PNAS NEXUS

PNAS Nexus, 2026, **5**, pgag022

https://doi.org/10.1093/pnasnexus/pgag022
Advance access publication 3 March 2026
Brief Report

How latent and prompting biases in AI-generated historical narratives influence opinions

Matthew Shu [a], Daniel Karell [b,c,*], Keitaro Okura [b] and Thomas R. Davidson [d]

[a]Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA
[b]Department of Sociology, Yale University, New Haven, CT 06511, USA
[c]Institution for Social and Policy Studies, Yale University, New Haven, CT 06511, USA
[d]Department of Sociology, Rutgers University, New Brunswick, NJ 08901, USA
*To whom correspondence should be addressed: Email: daniel.karell@yale.edu
Edited By David Rand

Abstract

Large language models (LLMs) can be used to persuade people on a range of issues, particularly through user-driven strategies such as personalizing messages and dialogues intended to change minds. However, their capacity to influence opinions through subtle, latent ideological framing remains relatively understudied. We investigate whether AI-generated historical summaries affect social and political opinions through a preregistered experiment ($N = 1{,}912$). Participants read Wikipedia or GPT-4o summaries of two historical events, with AI summaries maintaining factual accuracy while exhibiting different types of framing biases. Default AI summaries led to more liberal opinions compared with Wikipedia, demonstrating the persuasive capability of LLM's latent biases. Summaries purposefully induced with a liberal framing also led to more liberal opinions, regardless of readers' ideologies. Summaries constructed with a conservative framing produced conservative shifts primarily among conservative readers. These findings demonstrate that the use of AI for learning history can influence opinions through both intrinsic and intentional framing mechanisms, even when the content remains factually accurate. As AI becomes integral to information acquisition, recognizing pathways of influence based not only on user-manipulated content but also on models' latent biases is essential for understanding AI's broader societal impacts.

Keywords: large language models, generative artificial intelligence, AI history, AI framing, AI persuasion

Introduction

The rapidly improving capability of large language models (LLMs) to produce human-like text, images, and video has caused widespread concern about their potential to spread inaccurate information, conspiracy theories, and disinformation (1, 2). A growing body of research thus examines how exposure to LLMs can affect people's views, ranging from beliefs in blatant falsehoods to opinions about legitimate policy debates (2). Much of this work has focused on content intentionally crafted by LLM users to persuade, including propagandistic statements, personalized arguments, and conversations with AI agents (2–8).

This research overlooks a less overt but potentially significant mechanism of persuasion: the underlying, latent biases in LLMs. Biases can be introduced at multiple stages of the LLM training process, resulting in ideological leanings that can be difficult to remove (9–11). In some cases, these biases are unintentional; in other cases, developers can guide models to express particular perspectives. For example, xAI's Grok chatbot generated racist and anti-antisemitic material on X after its instructions were altered to permit "politically incorrect" claims.[a] To be sure, most latent biases, whether intentional or unintentional, do not yield such clearly prejudiced content. Rather, they can produce outputs that appear neutral but convey subtle nuances or emphases that potentially influence opinions (10). We examine this more elusive mechanism of AI persuasion, *latent bias*, and compare its effect to the more commonly studied *prompting bias*, where users explicitly employ or instruct AI to convey a particular perspective (2, 4–8).

Whereas prior work on prompting bias has focused on political and policy contexts (2, 5, 6, 8, 10) or outright misinformation (1, 4, 7), our empirical application examines the use of LLMs to elicit facts. Specifically, we consider historical events, since learning about history informs attitudes and policy preferences (12), people are increasingly using AI tools like ChatGPT and Google's AI Overview to obtain historical information (13),[b] and private companies are using AI to produce alternative information sources, such as xAI's "Grokipedia."[c] Our analyses test how both AI's latent and user-prompted framing of factually accurate historical content can affect individuals' opinions about social and political issues.

