# The Elusive Verification:
# How People Respond to AI Information About Politics

**Working paper, Rutgers Computational Social Science Lab: State of the Prompt Project**

By Katherine Ognyanova, Mark Aakhus, Kiran Garimella, Shagun Jhaver,
Eu Gene Lee, Jianing Li, Sarah Shugars

October 2026

**Abstract.** Americans are routinely urged to be mindful of the information they encounter. Engaged consumers are expected to check sources, evaluate credibility, examine evidence, and think critically. With conversational AI emerging as a new information source, these expectations now extend to content we receive from chatbots. Yet we know little about people's reactions to the answers that chatbots give them in everyday life. In this study, we analyze a year of ChatGPT logs from 73,852 members of an opt-in US consumer panel. We measure how users react to AI answers in 283,139 conversations about politics and government. We find that observed behavior fell short of normative expectations. Pushback and requests for evidence were rare, even when the stakes were high. When users did challenge an AI answer, they mostly did it by asserting views they already held rather than by seeking new evidence.

## Theoretical background

Conversational AI has become an important part of everyday life. By mid-2026, nearly three quarters (72%) of Americans had used an AI chatbot, and nearly one in five had asked those tools about politics (Ognyanova et al., 2026). Scholars, journalists, and policymakers have been concerned about AI as a source of social and political information. A 2026 audit found that chatbots often make factual errors when answering questions about elections (Brennan Center for Justice, 2026). Conversational AI can give politically biased answers (Peng et al., 2026; Santurkar et al., 2023), as well as shift people's political attitudes (Hackenburg et al., 2025; Lin et al., 2025).

One relevant area that remains understudied is how people react when they receive political information from AI tools. Do users tend to accept an AI answer as legitimate and persuasive, and do they ever examine, verify, or challenge the response?

Several bodies of work offer potentially divergent answers to that question. *Machine heuristic* research suggests that we may see machines as objective and accurate (Sundar, 2008; Sundar & Kim, 2019), and be more willing to defer to them. Work on *epistemic vigilance* indicates that we evaluate information in light of what we already believe and who it comes from, accepting some claims and rejecting others (Mercier, 2020). In a political context, the *motivated reasoning* literature predicts that we will be more critical of claims that go against our views (Taber & Lodge, 2006).

Information literacy and critical thinking frameworks recommend that we take little for granted. We are encouraged to assess the accuracy, authority, objectivity, and currency of a source before trusting it (Metzger, 2007). More recent guides recommend lateral reading, leaving a source to see what others say about it (Caulfield & Wineburg, 2023). AI literacy frameworks advise users to critically evaluate these new technologies (Long & Magerko, 2020).

In the context of AI, how users evaluate information has been examined primarily through surveys and experiments. We know little about people's actual behavior when they are engaged in an everyday

conversation with a chatbot. Large-scale publicly available sources of AI conversation logs do exist, but they come from users of free research tools and rarely include demographic information about participants (Zhao et al., 2024; Zheng et al., 2024). Most available estimates of user reactions come from analyses conducted by AI companies rather than independent researchers (Chatterji et al., 2025).

## Research Questions and Hypotheses

This study provides behavioral evidence on how users react to political information from AI. We examine 283,139 conversations about politics and government, drawn from a full year of ChatGPT use by 73,852 members of an opt-in US consumer panel. For each conversation, we identify user behavior after ChatGPT answered: whether they accepted the answer, disputed it, asked for evidence, brought in an outside source, or simply moved on. The first research question we ask here is:

**RQ1:** How often do users react to ChatGPT's answers about politics, and in what ways?

Research on online credibility suggests that in practice, information verification may be limited. Double-checking facts takes effort, and people tend to rely on cues rather than verify claims themselves (Metzger et al., 2010). We therefore expect that:

**H1:** Users will explicitly accept ChatGPT's answers more often than they verify those answers by asking for evidence or consulting another source.

