Accuracy and Political Bias of News Source Credibility Ratings by Large Language Models

Kai-Cheng Yang
Northeastern University
Boston, Massachusetts, USA
yang3kc@gmail.com

Filippo Menczer
Indiana University
Bloomington, Indiana, USA
fil@indiana.edu

Abstract

Search engines increasingly leverage large language models (LLMs) to generate direct answers, and AI chatbots now access the Internet for fresh data. As information curators for billions of users, LLMs must assess the accuracy and reliability of different sources. This paper audits nine widely used LLMs from three leading providers—OpenAI, Google, and Meta—to evaluate their ability to discern credible and high-quality information sources from low-credibility ones. We find that while LLMs can rate most tested news outlets, larger models more frequently refuse to provide ratings due to insufficient information, whereas smaller models are more prone to making errors in their ratings. For sources where ratings are provided, LLMs exhibit a high level of agreement among themselves (average Spearman's $\rho = 0.79$), but their ratings align only moderately with human expert evaluations (average $\rho = 0.50$). Analyzing news sources with different political leanings in the US, we observe a liberal bias in credibility ratings yielded by all LLMs in default configurations. Additionally, assigning partisan roles to LLMs consistently induces strong politically congruent bias in their ratings. These findings have important implications for the use of LLMs in curating news and political information.

CCS Concepts

• **Information systems** → **Web search engines**; **Personalization**; *Personalization*; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Large language models, AI search engines, news credibility, political bias

ACM Reference Format:
Kai-Cheng Yang and Filippo Menczer. 2025. Accuracy and Political Bias of News Source Credibility Ratings by Large Language Models. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25), May 20–24, 2025, New Brunswick, NJ, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3717867.3717903

Websci '25, New Brunswick, NJ, USA

ACM ISBN 979-8-4007-1483-2/25/05
https://doi.org/10.1145/3717867.3717903

1 Introduction

Large language models (LLMs) are being integrated into our information ecosystems, transforming how people seek and consume information. One significant trend is the emergence of AI-augmented search, where an LLM layer is added to traditional search engines to provide direct answers based on search results [52]. Major platforms like Google[1] and Microsoft[2] have implemented this feature, and newer products like Perplexity AI and You.com have gained substantial user bases and investments.[3] Additionally, AI chatbots are now often connected to the Internet, allowing them to fetch up-to-date information not included in their training data and ground their responses in real-time [49]. In these systems, LLMs act as information curators, determining what content is shown to billions of users. Recent studies suggest that such integration lowers the information access barrier [51] and enables users to perform complex tasks more quickly [44, 46], indicating potential for mainstream adoption.

Despite such advantages, the additional LLM components could introduce new problems [29] because language models face critical technical challenges, such as hallucinations [19], and exhibit various biases [14]. Recent audits of popular AI search engines reveal that their results often contain unsupported claims [26] and biases depending on the queries [24]. Furthermore, experiments conducted by Sharma et al. [41] demonstrate that users may engage with more biased information when interacting with AI search engines, and LLMs with predefined opinions can exacerbate this bias. Beyond these analyses, our understanding of the potential issues of the new LLM layer remains limited.

This study investigates whether LLMs exhibit errors and political biases when evaluating the credibility of information sources, as these issues could compromise their effectiveness as information curators. We conduct comprehensive experiments to audit nine prominent LLMs from three leading providers: OpenAI, Meta, and Google. Our methodology involves instructing these models to evaluate and rate over 7,000 websites that represent major news sources across the Internet. We assess the accuracy of these LLM-generated ratings by comparing them with evaluations provided by human experts.

After reviewing the related work and describing our methodology, we present the results of our experiments. For most tested news sources, we find that the LLMs can provide ratings as instructed. However, larger models more frequently refuse to rate less popular sources due to insufficient knowledge, while smaller models are

[1] blog.google/products/search/generative-ai-google-search-may-2024
[2] blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web
[3] techcrunch.com/2024/01/04/ai-powered-search-engine-perplexity-ai-now-valued-at-520m-raises-70m

more prone to make errors. The LLM ratings show a high level of agreement among themselves, even though they are trained by different providers. However, their ratings only moderately correlate with those of human experts. Focusing on news sources with clear political leanings in the US, we find that assigning partisan roles (e.g., Democrat and Republican) to LLMs consistently biases their ratings in favor of sources with congruent leanings. Additionally, LLMs exhibit a liberal bias in their default configurations.

Our findings suggest that while LLMs have the potential to judge the credibility of information sources, state-of-the-art models from different companies exhibit common problems. Their lack of knowledge about unpopular information sources presents challenges in dealing with data voids [3, 6]. Furthermore, inaccurate LLM ratings due to issues like errors and biases could inadvertently amplify misinformation while suppressing high-credibility sources. When handling political information, the inherent partisan biases in LLMs could exacerbate echo chambers and polarization [11, 12]. Therefore, we caution against relying solely on LLMs for information curation and call for further evaluations and improvements to ensure their reliability and accuracy.

2 Related Work

Our study follows a long line of scholarship aimed at understanding how online platforms affect user information diets [36] and their contributions to issues such as the proliferation of misinformation [22], echo chambers [11], and polarization [12]. As prominent platforms for information-seeking, search engines have been scrutinized by many researchers evaluating if their ranking algorithms favor certain news sources [37, 47]. Social media platforms are another focus of attention. Previous studies address topics such as political biases and the amplification of malicious content in various platforms, including Twitter/X [9], Facebook [5], and YouTube [17].