We begin our study by evaluating latent biases. We compare default GPT-4o and Wikipedia summaries of two events that took place in the United States during the 20th century: a large

Competing Interest: The authors declare no competing interests.
Received: September 17, 2025. **Accepted:** December 10, 2025

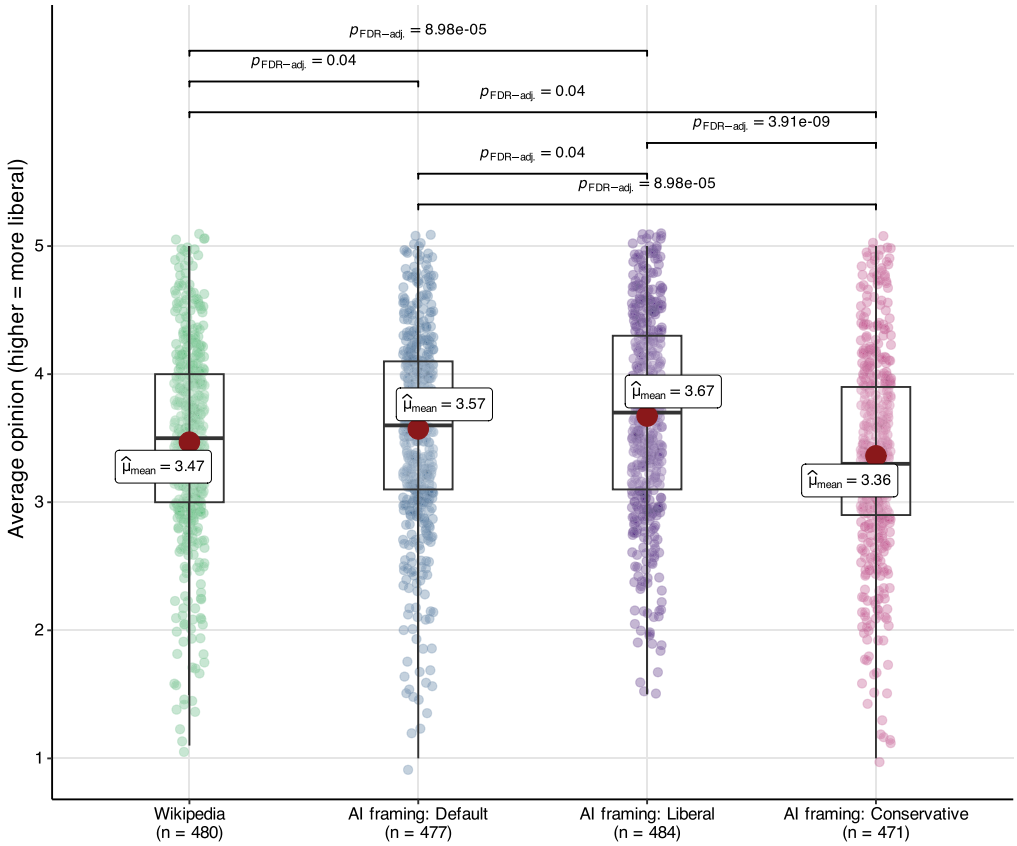

Fig. 1. Pairwise comparisons testing the effect of reading AI summaries of historical events on readers' opinions. Statistical significance calculations incorporated an adjusted false discovery rate (FDR) based on the Benjamini–Hochberg procedure.

labor strike and student protests calling for greater ethnic minority representation in academia. Following recent research on AI persuasion (6), the default summaries are generated by instructing the model to take on the role of a relevant expert (i.e. a "historian and teacher"). To assess the relative influence of latent and prompting biases, we compare similar summaries that purposefully frame the events in politicized ways. This prompting bias is introduced by following recent work that instructed LLMs to adopt a politicized perspective, or "role-play" (8). Namely, we use survey data to seed the model with "opinions," yielding summaries that portray the same historical events with either liberal or conservative framing. The first two research questions are thus:

> RQ1. What is the effect of default-framed AI summaries of historical events on readers' opinions, compared to Wikipedia summaries?

> RQ2. What is the effect of inducing ideological framing in AI summaries of historical events on readers' opinions, compared to Wikipedia summaries?

People filter information through their own ideologies and belief systems. Motivated reasoning theory suggests a tendency to favor belief-confirming information (14), potentially leading to reinforcement from ideologically aligned frames. Meanwhile, encountering counter-attitudinal views might provoke defensive reactions and backlash (15). Therefore, we additionally examine heterogeneous effects of the AI summaries across readers' ideologies.

