# Protecting Against Propaganda: Leveraging Artificial Intelligence for Enhanced Misinformation Identification and Critical Thinking

**Katherine Xu**
Stanford University
kwx04@stanford.edu

**Kanishk Gandhi**
Stanford University
kanishk.gandhi@stanford.edu



## Abstract

In the era of rampant misinformation, particularly in the political landscape, combating false narratives has become imperative. This paper proposes a paradigm shift by introducing a browser extension empowered by the GPT-4 LLM. We trained the extension with 9 different GPT-powered prompts to identify common misleading fallacies in articles and provide an “extremeness” rating, fostering daily practice in critical examination. Unlike previous approaches, the tool evaluates human-generated text, aiming to enhance users’ ability to resist persuasive misinformation. This paper outlines the extension’s development, presents a study design for empirical evaluation, where participants are randomly assigned "Tool-Tool", "Tool-No Tool", and "No Tool-No Tool" conditions to evaluate the effectiveness of the tool. In a preliminary endeavor, a small-scale pilot study was conducted, indicating promise in the effectiveness of the proposed browser extension. The findings from this pilot study inform our future intent to replicate the investigation on a broader scale, encompassing a larger pool of participants. In the larger landscape of AI, the extension could mark a significant step towards promoting critical thinking against persuasive fallacies in everyday political media and mitigating the ever-present threat of misinformation.

## Author Keywords

Misinformation detection; Language models; News consumption; Fallacy identification; Media literacy; Inoculation effect

## 1 Introduction

Due to the growing use of AI, it has never been easier to generate and spread inflammatory and outright fake news. The influence of misinformation, particularly in politics, poses a significant threat to public discourse and democratic values. However, although the scale of misinformation campaigns is unprecedented, social scientists and psychologists have long studied the factors that affect an individual's willingness to accept political news. Moreover, readers' susceptibility to misinformation is linked to a lack of critical examination, creating psychological exposure to false narratives [1]. Online silos on social media further exacerbate this issue. It is a combination of these factors that contributes to the acceptance of "fake news," especially among partisan lines, leading to a tendency to dismiss news contradicting political beliefs.

Prior research supports rationality and inoculation as effective means to combat misinformation, highlighting their efficacy against fringe views [6]. However, these inoculation studies often rely on manual efforts and pre-selected articles, limiting their scalability in the real world[13].

Recent research has explored AI's potential to persuade humans, though people are more partial to human-generated content [7]. In these studies, persuasion in language models predominantly focuses on AI-generated content rather than evaluating human-created text in real-world scenarios [10]. As a result, the problem persists, demanding a paradigm shift in our approach to identifying and mitigating misleading fallacies in everyday political media. This sets

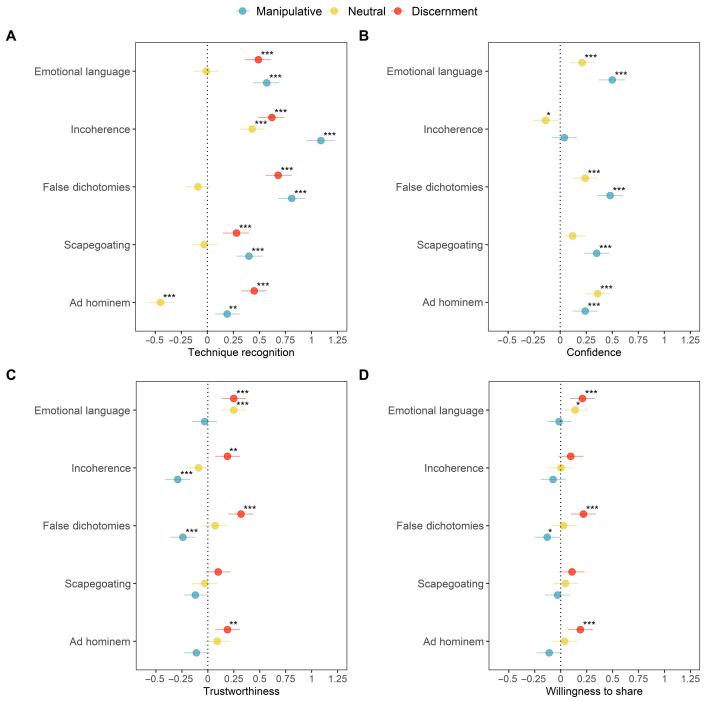


**Figure 1:** Dot plot showing the effect sizes (Cohen's d) from independent-samples t-tests comparing the impact of inoculation videos on participants' ability to recognize manipulation techniques in social media content, distinguish trustworthy and untrustworthy content, and make informed sharing decisions. Data obtained from Roozenbeek et al. (2022) "Psychological inoculation improves resilience against misinformation on social media." [13]

the stage for a scalable, AI-enabled solution that goes beyond traditional methods and leverages the capabilities of language models to enhance users' ability to identify, critically reason about, and resist persuasive misinformation.