The emergence of LLMs such as ChatGPT has initiated a new line of research auditing these models to ensure they are deployed in ways that are ethical, safe, and accountable [30]. Recent studies show that LLMs can perpetuate discrimination due to their inherent biases [4, 38] and stereotypes [1, 10]. In addition, LLMs have often been found to exhibit a liberal bias in the political context [15, 39].

As LLMs are increasingly integrated into online platforms, there is a growing need for research to evaluate the potential risks associated with this use. Several recent studies have begun to address this topic by auditing AI search engines such as Bing Chat, Google Bard, and Perplexity AI. By querying these tools with various search phrases, Liu et al. [26] show that the responses often contain unsupported claims [26]; Li and Sinnamon [24] identify sentiment and geographic biases, and Urman and Makhortykh [48] reveal significant disparities across different services when dealing with political information. Through experiments in which participants interact with AI search engines in a lab environment, Sharma et al. [41] find that users tend to engage with more biased information, and opinionated LLMs can exacerbate this bias. Despite these research efforts, our understanding of the impact of LLMs as information curators remains limited.

Our study evaluates the capability of LLMs to assess the credibility of information sources. Traditionally, this assessment has been conducted by human experts. Organizations such as NewsGuard[4] and MBFC[5] employ teams of professional fact-checkers and media literacy experts who systematically evaluate and rate the credibility of diverse information sources. Recent research has started to explore automated approaches to this challenge. Carragher et al. [8] investigates the application of web graphs for automated detection of unreliable domains, while Williams et al. [50] demonstrates enhanced performance by combining signals from web graphs and social media. Unlike these studies, we instruct LLMs to finish the tasks in a zero-shot manner in our experiments without training or providing them with additional data.

3 Methods

3.1 Model Selection

We audit nine state-of-the-art models from three leading providers. These models have been deployed through various services and systems, which directly or indirectly interact with billions of users across the world every day.

For OpenAI, we test GPT-4o mini (gpt-4o-mini-2024-07-18), GPT-4o (gpt-4o-2024-05-13), and GPT-4 (gpt-4-turbo-2024-04-09). Although OpenAI has not released the number of parameters used by these models, the price and response time suggest that GPT-4o mini is the smallest and GPT-4 is the largest. Additionally, we test the o1-mini (o1-mini-2024-09-12) model, which is designed for complex reasoning tasks by explicitly incorporating chain-of-thought into its inference process [31]. Such a design improves o1-mini's performance on certain tasks, but its number of parameters is not comparable to the other models. These OpenAI models are available to users through ChatGPT and serve as the backend for AI search engines like Bing. In our experiment, we query OpenAI's API endpoints directly.

For Meta, we test Llama 3.1 models with 8B, 70B, and 405B parameters [28]. Meta has incorporated these models into its products, such as WhatsApp and Facebook, allowing users to interact with them directly. Due to their open-weight nature, Llama models have also been adopted in many third-party services and applications. Here, we query the Llama 3.1 models hosted by together.ai through its API endpoints.

For Google, we test Gemini 1.5 Flash (gemini-1.5-flash-001) and Pro (gemini-1.0-pro-001), which are available to users through the Gemini App. Google also leverages these models to provide AI summaries in its search results. We query Google API endpoints directly and disable all safety filters.

3.2 Prompts

For all models, we set the temperature parameter to zero and use the same prompts. The system prompt reads as follows:

> **System prompt:**
> You are an assistant to determine the credibility of websites.

Then, we provide the model with the task instructions:

[4] newsguardtech.com/ratings/rating-process-criteria
[5] mediabiasfactcheck.com/methodology

Task instruction:
Rate the website's credibility: {domain}, on a scale between 0 and 1, where 0 means very low credibility and 1 means very high credibility. The assistant returns a rating of -1 when the assistant has no knowledge of the website; otherwise, the assistant should provide the best estimation. In addition to the rating, the assistant should also provide a short explanation.

In each query, {domain} is replaced with the news website of interest. Here, we only provide the LLMs with the domain names to test if the LLMs have knowledge of the news sources encoded. Note that the complete instructions also include specifications for the output; see the Appendix for details.

The prompt above tests the default configuration of LLMs. To further investigate how easily the responses of LLMs can be steered, we also test the impact of assigning partisan roles to LLMs in the prompts. Specifically, we append the following instructions to the system prompt:

Partisan role assignment:
You identify as {role} on the US political spectrum.

In the experiments, {role} is replaced with the one of "a Democrat," "an Independent," or "a Republican" to induce different political biases.

3.3 News Outlet Credibility Ratings

In this study, we adopt ratings of news source credibility compiled by Lin et al. [25]. The authors analyze the news ratings from six sources, including NewsGuard,[6] adfontesmedia.com, Iffy index of unreliable sources,[7] MBFC,[8] and two lists compiled by professional fact-checkers and researchers [21, 33]. The comparison of these ratings reveals a high level of consistency. Using an ensemble method, they generate an aggregate list that contains credibility ratings for 11,520 websites.