If disagreement is grounded in motivated reasoning, it may take the form of counterarguing rather than requests for evidence (Taber & Lodge, 2006). Our next question addresses forms of pushback:

**RQ2:** When people challenge an answer, what form does the pushback take?

We further examine who pushes back and when:

**RQ3:** Which features of the conversation and user are associated with acceptance, pushback, and requests for evidence?

Motivated reasoning predicts that users who already hold a view will be more likely to dispute an answer that goes against it. Pushback may also be more common for contested topics such as ideology and political figures, where partisans perceive even balanced information as biased against their side (Vallone et al., 1985) and people trust algorithms less (Castelo et al., 2019). Epistemic vigilance predicts that scrutiny will increase when being wrong is costly or risky (Mercier, 2020). Thus, disputing and verification might be more common in high-stakes conversations. We therefore propose the following hypotheses:

**H2:** Users who state a view before ChatGPT answers will be more likely to dispute the answer.

**H3:** Disputes will be more common in conversations about contested topics, such as ideology and political figures.

**H4:** Disputing and verifying information will be more common in high-stakes conversations than in low-stakes ones.

## Method

The study uses data from a US commercial consumer panel whose members consented to the passive collection of their ChatGPT conversations. The corpus is licensed for research and contains 3.9 million conversations held by 73,852 panelists between July 2025 and June 2026. The data include the panelists' self-reported demographics.

Table 1. *Demographic composition of the panel and of the political subset (% of panelists with a recorded value). The political subset includes panelists with at least one political conversation. For 3.0% of the panel and 2.6% of the political subset, we have no demographic record.*

| Variable | | Panel (N = 73,852) | Political subset (n = 33,926) |
|---|---|---|---|
| Gender | Female | 66.5 | 67.6 |
| | Male | 33.5 | 32.4 |
| Age | 13-17 | 5.4 | 4.8 |
| | 18-24 | 23.1 | 23.4 |
| | 25-34 | 26.2 | 27.4 |
| | 35-44 | 23.5 | 24.0 |
| | 45-54 | 14.3 | 13.6 |
| | 55-64 | 6.1 | 5.6 |
| | 65+ | 1.5 | 1.3 |
| Race and ethnicity | White | 44.4 | 43.4 |
| | Black | 29.4 | 29.7 |
| | Hispanic or Latino | 13.7 | 14.0 |
| | Asian | 2.8 | 3.5 |
| | Native American or Alaskan | 1.9 | 1.8 |
| | Middle Eastern | 0.7 | 0.8 |
| | Pacific Islander | 0.6 | 0.6 |
| | Other | 6.5 | 6.2 |
| Hispanic | Yes | 20.9 | 21.5 |
| | No | 79.1 | 78.5 |
| Education | Some high school or less | 18.3 | 15.4 |
| | High school | 34.9 | 31.6 |
| | Some college or associate degree | 31.3 | 34.1 |
| | Bachelor's degree | 10.9 | 13.1 |
| | Graduate study or degree | 4.7 | 5.8 |
| Household income | Under $25,000 | 42.9 | 38.0 |
| | $25,000-34,999 | 15.8 | 15.5 |
| | $35,000-49,999 | 12.9 | 13.5 |
| | $50,000-74,999 | 13.6 | 15.2 |
| | $75,000-99,999 | 6.8 | 7.9 |
| | $100,000 or more | 8.0 | 9.8 |

Political conversations were identified by a large language model (LLM) classifier validated against human coding (precision 0.85, recall 0.92). The political set includes public-affairs discussions as well as personal dealings with government, such as questions about taxes, benefits, or court procedures. The classifier identified 283,139 political conversations held by 33,926 panelists. The panel is not a probability sample. It overrepresents women and lower-income panelists (Table 1).

An open-weight large language model (*gpt-oss-safeguard-20b*) then read each political conversation in full, including ChatGPT's replies, and coded the setting and the user's reactions. The coding covers *stakes* (low, moderate, or high), *prior view* (whether the user stated a view or asked a leading question before ChatGPT answered), and *compliance* (whether ChatGPT did what was asked fully, partly, or not at all). Conversation *scope* (societal, personal, or mixed) was previously identified through the political classification.