> RQ3. How do the AI summaries' effects vary based on the correspondence between readers' political ideology and the summaries' ideological framing?

Results

We address RQ1 and RQ2 by comparing the opinions of readers of AI summaries to the opinions of those who read the Wikipedia summaries. Figure 1 shows that both the AI summaries with default framing and the AI summaries with liberal framing led to more liberal opinions compared to the Wikipedia summaries (Default: mean difference = 0.1, Cohen's $d = 0.14$, $P < 0.05$; Liberal: mean difference = 0.21, Cohen's $d = 0.28$, $P < 0.001$). Meanwhile, readers of AI summaries with conservative framing reported more conservative opinions relative to Wikipedia (mean difference = −0.11, Cohen's $d = -0.13$, $P < 0.05$). Similar results were observed when analyzing each event separately.

The effect sizes are similar in magnitude to those in prior research (6). On the response scale, average opinions after reading each summary fell between 3 (moderate) and 4 (somewhat liberal). The average opinion after reading either the default (3.57) or liberal AI text (3.67), compared with the average opinion after reading the Wikipedia text (3.47), represents a shift from a moderate stance to leaning towards a somewhat liberal stance. The average opinion after reading the conservative AI text (3.36) indicates a slight shift towards more moderate opinions.

To examine whether readers' political ideology moderated how politically framed AI summaries affected their opinions (RQ3), we conditioned our preceding analysis on self-reported ideology. The average marginal effects (AMEs) across reader ideology in Fig. 2

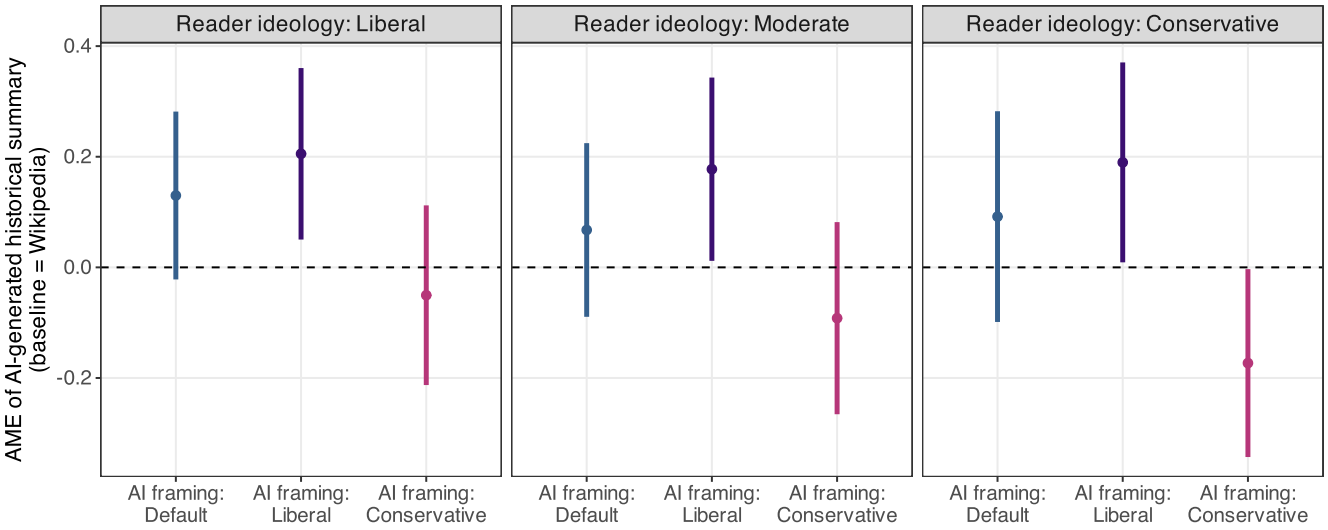

Fig. 2. The AME of reading AI summaries of historical events, relative to reading Wikipedia summaries, conditional upon readers' ideology. Lines denote 95% CIs. Calculations of statistical significance incorporate the Benjamini–Hochberg procedure.