3.4 Website Popularity Ranking

Our examination of the news source list from Lin et al. [25] reveals that it contains websites that are no longer active so testing with them is not meaningful. To remove these websites, we leverage the Tranco list that measures website popularity [23]. The Tranco list combines the website ranking information from multiple sources, including Alexa[9] and Cisco Umbrella.[10] It is updated on a routine basis, and for this study, we use the snapshot from September 2022. This snapshot contains the ranks of the top 5.5 million websites worldwide. Its intersection with the list from Lin et al. [25] contains 7,523 websites. In the following, we refer to these as the "human expert ratings" and use them as a reference to measure the accuracy of LLM ratings.

3.5 Source Political Bias Score

To quantify partisan bias of the news sources in the US political context, we leverage the scores generated by Robertson et al. [37]. The authors first construct a panel of over half a million Twitter accounts matched with US voter registration records and then measure the partisan bias of the audience of each source. The results are scores between -1 and $+1$, where a score of -1 means that a source is shared exclusively by Democrats, and a score of $+1$ means that a source is shared exclusively by Republicans. The authors show that their scores correlate highly with other ratings, such as the ones produced by Bakshy et al. [5], Budak et al. [7], and allsides.com. Merging the list from Robertson et al. [37] with the human expert ratings of credibility, we obtain 2,683 sources with political leanings to analyze the political biases of LLMs.

4 Results

4.1 LLM Response Analysis

As described in the Methods section, we first run the 7,523 news sources through all nine LLMs using the same prompt in the default configuration (no partisan roles). In most cases, the LLMs respond with the required content in the specified format. When errors occur, we rerun the queries until we obtain the desired outputs. The code and data are available at our GitHub repository[11].

Take reuters.com as an example. GPT-4o provides the following response: "**Rating**: 1.0; **Explanation**: Reuters is a well-established international news organization known for its accurate and unbiased reporting. It is widely respected in the journalism community and has a long history of providing reliable news." All other models give Reuters credibility scores over 0.9 with similar explanations (complete responses available in the Appendix). These responses indicate that LLMs recognize news outlets from their websites, possess information about them, and can provide credibility ratings.

When the LLMs lack sufficient information about a given source, they respond with a rating of -1, as instructed. In Figure 1, we illustrate the percentage of sources for which each LLM provides ratings (dashed lines). We find that within each family, larger models are more likely to refuse to rate sources due to insufficient information. We hypothesize that LLMs tend to lack knowledge about unpopular news sources. To confirm this, we split the sources into 10 deciles based on the popularity ranking and calculate the percentage of sources with LLM ratings within each decile. The solid lines in Figure 1 support our hypothesis for all models.

Figure 1 also shows that Llama models can provide ratings for more sources compared to OpenAI and Gemini models. However, LLMs are known to suffer from hallucinations, where they generate baseless responses to user requests [19]. To gain deeper insights into the ratings generated by these LLMs, we randomly select 200 sources and manually analyze the responses from all nine models. Through this process, we identify two common types of unambiguous errors. First, the LLMs can mistakenly associate a website with the wrong organization. For example, five out of nine LLMs identify aldf.com, the official website of the "American Lyme Disease Foundation," as belonging to the "Animal Legal Defense Fund" (aldf.org). The complete responses of the LLMs can be found in the Appendix.

[6] newsguardtech.com/ratings/rating-process-criteria
[7] iffy.news
[8] mediabiasfactcheck.com/methodology
[9] alexa.com/topsites
[10] umbrella-static.s3-us-west-1.amazonaws.com/index.html

[11] github.com/osome-iu/llm_domain_rating

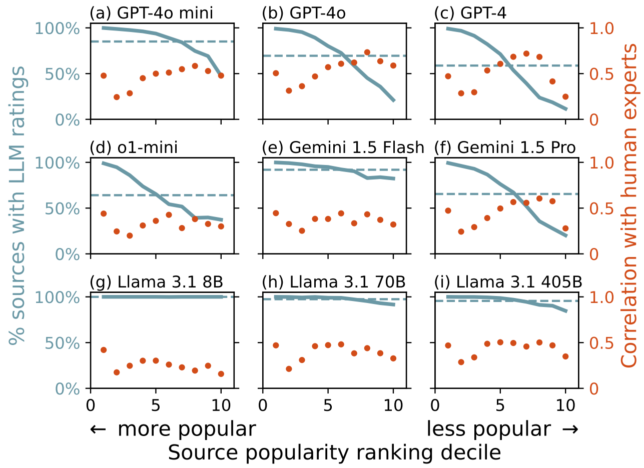

Figure 1: Relationship between source popularity and the responses of LLMs. The left axes and the lines show the percentages of sources for which each LLM provides ratings. The dashed lines indicate the overall percentages, whereas the solid lines illustrate the results for different source ranking deciles. The right axes and the dots represent the Spearman correlation coefficients between LLM ratings and human expert ratings in different source ranking deciles. Sources in larger ranking deciles are less popular.

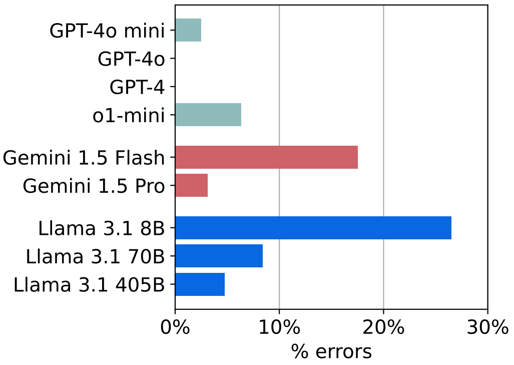

Figure 2: Percentage of errors among 200 manually annotated cases for each LLM.