The model coded nine possible user reactions to a ChatGPT answer: (1) accepting it, (2) asking what it means, (3) correcting ChatGPT's reading of the request, (4) disputing it, (5) asking for evidence, (6) bringing in an outside information source, (7) asking for the other side, (8) objecting to how ChatGPT answered, or (9) responding to a model refusal. No reaction was recorded if the user did not follow up, made a new request, or changed the topic. A single conversation could contain multiple user reactions.

A second LLM coding pass revisited conversations with user reactions and collected additional information. That pass could also discard previously identified user reactions if it found insufficient evidence for them. Our results include only confirmed user reactions. Agreement between two independent runs of the LLM coding was substantial ($\kappa$ = 0.75-0.77 for stakes, acceptance, and disputes). Validation against human coders for this stage is currently in progress.

We report the share of conversations containing each user reaction. We also estimate logistic regressions for disputing, checking (asking for evidence or bringing in a source), and explicit acceptance among the 200,155 conversations in which the user followed up after the initial ChatGPT answer. The regression models include the stakes, the user's prior view, the scope of the conversation (personal, public affairs, or mixed), its political domains, whether ChatGPT complied with the request, and the user's age, gender, and education, with standard errors clustered by panelist.

Since the panel is not a probability sample, we also weighted adult panelists to US population benchmarks from the 2024 American Community Survey. Weighting changed the estimates very little: the share of conversations with each user reaction changed by less than 0.3 percentage points, and the main associations in the regression models retained their direction and were of similar size. We report unweighted results below.

## Results

Exploring RQ1, we found that most ChatGPT answers received no reaction. Users wrote again after a ChatGPT answer in 72% of political conversations, but only 19% contained any reaction. Most follow-ups were simply new requests or changes of topic.

The most common confirmed reaction was explicit acceptance (7.1% of conversations). Moreover, nearly two thirds of acceptances (63%) were simply "thanks" or "ok." As predicted by H1, verification was rare. Users asked for evidence in 0.22% of conversations and brought in an outside source in 0.36%, while 1.6% disputed an answer (Figure 1). People rarely pushed back: among the 33,926 panelists with at least one political conversation, 7.6% ever disputed an answer and 3.6% ever asked for evidence or brought in a source. Pushback was also concentrated: half of all disputed conversations came from 527 users, 1.6% of those who discussed politics with ChatGPT.

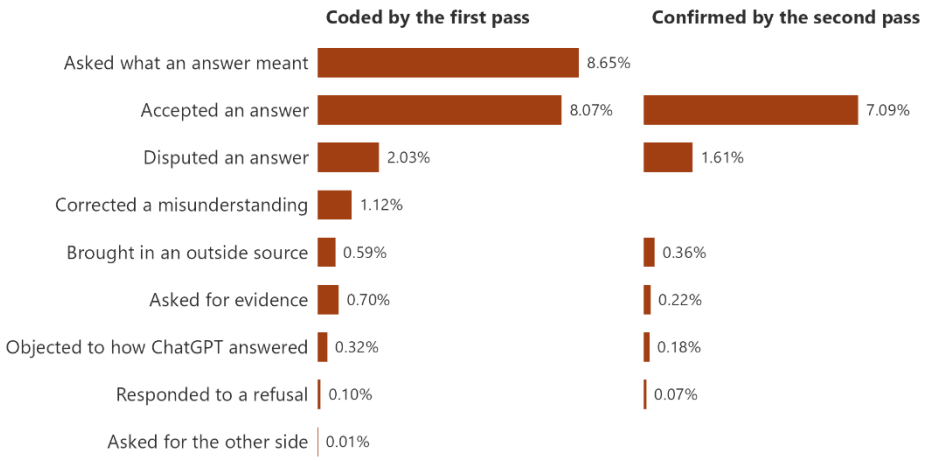


Figure 1. *Share of political conversations with each reaction, coded by the first LLM pass and confirmed by the second LLM pass. The second pass examined six of the nine reactions; the other three were not coded a second time.*

RQ2 asked how people challenge ChatGPT answers (Table 2)**.** When users disputed an answer, they usually told the model what they believed instead (80% of disputes). They commonly gave no reason (44%) or cited their own knowledge or experience (39%). Disputes rarely settled the matter. Most (69%) ended up unresolved, 7.5% with the user rejecting the answer, and 4.6% with the user accepting what ChatGPT had said.