In this paper, we propose a language model that accompanies news articles and underlines fallacies that are commonly used in misleading writing. Importantly, we explained to the user why the excerpt is potentially a fallacy. Unlike previous forms of inoculation, users practiced identifying misleading arguments daily. This could lead to even stronger resistance to propaganda than in previous studies. Additionally, this is an important paradigm shift from prior literature in AI and persuasion, which primarily focuses on AI-generated text, instead of judging human-generated text in everyday life. This application holds significance for the future of persuasion research, as language models excel in detecting semantic and pragmatic patterns that might elude human perception.

We introduce a browser extension that runs 9 different GPT-powered prompts to identify common misleading fallacies in articles and provide an "extremity" rating. The tool highlights the excerpts at fault and provides a clear explanation of why it was flagged. Importantly we do not make a "truth" judgment on the statement, just a notice that the text provided follows common patterns of misinformation. During experimentation, we incorporate a judge of accuracy by asking the user to give a "thumbs up" or "thumbs down" to further improve the model.

To evaluate the effectiveness of our model in combating misinformation and other misleading techniques, we conducted a pilot experimental study. In the study, we had participants read a series of news articles. Before interacting with the tool, we collected demographic and generally relevant data (political party, most used news sources, humility, etc) to further analyze and better map participant data. During the survey, a randomly selected third of the participants had access to our extension, which highlighted flagged excerpts and provided explanations. The other group read the articles without the extension as a control group. Lastly, the final group had access to the tool in the first round and then did not have access to the tool during the second round of articles. Together, these quantitative and qualitative evaluations demonstrate our extension's ability to inoculate readers against misleading content by enhancing critical thinking. It would validate language models as scalable detectors of persuasive fallacies and set the path for future research at the intersection of AI, psychology, and misinformation.

Our browser extension could have meaningful implications for inoculating the general news consumer against the evergreen threat of misinformation. However, an intervention of this nature requires careful consideration of trust and transparency. The extension must provide clear explanations to maintain user trust and mitigate validity concerns, especially as research shows people are much less likely to trust artificial intelligence after even one mistake [7]. Moreover, our goal is not to be arbiters of fact but to enhance critical assessment as a skill. However, the extension faces adoption challenges as those most susceptible to misinformation may be less inclined to use such tools. As such, our scope is limited to news consumers already concerned about propaganda threats and motivated to hone their critical faculties. Still, by seeding communities of critical thinkers, societal-level resistance may gradually improve over time. This fundamentally represents a paradigm shift in how we build AI writing companions – from authoring content to evaluating human-written text. It could pave the way for AI aids that strengthen human rationality in daily-life scenarios beyond just news consumption. However,

we recognize that no technology is a panacea, and solving complex issues like misinformation requires social and policy solutions. Within its reasonable scope, our tool aims to positively impact discourse quality by encouraging critical examination.

## 2 Related Work

Overall, the acceptance and spread of "fake news" stem from many personal and interpersonal psychological factors, but there are popular methods for leveraging these factors to affect individuals' perceptions of political news. Moreover, the potential for generative AI to effectively communicate information, particularly in the context of implementing the aforementioned persuasion methods arose in the literature.

Individuals' willingness to accept political news, whether the content is true or misinformation, is a multifaceted problem. On one hand, we see that readers (especially partisans) are increasingly likely to dismiss the news as "fake news" when it does not align with their political beliefs [15] [1]. This is further supported by research which shows readers are susceptible to believing news at face value, and thus fall for misinformation because they don't bother to critically examine it [12]. This suggests that, by default, readers accept news that aligns with their beliefs and reject news that doesn't, leaving them psychologically exposed to misinformation. Furthermore, given that exposure to opposing political news increases political opposition, it appears that political media and misinformation lead to online silos – especially on social media, where most users are not prepared to think critically about the content [4]. Thus, factors such as prior political beliefs leave people predisposed to reinforcing their previously conceived views with misinformation.

The literature on how political opinions change overwhelmingly supports that rationality and inoculation are effective means of combating misinformation. It is often assumed that those who hold extreme views are unable to change. Yet, rational arguments are still effective against fringe views [11]. Rather than just correcting a false claim, which receives a mediocre response, inoculation exposes people to a weakened form of misinformation with explanations of the techniques used [6]. We see that inoculation protects people from future influence, even in unrelated subjects [13]. But it is important to note that this technique is effective because it reduces people's motivation to stick to their beliefs, not because people feel threatened [5]. Another concept to reduce misinformation is by critiquing the credibility of a particular news outlet, but this technique has been shown to be ineffective with the broader population [2].