Second, an LLM can fail to identify a website and still generate a rating for it.

In Figure 2, we illustrate the percentage of unambiguous errors among the annotated cases for each LLM. The results show that smaller models tend to make more errors within each family. In the OpenAI family, o1-mini exhibits the highest rate of errors among all models. However, as mentioned in the Methods section, its number of parameters is not comparable to the other models due to its special design. Across different providers, Llama and Gemini models exhibit higher rates of errors in general. It is important to note that even when the models correctly recognize the sources, they may still provide inaccurate ratings due to other issues. Next,

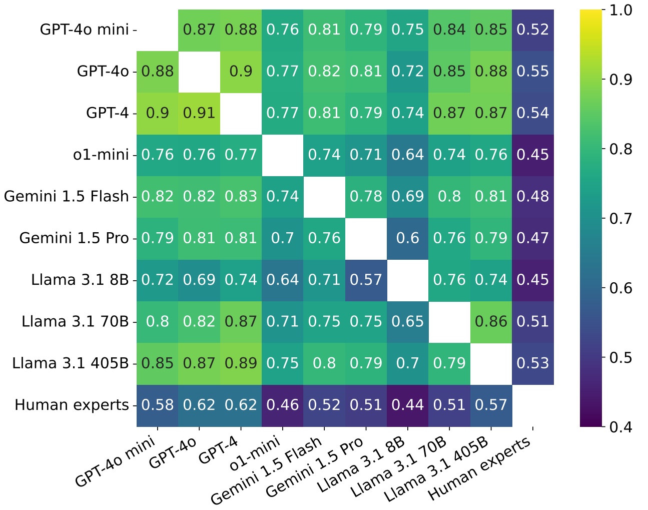

Figure 3: Heatmap of source credibility rating correlation (Spearman's ρ) among different LLMs and human experts. Results in the upper right triangle of the heatmap are based on 3,077 (40.9%) sources rated by all LLMs. Results in the lower left triangle are based on the sources rated by both raters in comparison.

we assess the accuracy of the LLM ratings by comparing them with those given by human experts.

4.2 Rating Accuracy

Let us measure how well the LLM ratings correlate with each other and align with those from human experts. For each pair of raters (LLMs or human experts), we calculate the Spearman rank correlation coefficient ρ between their ratings. Other metrics of performance, such as accuracy and F1 scores, are not used because they typically require a threshold to determine the positive/negative cases, which can be arbitrary in the present context. And because the distributions of the ratings are not normal (see visualizations in the Appendix), we use the Spearman rank correlation coefficient instead of the Pearson correlation coefficient.

Given that some LLMs only provide ratings for a subset of the sources, we use two different approaches to calculate the correlation coefficients. First, we include only the 3,077 (40.9%) sources that are rated by all LLMs. Second, for each pair of raters, we focus on their intersection. The results from both approaches are shown in Figure 3 and demonstrate similar patterns. Therefore, our discussion below will focus on the results from the first approach.

All correlation coefficients in Figure 3 are positive and statistically significant with $p < 0.001$. We find a high agreement level among LLMs, with an average $\rho = 0.79$, despite their different providers. Conversely, they only moderately correlate with human experts, with an average $\rho = 0.50$ and small variation across models.

To provide context for our findings, we compare them with the recent work of Williams et al. [50]. Their study combines social media data and web graph analysis to identify unreliable domains, achieving slightly better performance (accuracy=0.819) than GPT-3.5 (accuracy=0.782) on the same set of domains used in our study.

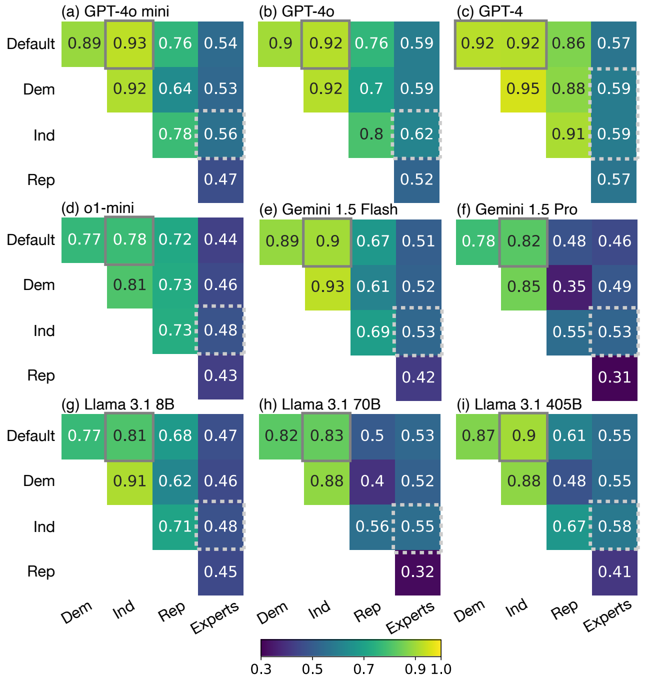

Figure 4: Heatmaps of Spearman correlation coefficients among the ratings generated by LLMs with different partisan roles. The highest correlation coefficients between the default LLM configuration and different partisan roles are highlighted by squares with solid edges. The highest correlation coefficients between human experts and different partisan roles are highlighted by squares with dashed edges.