show that the default AI summaries did not affect readers' opinions differently from the Wikipedia summaries for liberal, moderate, or conservative participants. In contrast, the liberal-framed AI summaries led to more liberal opinions for participants across all ideology groups (liberals: AME = 0.205, $P < 0.05$; moderates: AME = 0.177, $P < 0.05$; conservatives: AME = 0.190, $P < 0.05$). The AI summaries with conservative framing resulted in more conservative opinions, although this effect was statistically significant ($\alpha = 0.05$) only for those who already had a conservative outlook (conservatives: AME = −0.173; $P < 0.05$). We conducted Bayesian analyses to evaluate the support for the null findings and found moderate support for all null results and strong support for the null effect of the conservative summary on liberals (see Section G of the Supporting Information [SI] document).

Discussion

Our results suggest that using popular AI tools to learn about history can influence people's opinions on social and political issues. Default GPT-4o-generated summaries of historical events made readers' opinions more liberal compared with Wikipedia summaries, demonstrating the persuasive effects of LLMs' latent biases.

While the experiment tested only two events, supplementary analyses showed that GPT-4o tended to generate liberal-framed default summaries for a wide range of similar events, including anti-union protests and anti-Black race riots. This framing persisted even when we used a simpler system message that did not instruct the model to take on the role of a relevant expert to generate the summaries (see Section L the SI). This suggests that conservative framing in content generated by GPT-4o may require explicit prompting, whereas liberal framing could be the result of both latent and prompting bias.

The main results also indicate that AI summaries offering factually accurate yet ideologically framed historical accounts led to alignment between readers' opinions the summaries' ideological framing. Readers of the liberal summaries reported more liberal opinions, regardless of their own self-reported ideology. Readers of the conservative summaries exhibited more conservative opinions overall, although this effect was primarily driven by conservative respondents.

This study is not without limitations. We focused on two events to ensure precise estimates, so further work is needed to examine whether the findings generalize to a broader range of events and historical periods, as well as other domains beyond history. The degree of latent bias may also vary across models, and future work should consider the impacts of different types of training data and posttraining alignment techniques (11). More research is also needed to understand how the interaction of latent and prompting biases shapes AI's persuasiveness. For example, one explanation for the observed ideological asymmetry (Fig. 2) is that the liberal texts were more congruent with the latent bias of the model, whereas the conservative prompt may have conflicted with the underlying representations used to generate a response (9).

This study advances our understanding of the persuasive effects of AI and, more generally, its potential risks. Previous research on AI persuasion has largely studied the influence of AI material and interactions that are designed to persuade (2, 5, 6, 8)—what we conceptualize as prompting bias. Complementing this work, our findings identify the persuasiveness of AI's latent biases (10) and the implications of using AI to obtain factual information. The presence of latent biases means that simply using chatbots to learn about the past can influence people's opinions at a magnitude comparable to prompting-biased AI material (Fig. 1). Beyond AI's persuasiveness, our study raises important questions about how the increasingly routine and widespread reliance on AI tools to acquire and synthesize information and knowledge (13) can shape our understanding of the social world.

Materials and methods

We conducted a preregistered survey experiment approved by the Institutional Review Board at Yale University (#2000037333). All participants provided informed consent before joining the study. The experiment had 1,912 participants, sampled proportionally to the United States population along several sociodemographic dimensions.

Each subject was randomly assigned to read summaries of two historical events, the Seattle General Strike (SGS) and the Third World Liberation Front (TWLF) student protests, taken from Wikipedia or generated by GPT-4o with either a default AI framing, a liberal framing, or a conservative framing. The SGS was the first "general" strike in the United States, during which 65,000 workers in Seattle stopped work from 1919 February 6 to 11. The TWLF protests involved university student groups in 1968 that campaigned for curricula representing the histories and cultures of students of color, leading to the establishment of Ethnic Studies departments. The SI contains further discussion of these events and information on the construction of the summaries. After reading each summary, participants reported their opinions on related social and political issues, such as their views on the appropriateness of strikes (relevant to the SGS) and the use of curricula to advance social justice causes (relevant to the TWLF). Participants' answers were scored on a five-point scale (1 = extremely conservative; 5 = extremely liberal), and the mean across both events forms our outcome measure.