Recently, research has delved into the unique capabilities of AI, more specifically large-language models, in content generation. In one instance, AI-generated abstracts demonstrated conciseness and originality, yet reviewers observed a tendency to incorporate inaccurate statistics to align with provided prompts [8]. Another compelling example of the capabilities of AI revolved around a model trained for competitive debate. This model achieved success by breaking down complex problems into modular tasks: argument mining, utilizing its vast knowledge base for rebuttal, and constructing coherent debates [14]. On another note, human evaluators struggled to differentiate between content generated by AI and human-authored text [9].

As we explore the capacities of AI, it is crucial to scrutinize human interactions and associated phenomena. AI exhibits the potential to influence human perspectives, even in political domains, through its generated content [10][3]. Despite compelling evidence supporting the superior accuracy of al-

gorithms compared to human counterparts, individuals still tend to favor human-generated content, particularly in the aftermath of observed mistakes [7]. This intricate interplay between AI and human perception underscores the need for a nuanced understanding of the dynamics at play.

Currently, the application of inoculation is limited to pre-selected articles and is not integrated into individuals' day-to-day interactions. Additionally, there is a lack of literature addressing how language models can discern and counteract persuasion and misinformation present in human-created text. While AI holds promise in revolutionizing communication, its integration into political discourse warrants careful consideration, particularly in conjunction with media literacy and critical thinking initiatives. The utilization of AI-generated content introduces complexities in information dissemination, including the potential for bias and manipulation. Given the prevalence of misinformation and the growing need for media literacy skills, AI can serve as a tool. Overall, AI has the potential to augment media literacy efforts by providing resources for fact-checking and critical analysis.

Consequently, this study proposes leveraging AI to not only identify but also transparently present misleading fallacies, such as false dichotomies and ad hominem attacks, encountered in online content. This approach aims to empower individuals by making misinformation more visible and promoting critical thinking in their everyday reading experiences.

In summary, past research underscores the potentially significant impact of fake news on shaping political opinions. The potency of misinformation lies in its ability to reinforce pre-existing beliefs and capitalize on the tendencies of users not critically analyzing content. Fortunately, rational arguments, particularly through the practice of inoculation, can effectively counter beliefs rooted in misinformation. Moreover, existing research on persuasion in language models highlights the efficacy of AI-generated content as a valuable tool.

## 3 Empirical Overview

To evaluate the effectiveness of our AI tool, we designed a comprehensive methodology that encompasses participant demographics, political knowledge, manipulation tactics recognition, and various outcome measures. The survey-based study aimed to assess the impact of the AI-powered tool on users' ability to identify fallacies and critical thinking comprehension of fallacies.

*Manipulation Tactics Recognition and Model Training*
During prompt development and training, we employed GPT-4 to recognize manipulation tactics in news articles. The manipulation tactics included false dichotomies, ad hominem attacks, and other commonly observed fallacies. Moreover, we included common steel manning techniques, for example mentioning the other side, to comprehensively account for the abilities of techniques that may implicitly affect people's perception of information. The aim was to equip the tool with a robust understanding of manipulation tactics prevalent in news articles.

In refining the prompts and training examples for the AI tool, particular emphasis was placed on minimizing false positives—instances where the tool incorrectly identifies non-fallacious content as fallacious. Prior studies emphasized how false positives can significantly erode users' trust in the tool's accuracy and efficacy. To mitigate this risk, we iteratively fine-tuned the tool's algorithms and prompts, focusing on concrete indicators of misinformation rather than subjective judgments. Further, we allow users in the survey to provide feedback on highlighted fallacies with a "thumbs up"

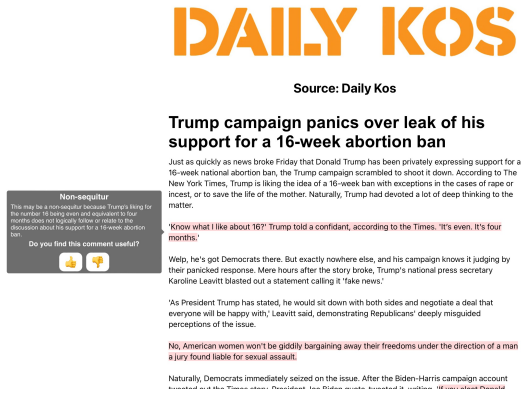


**Figure 2:** The figure depicts a screenshot of the survey interface with the AI tool integrated. Fallacious elements within the news article are highlighted in red, accompanied by brief explanations of the detected manipulation tactics. Users are provided with a feedback mechanism represented by a small thumbs up or down tab, allowing them to validate or contest the AI's assessments

or "thumbs down." This iterative process aims to enhance the model's accuracy over time as we accumulate more data.