Table 2. *Share of conversations with each user reaction (percentages within each reaction category).*

| User reaction | Reaction type | % |
|---|---|---|
| Disputed (n = 4,561) | Says what is true instead | 80.0 |
| | Doubts without a counterclaim | 18.1 |
| | Points to a specific error | 6.4 |
| Reason given for a dispute | None | 44.4 |
| | Own knowledge or experience | 38.5 |
| | Values or opinion | 10.1 |
| | An outside source | 5.1 |
| Outside source (n = 1,008) | A website | 37.7 |
| | An official source (agency, law, form) | 20.5 |
| | News media | 16.5 |
| | A professional the user consulted | 7.3 |
| Asked for evidence (n = 626) | Sources or citations | 74.0 |
| | To check again or look it up | 16.8 |
| How disputes ended | Unresolved | 68.8 |
| | User rejected the answer | 7.5 |
| | User accepted the answer | 4.6 |
| | No signal either way | 19.1 |

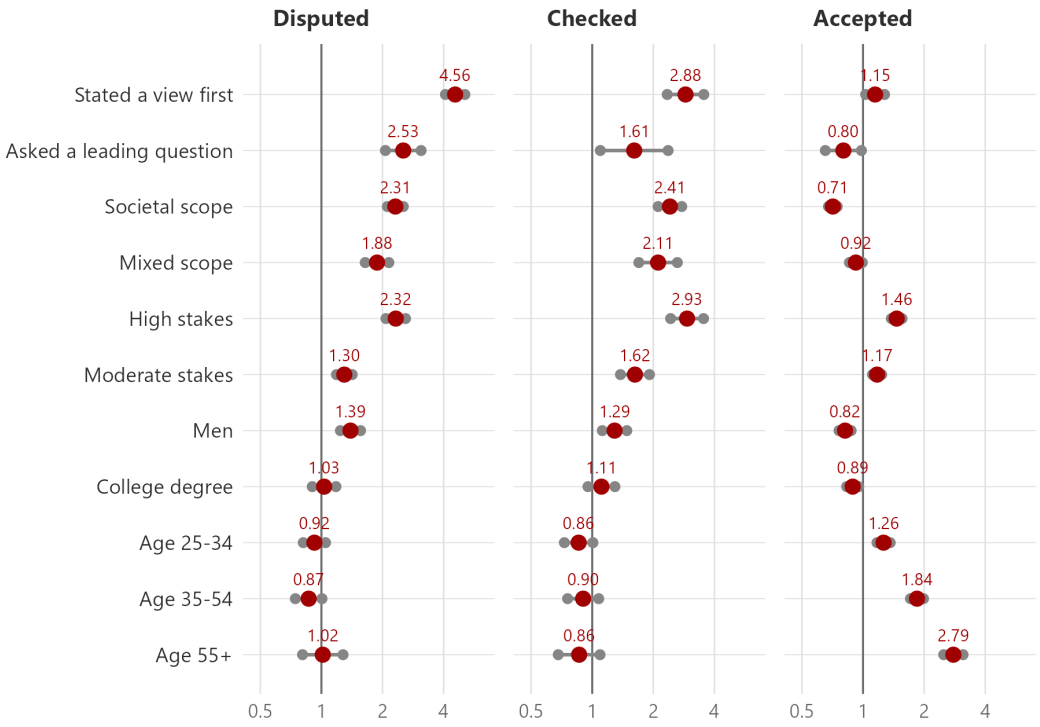


Figure 2. *Adjusted odds ratios from logistic regressions of disputing, checking (asking for evidence or bringing in an outside source), and accepting an answer, among conversations in which the user wrote again (n = 200,155; 28,713 panelists). Models also include political domains and whether ChatGPT complied. Red dots represent odds ratios, and gray lines indicate 95% confidence intervals from panelist-clustered standard errors, on a log scale. Reference categories: no prior view, personal scope, low stakes, women, no college degree, ages 13-24.*

