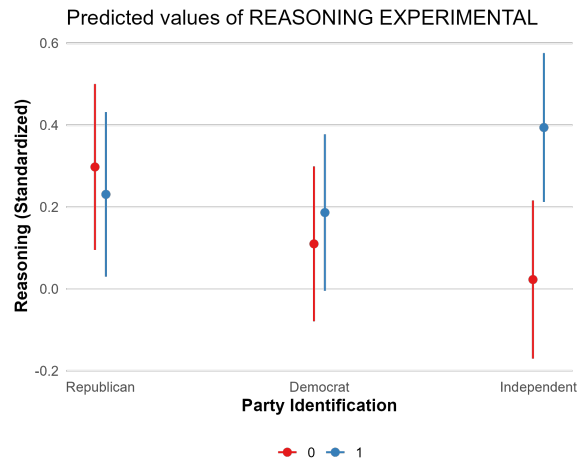
Figure 5: Study 1 — Predicted comment quality measures by PartyID. Panels show interaction-model predictions (95% CI).



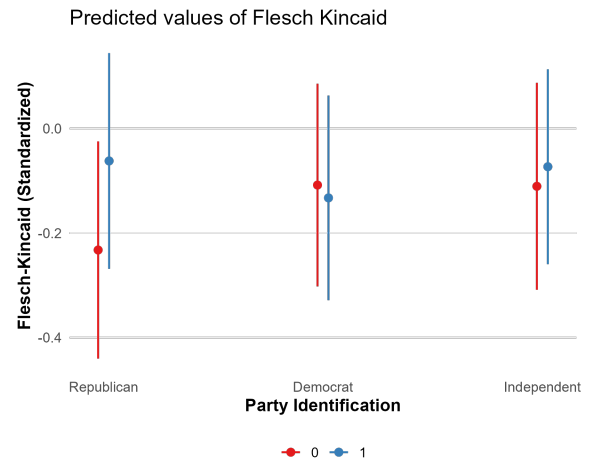
(a) Feeling thermometer: Democratic Party



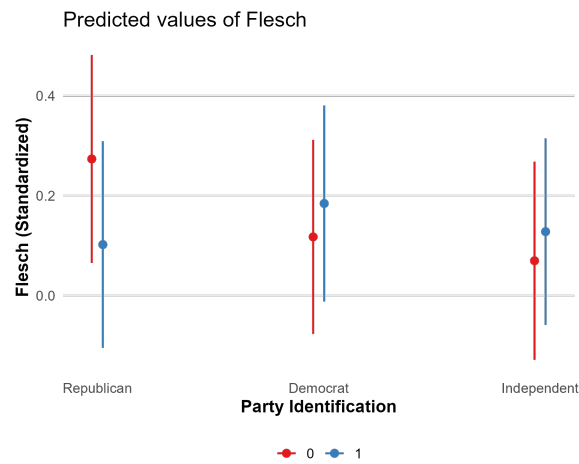
(b) Feeling thermometer: Republican Party