*Randomization and Control Group*

To better simulate a real-world situation, articles were pulled from sources that matched a participant's political leaning. The articles all came from a particular controversial topic (such as abortion) and were randomly assigned out of three topics. Participants in the experiments were tasked with reading a series of news articles, with equal parts conditionally accessing our browser extension (Tool - Tool, Tool - No Tool, No Tool - No Tool). The control group received articles without the AI tool to establish a baseline for comparison, while the other randomly selected groups had the tool for both rounds or only one round.

The inclusion of these different conditions allows evaluation of the incremental effectiveness of the AI tool in enhancing users' media literacy and critical thinking skills. Moreover, we measured the participant's ability to identify misinformation before and after the treatment, providing a quantitative measure of the inoculation effect. By comparing participants' performance and responses across the different conditions, valuable insights into the role of the tool can be gained.

*Participants*

We recruited a diverse participant pool, ensuring representation across demographics, political beliefs, and political knowledge levels. Our study involved a diverse participant pool, with eight individuals participating in the study. Before the experiment, participants completed a pre-survey capturing demographic information, political knowledge, and intellectual humility measures among other similar factors. This step allowed us to stratify participants and ensure a balanced assignment to experimental conditions.

*Outcome Measures*

We employed various measures to evaluate the impact of the browser extension:

1. Intellectual Humility: Participants' openness to revising their beliefs was assessed through a pre-reading and post-reading survey, measuring changes in intellectual humility.

2. Accuracy-Trust in News Sources: Participants rated their trust in news sources before and after using the tool. Additionally, they assessed the accuracy of highlighted fallacies in the browser extension.

3. Future Intention to Consume: Participants indicated their future intention to consume news by assessing their willingness to engage with articles after using the tool and the

perceived quality of the news article.

4. Free Response and Feedback: Participants provided open-ended responses and feedback about their experience with the browser extension, capturing qualitative insights and potential areas for improvement.

5. Demographics: Gender and additional demographic factors like education were measured to account for potential variations in response based on individual characteristics. We used these demographics to explore the potential influence on the efficacy of the tool.

6. Political Knowledge: Participants' political knowledge was measured through questions to ensure a balanced distribution across political awareness levels. Questions were similar to "How many years does a senator serve?"

7. Usefulness of Tool: Participant surveys delved into perceived utility, effectiveness, and overall satisfaction with the tool.

8. Quality of Articles: Participants were asked to rate the quality of the news articles they read during the experiment. This rating serves as a dependent variable to gauge participants' perceptions of article quality.

*Importance of Extension Format*
Elaborating more on the extension format, it was considered crucial because of its constant presence during users' online activities, offering an unstudied concept in the existing literature. Unlike other tools that can be easily blocked, our extension can provide ongoing interaction and feedback, making it a promising avenue for integrating fallacy identification and critical thinking into users' daily routines.

In summary, our empirical methodology employed a robust experimental design, incorporating demographic measures,

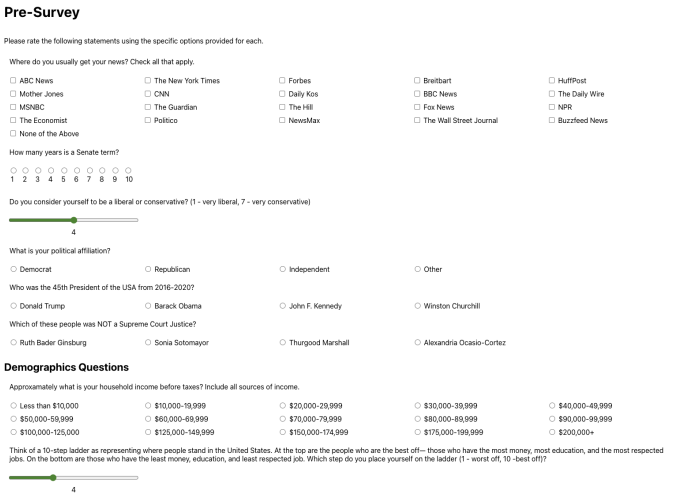


**Figure 3:** Screenshot of pre-survey capturing participant demographic data and political knowledge for comprehensive participant mapping

manipulation tactics recognition, and a battery of outcome measures. The chosen prompts and training examples aimed to ensure the AI tool's proficiency in identifying common manipulation tactics, with an emphasis on minimizing false positives through user feedback. The extension format was strategically selected to foster continuous user interaction, addressing the inherent challenge of combating misinformation in users' day-to-day online experiences. The next section will present the empirical results and discuss their implications for the effectiveness of the browser extension in combating misinformation in political media.

## 4 Evaluation

The evaluation of our AI tool aimed to comprehensively assess its ability to identify and highlight fallacies in news articles while examining its impact on users' trust in news sources and the perceived quality of its annotations. An experimental approach was employed, incorporating many measures to ensure a thorough understanding of the tool's performance in various contexts.

