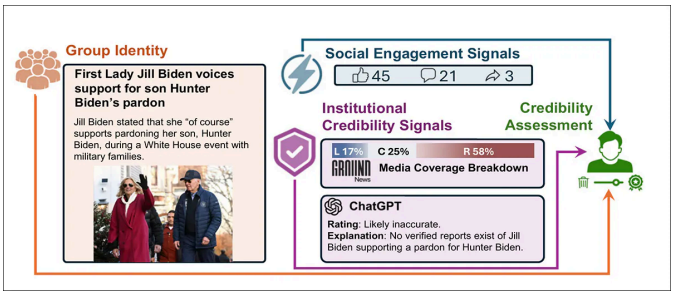
Research Suggests ChatGPT Has High Credibility as a News Provider

Originally published at https://www.unite.ai/research-suggests-chatgpt-has-high-credibility-as-a-news-provider/

New research suggests that ChatGPT's fact-check labels beat likes, shares, and even trusted news brands when it comes to shaping what people believe and want to share online.

A new study of 1,000 people has found that when ChatGPT gives a credibility score to political news, it often changes what people believe, and whether they wanted to share it – no matter what their initial political views. More traditional influence factors such as likes or shares didn't have much effect, but the AI's judgment strongly shaped how trustworthy the news seemed:



From the new paper, an illustration of how people evaluate a news headline, based on three interacting cues: whether the story aligns with their political identity; how much engagement it appears to have on social media; and the credibility signals provided by institutions or AI systems. These influences combine to shape both perceived accuracy and the likelihood of sharing the content. Source

The extent to which AI summaries are taken as credible news sources is perhaps one of the most important topics in the media for many years, not least because Google's AI summaries have drained traffic away from most major online media outlets in 2025, with no certainty where this transfer of power might be headed in the long or even medium term.

The authors of the new work state:

> *'These findings highlight both the potential and risk of algorithmic feedback in shaping public understanding. AI-generated cues can help mitigate bias and enhance credibility discernment, but their influence varies by political identity and carries ethical risks related to overreliance […]*
>
> *'…The outsized influence of ChatGPT highlights a critical trade-off: while persuasive, AI feedback may displace critical thinking if not carefully framed.'*

The authors suggest that future research should focus on improved AI credibility interventions that are more independent and supportive of known facts, and they observe that this is especially important in topics that are polarizing.

Besides the apparent ascendance of ChatGPT as an 'authority', and the unexpectedly low influence of sharing signals in the tests (which the authors think may be attributable to the slightly sterile test conditions), another interesting outcome from the researchers' tests relates to demographics and gender:

> *'Women and racially minoritized participants (particularly Black and Latine users) responded more positively to AI-based feedback than institutional labels.'*

The new work is titled *AI Credibility Signals Outrank Institutions and Engagement in Shaping News Perception on Social Media*, and comes from three researchers at the University of Notre Dame.

Let's take a closer look at the paper's methods and conclusions.

Method

Four hypotheses were tested: that people would rate headlines as more accurate when those headlines reflected their political views; that AI-generated credibility scores would be more influential than signals from established institutions; that headlines with high engagement (i.e., measured through likes, shares, or comments) would be seen as more trustworthy; and that users would tend to accept the AI's judgment about a headline's accuracy, *even when it conflicted with their own beliefs*.

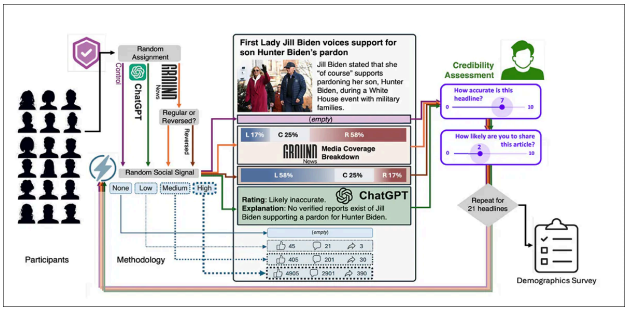
Participants in the study were sourced from the Prolific platform, and were required to be fluent in English, as well as regular consumers of news; and a process of randomization ensured that the participants represented a diversity of racial and gender viewpoints.

In the experiment*, the participants were divided into four groups, each shown a different type of feedback alongside the news headlines.

In the first group (the *control* condition) no extra information was given about how credible the headline might be; in the second group, headlines were labeled with a bias rating from GroundNews, a service that classifies news sources as left-, center-, or right-leaning.

A third group saw the same GroundNews labels, but *deliberately flipped*, creating a mismatch intended to test whether users would detect the distortion.

The final group was shown credibility assessments written by ChatGPT, offering a short explanation and a rating such as *'likely inaccurate'*:



Conceptual schema for the experiment: each participant was shown political headlines with varying combinations of credibility labels and social engagement signals. Responses were collected on how accurate each headline seemed, and how likely it was to be shared, with the full sequence repeated across 21 items.