(c) Approval: Democrats in Congress



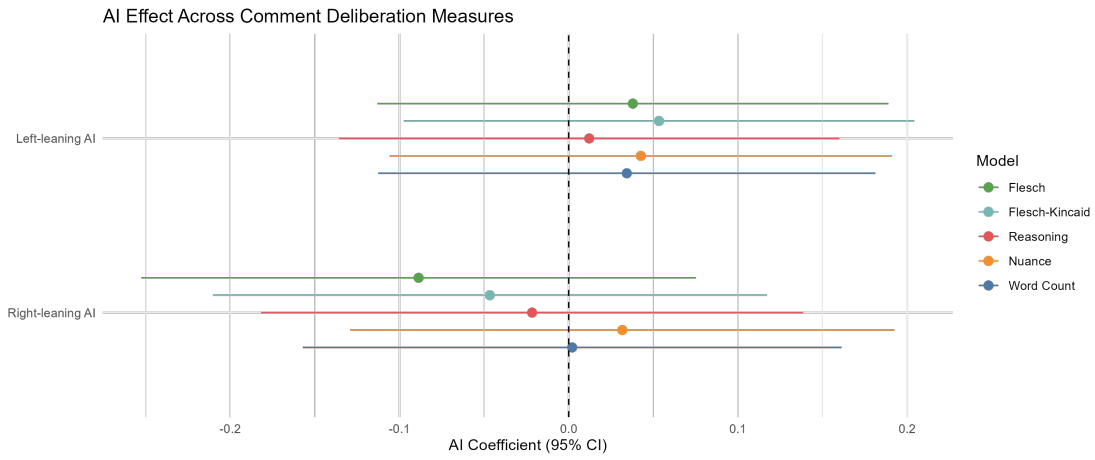
(d) Approval: Republicans in Congress



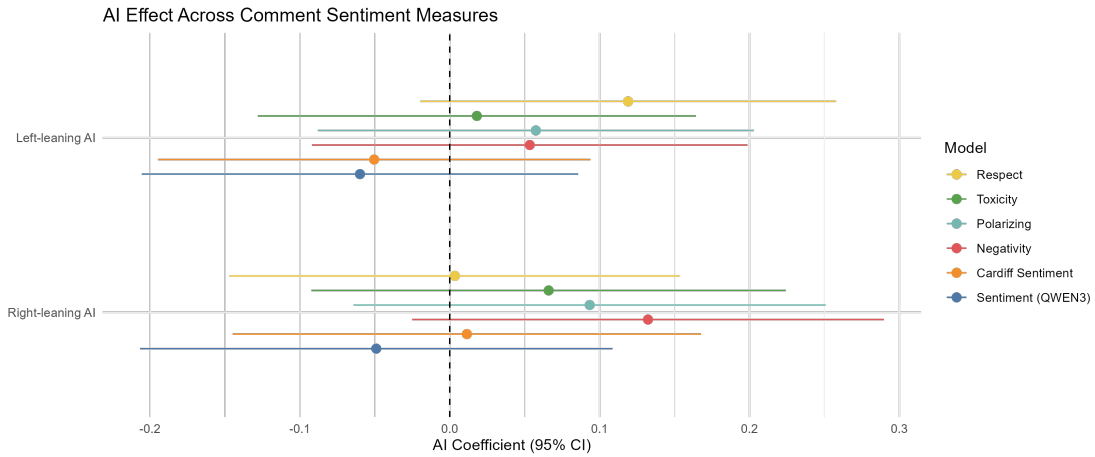
(e) Approval: President

Figure 6: Study 2 — Predicted comment quality measures by PartyID. Panels show interaction-model predictions (95% CI).

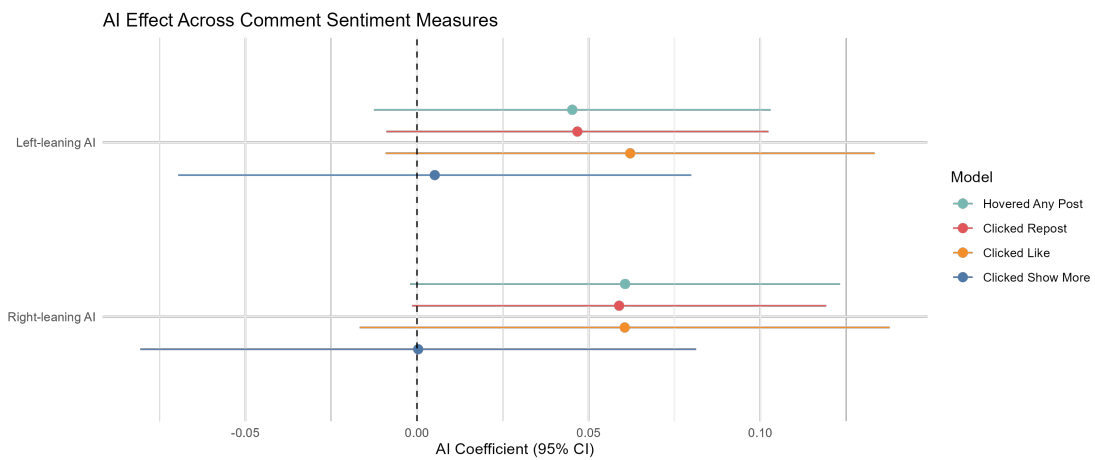
7.1 Main Results Study 1 - AI Treatment Subdivision



(a) Comment Quality by AI Partisan Lean



(b) Comment Sentiment by AI Partisan Lean



(c) Platform Engagement by AI Partisan Lean

Figure 7: Estimated AI treatment effects (95% confidence intervals) on in Study 1, relative to Neutral AI treatment.

As Study 1 contained three different partisan lean treatments for the AI summary, we present the disaggregated results here, with coefficients shown relative to the Neutral AI treatment. Across all measures, there are no significant differences between the partisan and neutral AI treatments. In the platform engagement measures, there is some evidence that both partisan AI treatments increased engagement, especially for the right-leaning treatment, but no coefficient reaches the traditional threshold of significance.

8 Data and Code Availability

De-identified replication data, analysis scripts, and stimulus materials will be made publicly available upon publication at:

[Anonymized Dataverse link].

All analyses were conducted in R version [4.3.3].

References

- Andrew O. Ballard, Ryan DeTamble, Spencer Dorsey, Michael Heseltine, and Marcus Johnson. Dynamics of polarizing rhetoric in congressional tweets. *Legislative Studies Quarterly*, 48(1):105–144, 2023. doi: <https://doi.org/10.1111/lsq.12374>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/lsq.12374>.
- 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, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL <https://aclanthology.org/2020.findings-emnlp.148>.
- Michael Heseltine. Polarizing online elite rhetoric at the federal, state, and local level during the covid-19 pandemic. *American Politics Research*, 52(3):187–202, 2024. doi: 10.1177/1532673X231220647.
- Michael Heseltine. Comparing large language models for text classification: Model selection across tasks, texts, and languages. *OSF Preprints*, 2025. URL https://osf.io/preprints/osf/jvgc5_v2.
- Michael Heseltine and Bernhard Clemm von Hohenberg. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1): 20531680241236239, 2024. doi: 10.1177/20531680241236239.
- Alyssa Lees, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. *arXiv preprint arXiv:2202.11176*, 2022.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL <https://arxiv.org/abs/1908.10084>.