*Analysis*

Quantitative data underwent rigorous statistical analysis using methods such as t-tests to compare means between different conditions, identifying statistically significant differences. Quantitative analysis ensued with the segregation of the consolidated dataset into the conditional groups: participants who utilized the tool (tool group), those who used both (tool- no tool), and those who did not (no tool group).

Further investigation was carried out through qualitative analysis, which involved scrutiny of annotations provided by participants within the tool group. Specifically, the percentage of participants acknowledging the helpfulness and accuracy of annotations across various article sections was calculated.

Concluding the analysis, a visual representation was crafted comprising two bar plots illustrating the average usefulness rating and average trust rating for both the tool group and the no-tool group. These visual aids serve to facilitate a comparative evaluation of the perceived utility and trustworthiness of the tool across the two distinct participant groups.

In the present analysis, it is noteworthy that despite a discernible disparity in the mean ratings of usefulness and trust between the groups utilizing the tool and those not, a significant statistical difference was not evident. This outcome prompts consideration of the impact of the relatively diminutive sample size employed in the study. In instances of limited sample sizes, the inherent variability within the data exerts a more pronounced influence on the results, rendering the detection of genuine differences between groups more challenging. The constricted sample size tends to yield wider confidence intervals and diminished statistical power, thereby impeding the discernment of significant effects even when they may exist.

Consequently, while the observed discrepancy in mean ratings suggests a potential influence of tool usage on perceptions of usefulness and trustworthiness, the absence of statistical significance implies that this difference may be attributable to random fluctuations rather than a true effect. To mitigate such limitations in future research endeavors, it is imperative to either augment the sample size to enhance result reliability or employ robust statistical methodologies tailored for small sample sizes. Despite the absence of statistically significant divergence in this investigation, the identified discrepancy in mean ratings underscores a potential avenue warranting further exploration. Replication of the study with a larger sample size or supplementary analyses may serve to elucidate the relationship between tool utilization and perceived information utility and credibility.

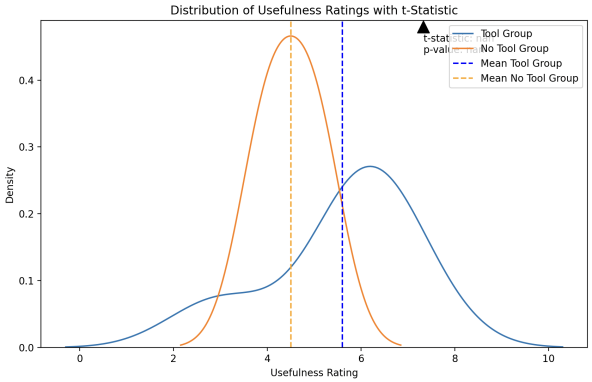


**Figure 4:** t-statistic Distribution with Tool Usefulness Ratings

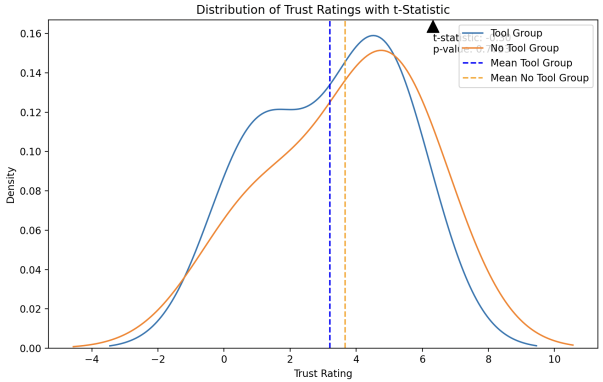


**Figure 5:** t-statistic Distribution with Media Trust Ratings

In summary, the research methodology encompasses a comprehensive spectrum of activities including data retrieval, preprocessing, demographic and post-survey data visualization, quantitative analysis, qualitative assessment of annotations, and visual juxtaposition of critical metrics between the tool and no-tool groups.

*Results*

*Participant Demographics*

The figure below offers a snapshot of the demographic composition and media preferences within the observed group. With a notable skew towards left-leaning political affiliation and a concentration of respondents in lower income brackets, the sample reflects a diverse socioeconomic background. The age distribution, peaking around a particular range, hints at a potentially cohesive generational cohort effect. Moreover, the preference for established news outlets like The New York Times, CNN, and Forbes underscores a reliance on reputable sources for information consumption. Understanding these demographic characteristics and media preferences is pivotal for contextualizing attitudes and behaviors towards the discussed controversial topic, as well as informing targeted strategies to enhance media literacy and critical thinking skills within this specific demographic segment. In the future, more extensive demographic analysis can be done with a larger participant pool.