Data and Tests

Each participant was shown a sequence of 21 political news headlines. For each headline, the level of social engagement was varied; sometimes it showed no likes or shares, sometimes many. These engagement signals were randomized to avoid fixed patterns.

The headlines themselves came from a mix of political perspectives, and were tagged as *left-leaning*, *right-leaning*, or *centrist*.

After each headline, participants were asked how *accurate* they thought it was, and whether they would consider sharing it. Because each participant had already stated their own political affiliation, it was possible to analyze whether headlines from the same side of the political spectrum were rated more favorably.

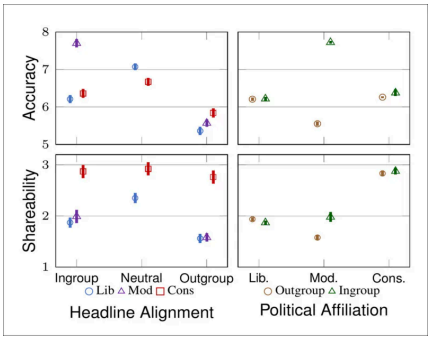
Participants scored each headline twice: once for how accurate it seemed, and once for how likely they would be to share it, with their responses measured on a 0-10 scale.

The researchers then combined these answers with additional data: demographic and media-use information from each participant; political tags for each headline; the type of credibility signal shown; and the level of social engagement assigned.

Political Identity's Influence

In the first test, which probed whether people rate headlines as more accurate when the political stance of the headline matches their own views, the results indicated that participants were more likely to believe headlines that matched their political views – but that this depended on the group.

Moderates showed the strongest bias toward their own side, while liberals and conservatives tended to trust center-leaning headlines the most; and overall, neutral headlines were rated as more accurate than left or right-leaning ones.



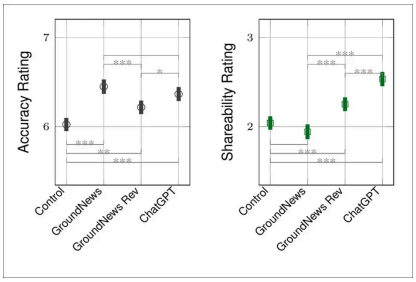
An illustration of how accuracy and shareability ratings changed depending on both the headline's political stance and the participant's political affiliation. Moderates were most likely to rate 'ingroup' headlines as accurate, while liberals and conservatives favored center-aligned headlines. Sharing behavior followed a similar pattern, with limited ingroup bias outside the moderate group.

In the results of this first experiment, a modest ingroup effect on shareability was found; but only for moderates, who were more inclined to share politically aligned (i.e., neutral) headlines. Liberals and conservatives showed no such tendency.

Analysis of variance (ANOVA), a method used to spot differences between groups, showed that alignment affected sharing only when it interacted with political identity. Credibility and sharing were linked, but only moderates showed a clear pattern across both.

Institutional vs. AI Credibility

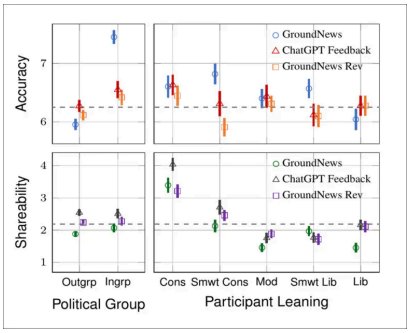
The next test asked whether people trust AI ratings more than traditional sources such as news-rating websites – especially when the ratings might disagree with their politics:



Credibility and shareability ratings across feedback sources: all three signals increased accuracy compared to control, with GroundNews yielding the highest ratings. However, ChatGPT produced the largest gains in shareability, suggesting it had broader persuasive impact. Error bars show 95% confidence intervals, and asterisks mark significant pairwise differences.

All feedback increased perceived accuracy; but GroundNews was most effective when it aligned with the user's politics.

ChatGPT raised accuracy ratings across the board, indicating it was viewed as more neutral. Conservatives were less swayed by GroundNews, yet responded to ChatGPT similarly to other groups:



Here we can see the effects of ChatGPT feedback on perceived accuracy. The results indicate that trust in institutional signals depends on alignment, while trust in algorithmic signals does not. ChatGPT boosted both credibility and shareability across groups – especially for conservatives.

Social Metrics Have Little Impact

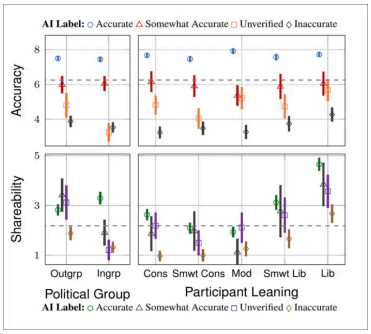
The third analysis tested whether visible social engagement cues such as likes, shares, and comments would boost credibility or shareability by acting as social proof; but no such effect was observed.