To address RQ1 and RQ2, we conduct pairwise comparisons between each treatment. The results were adjusted using the Benjamini–Hochberg procedure to account for multiple hypothesis tests. For RQ3, we use an ordinary least squares regression model and the Lin estimator to predict average opinion with an interaction term comprising the type of summaries and reader ideology. Further methodological details are provided in the SI.

Notes

[a] https://www.npr.org/2025/07/09/nx-s1-5462609/grok-elonmusk-antisemitic-racist-content. Accessed July 2025.
[b] https://www.nytimes.com/2025/06/16/magazine/ai-history-historians-scholarship.html. Accessed December 2025.
[c] https://www.nytimes.com/2025/10/27/technology/grokipedialaunch-elon-musk.html Accessed October 2025.

Acknowledgments

The authors thank the editor and anonymous reviewers for their valuable suggestions. The authors also thank Bart Bonikowski and the participants of New York University Department of Sociology's colloquium for their helpful comments and feedback.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

This research was not supported by any external funding.

Author Contributions

Matthew Shu (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—original draft), Daniel Karell (Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Writing—original draft, Writing—review & editing), Keitaro Okura (Data curation, Investigation, Methodology), and Thomas R. Davidson (Formal analysis, Investigation, Methodology, Writing—original draft, Writing—review & editing)

Data Availability

The data and coding script used to conduct the analyses in this study can be found at https://osf.io/zy9eu/overview.

References

1 Spitale G, Biller-Andorno N, Germani F. 2023. AI model GPT-3 (dis)informs us better than humans. *Sci Adv*. 9(26):eadh1850.
2 Salvi F, Ribeiro MH, Gallotti R, West R. 2025. On the conversational persuasiveness of GPT-4. *Nat Hum Behav*. 9(8):1645–1653.
3 Costello TH, Pennycook G, Rand DG. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*. 385:eadq1814.
4 Goldstein JA, Chao J, Grossman S, Stamos A, Tomz M. 2024. How persuasive is AI-generated propaganda? *PNAS Nexus*. 3(2): pgae034.
5 Hackenburg K, Margetts H. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proc Natl Acad Sci U S A*. 121(24):e2403116121.
6 Argyle LP, *et al.* 2025. Testing theories of political persuasion using AI. *Proc Natl Acad Sci U S A*. 122(18):e2412815122.
7 Dash S, Xu Y, Jalbert M, Spiro ES. 2025. The persuasive potential of AI-paraphrased information at scale. *PNAS Nexus*. 4(7): pgaf207.
8 Hackenburg K, Ibrahim L, Tappin BM, Tsakiris M. *Comparing the persuasiveness of roleplaying large language models and human experts on polarized U.S. political issues*. AI & Society, 2025.
9 Rozado D. 2024. The political preferences of LLMs. *PLoS One*. 19(7): e0306621.
10 Potter Y, Lai S, Kim J, Evans J, Song D. Hidden persuaders: LLMs' political leaning and their influence on voters. In: Al-Onaizan Y, Bansal M, Chen Y-N, editors. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, November 2024. p. 4244–4275.
11 Bai X, Wang A, Sucholutsky I, Griffiths TL. 2025. Explicitly unbiased large language models still form biased associations. *Proc Natl Acad Sci U S A*. 122(8):e2416228122.
12 Ginsberg B, Bachner J. *Warping time: how contending political forces manipulate the past, present, and future*. University of Michigan Press, Ann Arbor, 2023.
13 Karell D, Shu M, Davidson T, Okura K. 2025. Generating the past: how artificial intelligence summaries of historical events affect knowledge. *Soc Sci Comput Rev*. https://doi.org/10.1177/0894439325140974.
14 Taber CS, Lodge M. 2006. Motivated skepticism in the evaluation of political beliefs. *Am J Pol Sci*. 50(3):755–769.
15 Bail CA, *et al.* 2018. Exposure to opposing views on social media can increase political polarization. *Proc Natl Acad Sci U S A*. 115(37):9216–9221.