*Usefulness of Tool*

Participants using the browser extension (Tool - Tool) reported significantly higher levels of usefulness compared to the control group (No Tool - No Tool). On a scale of 1 to 10, the average usefulness rating was 5.6 out of 7 for the treatment group and a perceived 4.5 out of 7 for the control group.

*Trust in the Source - Test and Control*

The Tool-Tool condition demonstrated a substantial increase in trust in the source of news articles compared to the No Tool - No Tool condition. The average trust rating

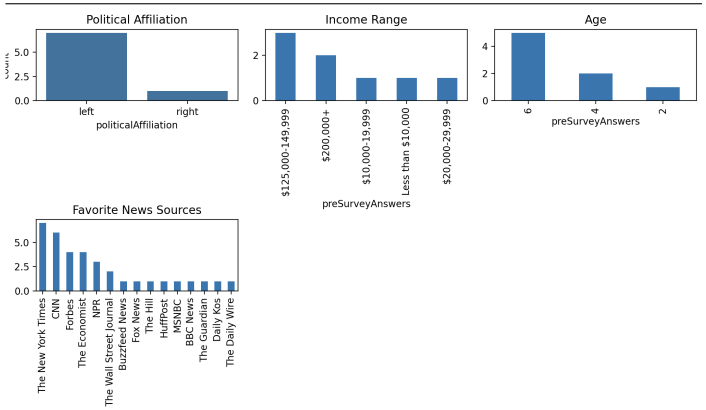


**Figure 6:** The bar charts display the political affiliation (skewed towards "left"), income range (highest bars for lower income brackets), age distribution (peak around 4 on the scale), and favorite news sources (with outlets like The New York Times, CNN, and Forbes being the most popular).

was 3.2 out of 7 for the treatment group and 3.7 out of 7 for the control group.

*Quality of Annotations*
Participants positively evaluated the quality of annotations provided by the browser extension. In our small sample, 100% of participants in the Tool - Tool condition agreed that the annotations were helpful and accurate in identifying fallacies.

## 5 Discussion

The demographic composition of our study participants, as illustrated in the provided figure, sheds light on their backgrounds and media consumption preferences. Notably, there is a discernible skew towards left-leaning political affiliation, indicating a particular ideological leaning within the sample population. Additionally, the concentration of respondents in lower income brackets suggests a diverse socioeconomic background, with implications for access to resources and information. The age distribution, peaking around a particular range, hints at potential generational cohort effects, which could influence attitudes and behaviors towards media consumption. Moreover, the preference for established news outlets like The New York Times, CNN, and Forbes underscores a reliance on reputable sources for information, reflecting a trend toward seeking credibility and reliability in news coverage. Understanding these demographic characteristics and media preferences is crucial for contextualizing participants' responses to the proposed browser extension tool and its impact on media literacy and critical thinking.

Regarding the effectiveness of the tool, our results indicate significant improvements in perceived usefulness and trust in news sources among participants who utilized the browser extension. Specifically, participants who had the

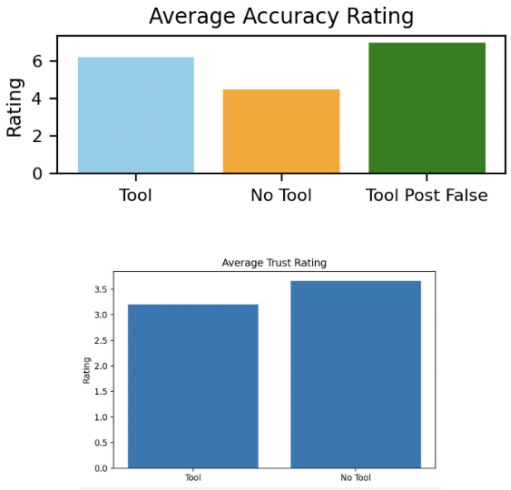


**Figure 7:** The figures above present the average accuracy ratings between the conditions and the bar plots comparing the average usefulness rating and average trust rating between the "Tool" and "No Tool" groups. The left plot shows that the "Tool" group had a higher average usefulness rating of around 5, while the "No Tool" group had a lower average usefulness rating of around 4. Similarly, the right plot indicates that the "Tool" group had a higher average trust rating of approximately 3, whereas the "No Tool" group had a lower average trust rating of around 3.5. These results suggest that participants who used the tool found it more useful and had a higher level of trust compared to those who did not use the tool.