Tests found that engagement levels, such as likes or shares, had no real effect on how accurate headlines seemed, and only a weak, unreliable effect on how shareable they *felt*; unlike algorithmic or institutional signals, these social cues did not appear to influence judgments in this setting, for the reasons mentioned earlier in the article.

AI Feedback Sways What People Trust

The fourth and final experiment tested whether users would adjust their credibility and sharing judgments in response to AI-generated labels: *accurate*; *somewhat accurate*; *unverified*; or *inaccurate*, all assigned by ChatGPT.

Participants responded strongly to AI-generated credibility labels. Accuracy ratings rose or fell in line with ChatGPT's feedback, with the largest effects seen when headlines were labeled *accurate* or *inaccurate*:



ChatGPT feedback influenced both accuracy and shareability ratings. Top: Accuracy scores rose with more positive labels, especially when headlines were marked 'Accurate', and dropped when labeled 'Inaccurate'. Below: Shareability followed a similar pattern but showed greater variation by group: liberals responded most strongly to negative cues, while conservatives showed more muted shifts.

Political identity shaped these effects, with users trusting ChatGPT more when its feedback aligned with their own views.

Sharing behavior followed a similar pattern: ingroup headlines flagged as accurate were shared most often, especially under ambiguous labels like *somewhat accurate*.

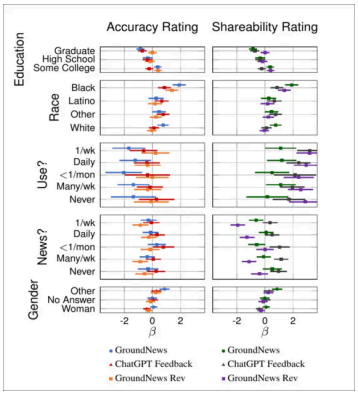
These results, the paper posits, suggest that AI feedback can shift user behavior; and that it also risks reinforcing partisan divides or discouraging critical thought.

Who Trusts AI the Most?

Follow-up analysis examined how user demographics shaped reactions to credibility labels. ChatGPT's feedback raised accuracy ratings overall, but the effect was weaker among highly-educated users, and frequent social media consumers, who showed more skepticism.

These same groups reacted negatively to GroundNews and Reversed cues, suggesting, the paper proposes, that overt bias markers may alienate more media-literate users.

By contrast, women and racially minoritized participants, especially Black and Latino users, responded more positively to ChatGPT than to institutional signals:



Demographic responses to feedback types, with each panel showing how a particular group responded to different credibility signals. ChatGPT's ratings had the strongest and most consistent influence on accuracy, while effects on

sharing were less uniform, with variation by race, gender, and media use.

Sharing behavior echoed this split: GroundNews reduced shareability most sharply among social media users and news junkies, while ChatGPT's effects were more mixed, even boosting shareability in some groups, with graduate degree holders especially responsive to *all* feedback types.

The authors conclude:

> *'These findings have direct implications for the design of credibility interventions in sociotechnical systems. Users are increasingly influenced by algorithmic feedback, which can override institutional cues and moderate partisan bias – but also risks promoting overreliance.*
>
> *'Institutional signals remain effective for some users, but their impact diminishes in politically polarized or [low-trust] environments. Meanwhile, engagement metrics such as likes and shares were largely ignored, suggesting reduced persuasive value when presented without social context.*
>
> *'To support equitable and informed news evaluation, AI-driven interventions must be transparent, explanatory, and designed to enhance user agency.*
>
> *'Future work should examine these mechanisms in more ecologically valid settings, evaluate alternative AI credibility framings, and develop adaptive systems that foster critical engagement across politically diverse audiences.'*

Conclusion

Given the tendency of all current generative AI systems to hallucinate and distort truth, it is arguably of some concern that the wildfire adoption (even if it is slowing down a little) of ChatGPT also represents an enormous leap of faith that the architectures of such systems can neither justify nor support.

One problem with trusting AI's representation of news is the lack of effective systems that can contextualize news sources as 'politically affiliated', or leaning towards one or other end of the political spectrum.

Even among the most reputable fourth-estate sources, the choice of what and what is not covered is in itself a political statement. Neither ChatGPT nor its stablemates is currently in any position to navigate these layers of interpretative bias, and the topic itself invites discussion rather than solid conclusions.

Another issue is that systems of this kind have arrived on the scene in one of the most polarized and divisive periods of human history in 80 years, and at a time when society is most willing to listen to 'alternative voices' – such as an entirely new genre of technology that is being touted as an essential filter of the world's truth, rather than what it actually is: a predictor of statistical probabilities, fed by high-volume quantities of partisan information.

* *Details of which the authors have made available online (see source paper, bottom of P2, for URLs). However, this data requires registration to view, and since I did not take the matter further at that point, I cannot confirm that it can be viewed entirely without payment and/or specific types of credential.*

First published Wednesday, November 5, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More