In examining when pushback was more common (RQ3), we found that challenging ChatGPT's answer was associated with stating a prior view, higher conversation stakes, and a public rather than personal scope. In adjusted models (Figure 2), users who stated a view in advance had more than 4 times the odds of disputing the ChatGPT answer (OR = 4.56, 95% CI [4.08, 5.09]), consistent with H2. Public affairs conversations were disputed more than those about the user's own dealings with government (OR = 2.31), as were high-stakes conversations (OR = 2.32), consistent with H4.

Men were more likely to dispute answers than women (OR = 1.39), while age and a college degree made no difference. Checking followed a similar pattern and increased with high stakes (OR = 2.93). Acceptance, however, increased steeply with age: relative to users aged 13-24, the odds of accepting an answer were 1.26 times as high at 25-34, 1.84 times at 35-54, and 2.79 times at 55 and older. Acceptance was also lower in public-affairs conversations (OR = 0.71) and slightly lower among college graduates (OR = 0.89).

Consistent with H3, conversations about ideology, political figures, international affairs, and civil rights were disputed 2-3 times as often as conversations about social welfare, education, or government operations (Figure 3). Conversations about political figures (OR = 1.87), ideology (1.48), the economy (1.44), and civil rights (1.35) remained more likely to be disputed in adjusted models, and welfare conversations were less likely (0.73). Somewhat surprisingly, elections were disputed no more than other topics in those models.

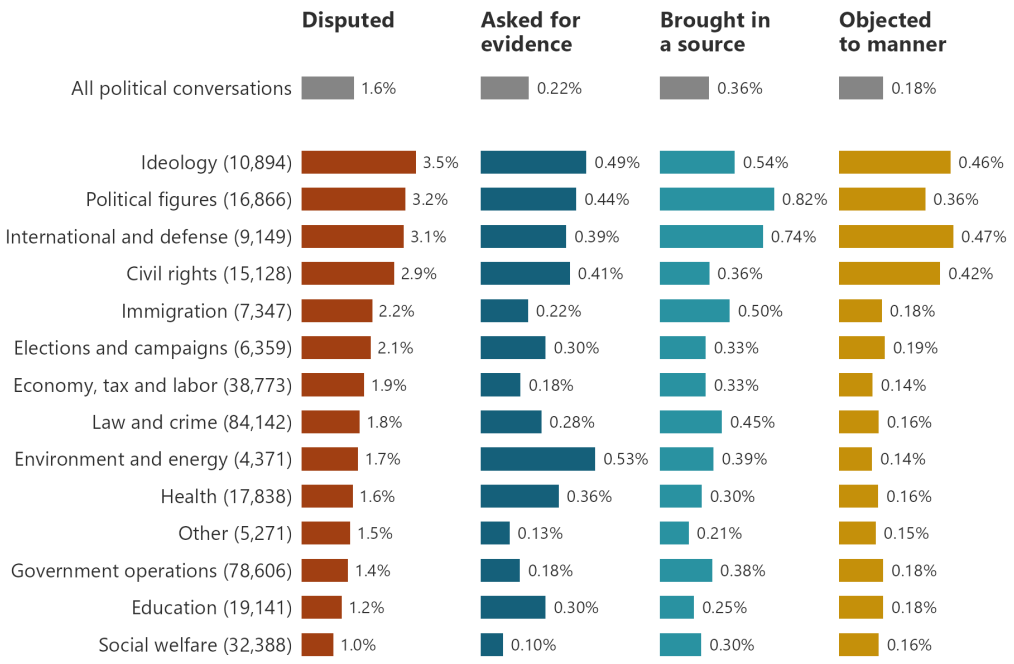


Figure 3. *Pushback by political domain, as a share of each domain's political conversations.*

## Discussion

The findings suggest that users rarely challenge or explore further answers about politics. Most panelists in our data neither explicitly accepted the answers nor checked them systematically: they simply saw the answer and moved on. The high verification standards of literacy education were rarely met. Even in high-stakes conversations, fewer than one in a hundred included a request for evidence or an outside source.