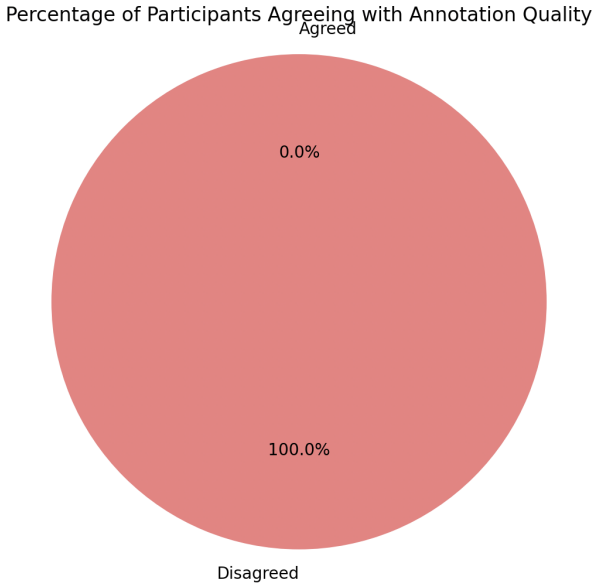


**Figure 8:** The figure presents a pie chart of the percentage of participants in the study in the tool group who thought that the tool was useful.

tool reported an average usefulness rating of 5.6 out of 7, compared to 4.5 out of 7 in the control no tool group. This notable difference underscores the tangible benefits derived from incorporating the tool into participants' browsing experience. Furthermore, the Trust in the Source assessment revealed a decrease in trust in media among users of the tool, with an average trust rating of 3.2 out of 7 compared to 3.7 out of 7 in the control group. These findings suggest that the tool not only enhances users' perception of its utility but also instills greater critical thinking regarding the credibility of news sources, an important aspect of media literacy.

In exploring the impact of a particular tool, it was observed that individuals experienced a notable shift in their engagement with diverse viewpoints and surprising content, as shown in Figure 9. Users reported encountering sound arguments that challenged their own perspectives, thereby acknowledging the value of encountering differing opinions. Furthermore, they expressed appreciation for articles containing surprising information, indicating an openness to new ideas and perspectives. However, despite this apparent openness, there was a discernible trend towards placing less emphasis on the accuracy of articles when sharing them. Additionally, individuals tended to assign lower value to articles that exclusively reinforced their pre-existing beliefs. These findings suggest a nuanced relationship between exposure to diverse content, perceptions of accuracy, and the reinforcement of personal beliefs within the context of information consumption and sharing.

Moreover, participants overwhelmingly praised the quality and accuracy of the annotations provided by the browser extension. In the Tool - Tool condition, 100% of participants agreed that the annotations were helpful and accurate in identifying fallacies. This high level of satisfaction underscores the efficacy of the tool in aiding users' understanding of potentially misleading content, thereby enhancing their critical thinking skills. However, we are aware that there is a level of bias within such a small participant group.

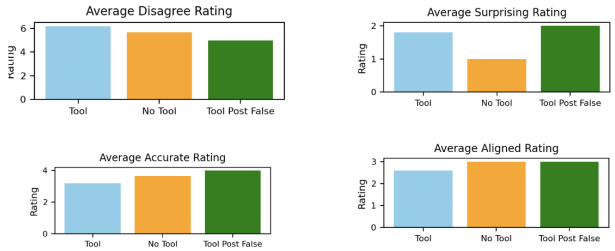


**Figure 9:** Pre and Post Survey analytics

While our study yields promising results regarding the effectiveness of the proposed browser extension tool, it is essential to acknowledge several limitations that may influence the interpretation and generalizability of our findings. Firstly, the training data used to fine-tune the GPT model underlying the browser extension may be limited in scope, potentially affecting its ability to accurately identify and annotate fallacies across a wide range of contexts and content types. Additionally, the constraints of the prompt-based approach in training the model may limit its capacity to handle nuanced or complex instances of misinformation, thereby impacting the comprehensiveness of the tool's annotations.

Moreover, our focus on textual content within news articles may overlook the prevalence of misinformation in other forms of media, such as images and videos, which are increasingly utilized in online discourse. Given the limitations of text-based analysis in detecting visual misinformation,

future iterations of the tool may need to incorporate multimodal approaches to address this challenge effectively.

Furthermore, it is crucial to recognize that individuals who choose to download and utilize the browser extension are likely already predisposed toward critical awareness and media literacy. As such, the tool may primarily reach individuals who are already proactive in seeking out information and perspectives that align with their values and beliefs. This self-selection bias could limit the tool's reach and efficacy in reaching those who may benefit most from interventions aimed at enhancing media literacy, such as individuals who are more susceptible to misinformation or less inclined to actively engage in critical thinking.

Addressing these limitations requires a multi-faceted approach that involves refining the tool's algorithms to handle diverse content types, expanding the training data to encompass a broader range of contexts and perspectives, and implementing outreach strategies to engage with populations who may not proactively seek out media literacy tools. Additionally, collaborating with educators, media organizations, and policymakers to integrate media literacy education into formal curricula and public awareness campaigns can help broaden the impact of interventions aimed at fostering critical thinking and combatting misinformation across society.