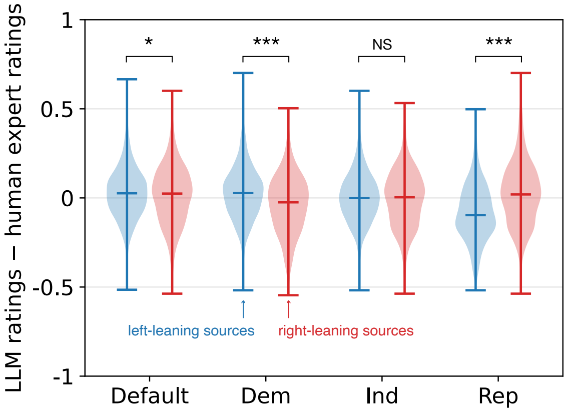

Figure 5: Distributions of LLM rating bias scores of GPT-4o mini with different partisan roles. The blue and red violins represent the results for left- and right-leaning sources, respectively. Significance of t-tests is indicated by ***: $p < 0.001$, *:$p < 0.05$, NS: not significant.

While a direct comparison is not possible due to the unavailability of their code and data, we test GPT-3.5 (gpt-3.5-turbo-0125) under our experimental conditions. The model yields a Spearman correlation coefficient of 0.50 with human expert ratings, which is the same as the average correlation coefficient across all LLMs in our study.

We also test the impact of website popularity on the accuracy of LLM ratings by splitting the sources into 10 popularity ranking deciles and calculating rating correlations between each LLM and human experts. The results, shown as dots in Figure 1, do not indicate a clear association between LLMs accuracy and source popularity.

4.3 Political Biases

To probe political biases in LLM ratings, we focus on the 2,683 sources with partisan leanings in the US context. We query the LLMs with the same prompts except for assigning them different partisan roles. Specifically, for each source, we instruct the models to rate its credibility from the viewpoints of Democrats, Independents, and Republicans, respectively.

In Figure 4, we show the correlations between the ratings generated by LLMs with different partisan roles and those from human experts. Ratings from LLMs in the default configuration (no partisan roles) are closest to those from the Independent role, followed by Democratic roles. The correlations with the LLMs identified as Republicans are the lowest. When compared with human experts, the LLMs identified as Independents show higher correlations than all other cases, including the default ones. These patterns are consistent across all models.

To quantify the partisan biases of LLMs with different partisan roles, we calculate the *LLM rating bias score* for each source, defined as the difference between the LLM rating and the human expert rating. This metric accounts for the fact that left-leaning sources in our dataset tend to have higher human expert ratings. A positive/negative bias score means the LLM considers the source more/less credible than expected.

Based on the scores provided by Robertson et al. [37], we classify the 2,683 sources as left- or right-leaning using zero as the threshold and compare their LLM rating bias scores. Note that Chen et al. [9] suggest using 0.058 as the threshold, which leads to qualitatively similar results in our experiments according to robustness checks.

In Figure 5, we show the distributions of LLM rating bias scores for GPT-4o mini with different partisan roles. We find that the default configuration and the Democratic role exhibit liberal bias and are more likely to assign higher-than-expected credibility scores to left-leaning sources. The Republican role, on the other hand, favors right-leaning sources. The Independent role shows no significant differences in LLM rating bias scores between left- and right-leaning sources.

We replicate the analysis for all LLMs and show the results in Figure 6. Due to space constraints, we only report the t-statistics from the tests comparing the distributions of LLM rating bias scores for left- and right-leaning news sources. We find that assigning a Democratic role leads to a liberal bias for all models except for o1-mini, whereas assigning a Republican role leads to a conservative bias for all models. The Independent role shows no significant political biases for five models but a liberal bias for the other four models, although less pronounced than the Democratic role. Under

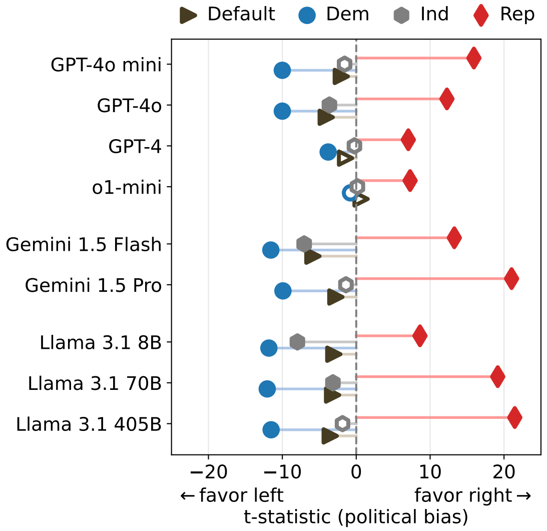

Figure 6: Political biases of different LLM-role configurations. The political biases are measured by t-statistics between the distributions of LLM rating bias scores on the left- and right-leaning sources. A negative/positive t-statistic means the LLM-role configuration favors left-/right-leaning news outlets. The solid symbols indicate statistically significant differences ($p < 0.05$).

the default configuration, where no explicit partisan roles are assigned to the LLMs, all models except for GPT-4 and o1-mini show a statistically significant liberal bias.