When users did challenge AI answers, this looked less like verification and more like defending a view they already held. Disputes came disproportionately from users who already had expressed a prior opinion about the matter. Challenges concentrated on contested topics, rarely relied on evidence, and seldom ended in agreement. This pattern is consistent with accounts of motivated reasoning (Taber & Lodge, 2006), as well as with research showing that online verification is rare (Metzger, 2007; Metzger et al., 2010).

These results have important implications for our understanding of the risks of conversational AI. The accuracy and objectivity of chatbots become especially relevant when users rarely contest what AI tells them. Without verification, errors are likely to go unnoticed. At the same time, the pushback that does occur comes from a small group of users with strong views. Since language models tend to conform when challenged (Sharma et al., 2024), chatbots may accept false premises if the user is sufficiently persistent.

The study has several limitations. We cannot observe whether users verify information outside ChatGPT conversations. Additionally, the panel is not a probability sample, and the analysis covers political conversations only. Our coding examines user response, not whether ChatGPT or the user was objectively correct. Images, which some users send in reply, are not available. Finally, the nature of the analysis means that the associations we report cannot be interpreted as causal.

## References


Brennan Center for Justice. (2026). *Does AI fight or fuel election disinformation?* https://www.brennancenter.org/our-work/research-reports/does-ai-fight-or-fuel-election-disinformation

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Caulfield, M., & Wineburg, S. (2023). *Verified: How to think straight, get duped less, and make better decisions about what to believe online*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226829845.001.0001

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). *How People Use ChatGPT* (Working Paper No. 34255). National Bureau of Economic Research. https://doi.org/10.3386/w34255

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, *390*(6777), eaea3884. https://doi.org/10.1126/science.aea3884

Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., Pennycook, G., & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, 1–8. https://doi.org/10.1038/s41586-025-09771-9

Long, D., & Magerko, B. (2020). *What is AI literacy? Competencies and design considerations*. 1–16. https://doi.org/10.1145/3313831.3376727

Mercier, H. (2020). *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, *58*(13), 2078–2091. https://doi.org/10.1002/asi.20672

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication*, *60*(3), 413–439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

Ognyanova, K., Singh, V., Ha, E., & Uslu, A. (2026). *National AI Opinion Monitor: AI attitudes, work, and regulation in America*. Rutgers University. https://www.naiom.net

Peng, T.-Q., Yang, K., Lee, S., Li, H., Chu, Y., Lin, Y., & Liu, H. (2026). Beyond partisan leaning: A comparative analysis of political bias in large language models. *Journal of Information Technology & Politics*, *0*(0), 1–18. https://doi.org/10.1080/19331681.2026.2646990

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). *Whose Opinions Do Language Models Reflect?* (arXiv:2303.17548). arXiv. https://doi.org/10.48550/arXiv.2303.17548

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. *The Twelfth International Conference on Learning Representations (ICLR 2024)*. https://arxiv.org/abs/2310.13548

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press.

Sundar, S. S., & Kim, J. (2019). Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, 1–9. https://doi.org/10.1145/3290605.3300768

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology*, *49*(3), 577–585. https://doi.org/10.1037/0022-3514.49.3.577

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., & Deng, Y. (2024). WildChat: 1M ChatGPT interaction logs in the wild. *The Twelfth International Conference on Learning Representations (ICLR 2024)*. https://openreview.net/forum?id=Bl8u7ZRlbM

Zheng, L., Chiang, W.-L., Sheng, Y., Li, T., Zhuang, S., Wu, Z., Zhuang, Y., Li, Z., Lin, Z., Xing, E. P., Gonzalez, J. E., Stoica, I., & Zhang, H. (2024). LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. *The Twelfth International Conference on Learning Representations (ICLR 2024)*. https://openreview.net/forum?id=BOfDKxfwt0