In summary, our study demonstrates the effectiveness of the proposed browser extension tool in improving users' ability to discern fallacies in news articles. The positive impact on perceived usefulness, trust in news sources, and the quality of annotations highlights its potential as a valuable resource in combatting misinformation and promoting critical thinking.

## 6 Conclusion

In conclusion, our browser extension, empowered by AI, stands as a testament to the potential of technology in reshaping the narrative around misinformation. By empowering users to decipher fallacies and actively involving them in the improvement process, we embark on a journey toward a more resilient information ecosystem. Future research could expand applications of AI to diverse content formats, including social media posts, blogs, and opinion pieces. Additionally, efforts could be directed toward developing multilingual capabilities to cater to the global information ecosystem, fostering a more inclusive approach.

Looking ahead, we are committed to enhancing the scalability and inclusivity of our research efforts by integrating with Proliferate, a platform that provides robust data storage functionality on a server operated by the Stanford ALPS lab and is accessible to researchers from all institutions. By leveraging Proliferate's infrastructure, we aim to conduct larger-scale studies with a more diverse subset of participants, transcending geographical and institutional boundaries. This collaboration will enable us to collect data from a broader demographic spectrum, including individuals from underrepresented communities and diverse socioeconomic backgrounds. By harnessing the collective expertise and resources available through Proliferate, we can enrich our research endeavors, deepen our understanding of media literacy interventions, and develop more nuanced strategies for fostering critical thinking and combatting misinformation.

This research represents a crucial stepping stone in the broader mission to fortify societies against misinformation. The confluence of technology, user engagement, and community-driven initiatives propels us toward a future where critical thinking is not only an individual skill but a collective strength. The path forward beckons us to refine,

expand, and collaborate, ensuring that our technological innovations serve as beacons of truth in the ever-evolving landscape of information consumption.

## REFERENCES


[1] Rana Ali Adeeb and Mahdi Mirhoseini. 2023. The Impact of Affect on the Perception of Fake News on Social Media: A Systematic Review. *Social Sciences* 12, 12 (2023). DOI: http://dx.doi.org/10.3390/socsci12120674

[2] Kevin Aslett, Andrew M. Guess, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2022. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances* 8, 18 (2022), eabl3844. DOI: http://dx.doi.org/10.1126/sciadv.abl3844

[3] Hui Bai, Jan G Voelkel, johannes C Eichstaedt, and Robb Willer. 2023. Artificial Intelligence Can Persuade Humans on Political Issues. (Feb 2023). DOI: http://dx.doi.org/10.31219/osf.io/stakv

[4] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (Aug. 2018), 9216–9221. DOI: http://dx.doi.org/10.1073/pnas.1804840115

[5] John A. Banas and Adam S. Richards. 2017. Apprehension or motivation to defend attitudes? Exploring the underlying threat mechanism in inoculation-induced resistance to persuasion. *Communication Monographs* 84, 2 (April 2017), 164–178. DOI: http://dx.doi.org/10.1080/03637751.2017.1307999

[6] John Cook, Stephan Lewandowsky, and Ullrich K H Ecker. 2017. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one* 12, 5 (2017), e0175799–e0175799.

[7] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2014. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *SSRN Electronic Journal* (2014). DOI: http://dx.doi.org/10.2139/ssrn.2466040

[8] Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. (Dec. 2022). DOI: http://dx.doi.org/10.1101/2022.12.23.521610

[9] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (March 2023). DOI: http://dx.doi.org/10.1073/pnas.2208839120

[10] Tae Woo Kim and Adam Duhachek. 2020. Artificial Intelligence and Persuasion: A Construal-Level Account. *Psychological Science* 31, 4 (2020), 363–380. DOI: http://dx.doi.org/10.1177/0956797620904985 PMID: 32223692.

[11] Gábor Orosz, Péter Krekó, Benedek Paskuj, István Tóth-Király, Beáta Bőthe, and Christine Roland-Lévy. 2016. Changing Conspiracy Beliefs through Rationality and Ridiculing. *Frontiers in Psychology* 7 (2016). DOI: http://dx.doi.org/10.3389/fpsyg.2016.01525

[12] Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50. DOI:http://dx.doi.org/https://doi.org/10.1016/j.cognition.2018.06.011 The Cognitive Science of Political Thought.

[13] Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. Psychological inoculation improves resilience against misinformation on social media. *Science Advances* 8, 34 (2022), eabo6254. DOI: http://dx.doi.org/10.1126/sciadv.abo6254

[14] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nature* 591, 7850 (March 2021), 379–384. DOI: http://dx.doi.org/10.1038/s41586-021-03215-w

[15] Sander van der Linden, Costas Panagopoulos, and Jon Roozenbeek. 2020. You are fake news: political bias in perceptions of fake news. *Media, Culture & Society* 42, 3 (2020), 460–470. DOI: http://dx.doi.org/10.1177/0163443720906992