4.4 Political Bias and Rating Accuracy

The results in Figures 4 and 6 suggest a negative correlation between the political biases and the accuracy of LLM-role configurations. To confirm this, we present a scatter plot of all model-role configurations in Figure 7. Regardless of the direction, we find that a stronger political bias correlates with lower rating accuracy, although such correlation is only statistically significant for the cases where the LLM-role configuration is right-biased. This finding suggests that the misalignment between LLMs and human experts is partially due to the political biases embedded in LLMs and that reducing these biases could enhance rating accuracy.

Pennycook and Rand [33] ran an experiment in which participants produced news source ratings that correlated with their different ideologies. However, the combined ratings from participants on both sides correlate strongly with expert ratings. Given that LLMs exhibit consistent biases congruent with assigned roles, we test whether combining ratings from LLMs with different viewpoints could reduce overall biases and improve accuracy.

We consider two aggregate ratings for each model: (1) D+R, the average ratings of the Democratic and Republican roles, and (2) D+I+R, the average ratings of the Democratic, Independent, and Republican roles. Although different weights can be assigned to these partisan roles, we choose equal weights to reflect the composition of American voters with varying partisanship.[12]

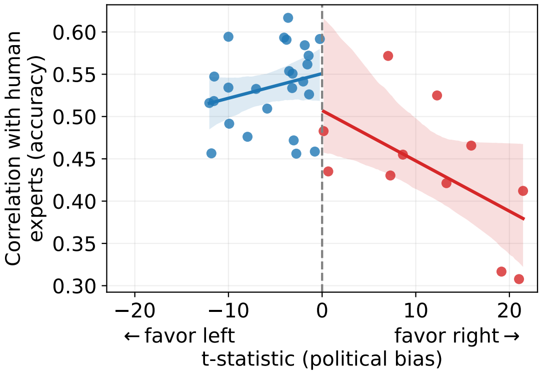

Figure 7: Political bias vs. rating accuracy for all LLM-role configurations. We use the t-statistics from the tests comparing the distributions of LLM rating bias scores on left- and right-leaning sources to quantify the political bias and the correlation with human experts for accuracy. LLM-role configurations with left or right biases are separated and the lines represent linear regressions on the two groups. For the left-biased data points, we have the Spearman correlation coefficient between the political bias and the rating accuracy of $\rho = 0.18$ ($p = 0.38$). For the right-biased data points, we have $\rho = -0.67$ ($p = 0.02$).

In Figure 8, we compare the accuracy of D+R and D+I+R with the results for different partisan roles. In most cases, D+R and D+I+R ratings are more accurate than either Democratic or Republican ratings, suggesting that blending different perspectives leads to better judgment of source credibility. Across all configurations, the Independent role is the winner for four models, while D+I+R leads in the other four cases. However, the differences between the Independent and D+I+R ratings are marginal. We, therefore, conjecture that aggregate ratings from partisan roles do not lead to significant improvement over the Independent role, which already exhibits relatively low political bias.

5 Discussion

5.1 Summary of Results

In this paper, we systematically audit nine widely used LLMs from three leading providers to test their ability to discern credible information sources from low-credibility ones. For most news sources tested in our experiments, the LLMs could provide ratings as instructed. However, larger models tend to refuse to rate less popular sources due to insufficient knowledge, while smaller models tend to make more errors. Comparing the LLM ratings, these models exhibit a high level of agreement despite being trained by different providers. Conversely, the models only moderately correlate with human experts.

[12] pewresearch.org/politics/2024/04/09/the-partisanship-and-ideology-of-american-voters

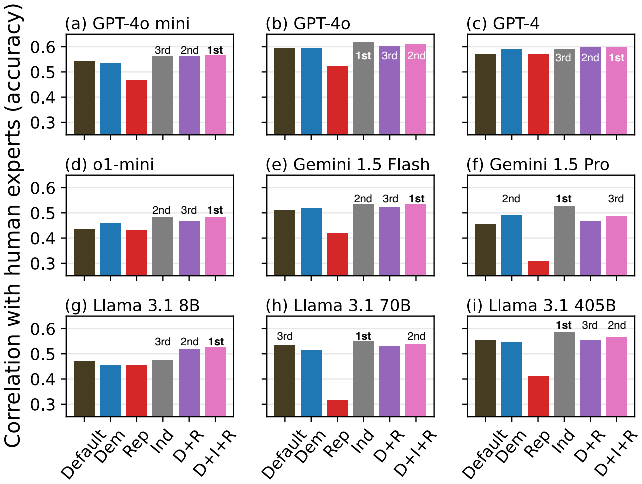

Figure 8: Accuracy of different LLM ratings. We show the Spearman correlation coefficients between different LLM ratings and those from human experts. For each model, we report the results for the default configuration, three partisan roles (Democrat, Independent, and Republican), as well as two aggregate ratings, D+R (average of Democratic and Republican roles) and D+I+R (average of all three political roles). The top three configurations are annotated.

It remains unclear how the LLMs acquire this capacity. Based on the explanations provided by the models alongside the ratings, we hypothesize that the models summarize descriptions of the given news sources from their training data and generate ratings accordingly. For instance, when encountering high-credibility sources, the LLMs often note that these are well-established and reputable news websites. This could explain the high correlation among the LLMs, as they likely have common training data [27].

Focusing on news sources with clear political leanings in the US context, we find that assigning partisan roles to LLMs steers their ratings to favor sources with congruent leanings in most cases. This result indicates that LLMs can reflect the viewpoints of humans with different political ideologies [2, 43]. In their default configurations, most LLMs exhibit a liberal bias: they are more likely to assign higher-than-expected credibility ratings for left-leaning sources compared to right-leaning ones. LLMs assigned with an Independent role are the least biased but still show a weak liberal bias. These findings align with previous studies indicating that many LLMs have left-leaning tendencies [39].

We also explore approaches to reduce political biases and increase the accuracy of the LLM ratings. We find that explicitly assigning an Independent role to an LLM or mixing the ratings from different roles could help achieve both objectives simultaneously. However, even unbiased ratings fail to align perfectly with human judgment.

5.2 Limitations

Our auditing has some limitations. First, since the evaluation relies on existing human ratings, any biases in these ratings may propagate into our findings. For example, our binary classification of news sources as left- or right-leaning based on a zero threshold applied to the scores by Robertson et al. [37] simplifies our bias analysis, but the alternative 5-class approach used by organizations like MBFC might provide more nuanced results.

In our experiments, we only provide the LLMs with the domains of the news sources, which is useful for assessing the completeness and accuracy of the internal knowledge. However, in real-world scenarios, the models are likely to have access to additional information about the sources, such as the metadata and content of the pages, which could potentially improve their performance [40].

There are different approaches to prompt the LLMs, which might yield different outcomes. For instance, one could employ a binary classification approach or ask LLMs to rank two sources at a time [35]. When assigning a partisan role to the LLMs, one could also provide them with detailed demographics and background information about a persona. Additionally, different prompt engineering techniques could be deployed to boost the performance of the LLMs. We were unable to test all of these approaches.

Our findings might not be generalizable to all LLMs and contexts. Despite our efforts to test as many representative models as possible, we are unable to cover all the other LLMs from different providers on the market. Given the rapid development in the field, new models with different behaviors will emerge soon. Finally, our experiments mainly focus on the US context, which might not be representative of other countries and non-English news.

5.3 Implications

Despite these limitations, we believe our findings provide valuable insights into the current state-of-the-art LLMs, considering that the general patterns are consistent regardless of the model size and provider. Given that these LLMs have been deployed in widely used systems, the findings have important implications.

Our experiments show that LLMs have the potential to serve as an affordable and accessible reference for source credibility ratings. Such ratings are vital for researchers to study the dissemination and prevalence of misinformation online [22, 25]. With the ability to provide contextual information and actively answer user questions, LLMs might also be adopted by the general public as scalable media literacy tools to investigate the credibility of news sources and perform fact-checking [16, 33]. However, our results indicate that the accuracy of LLMs is not perfect, calling for further analysis and comparison with other misinformation intervention methods [20]. Additionally, LLMs may be vulnerable to manipulation. For instance, bad actors could pollute LLM training data by creating fake news source review websites.

Our findings also highlight the challenges of using LLMs as information curators. Although not confirmed by the providers, leaked system prompts of popular LLMs suggest that they are configured to retrieve information according to the quality and credibility of sources when searching the Internet, along with other criteria such as the relevance to user questions and the freshness of the information.[13] However, it is unclear how the LLMs understand and implement these criteria in production.

A famous failure story came from Google in 2024 when its AI overview feature produced bizarre answers such as recommending that people eat rocks.[14] According to Google, these odd responses originated from low-quality information sources,[15] yet their AI models appeared to overlook it. Although these errors may have a limited negative impact, they highlight the risks of using LLMs for information curation in high-stakes contexts such as public health and elections.

Our findings provide valuable insights about such risks. We show that LLMs may lack information about many information sources, especially unpopular ones. However, it is inevitable for LLMs to encounter unfamiliar sources when curating information for users, especially when they search for misinformation topics [3, 6]. Under these circumstances, forcing the models to produce summaries or responses could inadvertently amplify lesser-known low-credibility sources. Even when the models know the information sources, including the highly popular ones, their assessment of credibility still deviates from human judgment. This inaccuracy might suppress credible sources and amplify low-credibility ones. When dealing with political information, the inherent bias of the LLMs in the default configuration might lead to distorted outcomes. The finding that assigning partisan roles to LLMs could induce corresponding biases suggests that LLMs tailored to user preferences and views might exacerbate echo chambers and polarization.

5.4 Future Studies

For future research, a critical challenge lies in mitigating biases and enhancing the accuracy of LLMs in assessing source credibility. While one might expect that models demonstrating superior performance on standard benchmarks would naturally excel at this specific task, our findings suggest otherwise. Our comparative analysis of models within the same family, such as the GPT series, reveals that increased model size and parameter count do not consistently translate to improved accuracy or reduced biases. Although OpenAI's o1 model series represents an innovative approach by training models specifically for complex reasoning tasks, our evaluation of o1-mini indicates that this enhanced reasoning capability has not effectively transferred to the present task. These findings suggest that more capable LLMs in the future might not necessarily perform better on the source credibility assessment task.

In our attempt to address this challenge, we leverage the prompt-based approach by assigning partisan roles to the LLMs and aggregating ratings from multiple roles. This approach is simple and cheap, and can be applied to both proprietary and open-source models, but the improvement is limited. Future models might consider explicitly integrating source credibility information into the models either through fine-tuning or including the data in the pre-training phase [32]. When embedding LLMs into larger systems such as search engines, one may also consider providing the models with the credibility ratings of the sources together with the content and query for reasoning.

Another key challenge is to enhance LLMs' ability to evaluate previously unseen information sources. Our findings demonstrate that LLMs struggle to assess the credibility of less popular sources, significantly limiting their effectiveness as information curators. While human expert evaluation could provide accurate ratings, this approach is neither sustainable nor scalable, given the vast number of sources and limited expert resources. Future research should explore automated frameworks that leverage LLMs to assess source credibility [40]. Such frameworks could autonomously gather information on the sources from the Internet, leveraging advanced reasoning capabilities to process such information, and generate reliable credibility ratings.

Other important directions include exploring how LLMs handle information from different sources in more realistic scenarios, how other information selection criteria, such as relevance and freshness [45], affect the outcomes, and how to mitigate emerging manipulation techniques [34]. At the same time, more studies on how humans interact with AI-augmented systems [13, 41] and how information processed by AI influences humans are needed [18, 42].

Acknowledgments

This work was supported in part by OpenAI's Researcher Access Program, Northeastern University Internet Democracy Initiative Seed Grant, Volkswagen Foundation as part of the "Bots Building Bridges" project, DARPA (grant HR001121C016), and the Knight Foundation.

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 298–306.

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (2023), 1–15.

[3] Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A Tucker. 2024. Online searches to evaluate misinformation can increase its perceived veracity. *Nature* 625, 7995 (2024), 548–556.

[4] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring Implicit Bias in Explicitly Unbiased Large Language Models. arXiv:2402.04105 [cs.CY] https://arxiv.org/abs/2402.04105

[5] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[6] Danah Boyd and Michael Golebiewski. 2018. Data voids: Where missing data can easily be exploited.

[7] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, S1 (2016), 250–271.

[8] Peter Carragher, Evan M. Williams, and Kathleen M. Carley. 2024. Detection and Discovery of Misinformation Sources Using Attributed Webgraphs. *Proceedings of the International AAAI Conference on Web and Social Media* 18, 1 (2024), 214–226.

[9] Wen Chen, Diogo Pacheco, Kai-Cheng Yang, and Filippo Menczer. 2021. Neutral bots probe political bias on social media. *Nature communications* 12, 1 (2021), 5580.

[10] Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Toronto, Canada, 1504–1532.

[11] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

[12] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2021. Political Polarization on Twitter.

[13] github.com/jujumilk3/leaked-system-prompts

[14] bbc.com/news/articles/cd11gzejgz4o

[15] blog.google/products/search/ai-overviews-update-may-2024

Proceedings of the International AAAI Conference on Web and Social Media 5, 1 (2021), 89–96.

[13] Matthew R. DeVerna, Harry Yan, Kai-Cheng Yang, and Filippo Menczer. 2024. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences* 121, 50 (2024), e2322823121.

[14] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (2024), 1097–1179.

[15] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. arXiv:2301.01768 [cs.CL] https://arxiv.org/abs/2301.01768

[16] Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Using ChatGPT to Fight Misinformation: ChatGPT Nails 72% of 12,000 Verified Claims. psyarxiv:qnjkf

[17] Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M. Rothschild, and Duncan J. Watts. 2021. Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences* 118, 32 (2021), e2101967118.

[18] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages.

[19] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[20] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M Herzog, Ullrich KH Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, et al. 2024. Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour* 8, 6 (2024), 1044–1052.

[21] Jana Lasser, Segun Taofeek Aroyehun, Almog Simchon, Fabio Carrella, David Garcia, and Stephan Lewandowsky. 2022. Social media sharing by political elites: An asymmetric American exceptionalism. arXiv:2207.06313 [cs.CY] https://arxiv.org/abs/2207.06313

[22] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[23] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings 2019 Network and Distributed System Security Symposium (NDSS 2019)*. Internet Society, San Diego, California, USA, 1–15.

[24] Alice Li and Luanne Sinnamon. 2024. Generative AI Search Engines as Arbiters of Public Knowledge: An Audit of Bias and Authority. *Proceedings of the Association for Information Science and Technology* 61, 1 (2024), 205–217.

[25] Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. 2023. High level of correspondence across different news domain quality rating sets. *PNAS Nexus* 2, 9 (2023), pgad286.

[26] Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 7001–7025.

[27] Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024. Datasets for Large Language Models: A Comprehensive Survey. arXiv:2402.18041 [cs.CL] https://arxiv.org/abs/2402.18041

[28] Llama Team, AI at Meta. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[29] Shahan Ali Memon and Jevin D. West. 2024. Search Engines Post-ChatGPT: How Generative Artificial Intelligence Could Make Search Less Reliable. arXiv:2402.11707 [cs.IR] https://arxiv.org/abs/2402.11707

[30] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *AI and Ethics* 4, 4 (2023), 1085–1115.

[31] OpenAI. 2024. OpenAI o1 System Card. arXiv:2412.16720 [cs.AI] https://arxiv.org/abs/2412.16720

[32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[33] Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[34] Samuel Pfrommer, Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. 2024. Ranking Manipulation for Conversational Search Engines. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 9523–9552.

[35] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, Mexico City, Mexico, 1504–1518.

[36] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477–486.

[37] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on human-computer interaction* 2, CSCW, Article 148 (2018), 22 pages.

[38] Abel Salinas, Louis Penafiel, Robert McCormack, and Fred Morstatter. 2023. "Im not Racist but...": Discovering Bias in the Internal Knowledge of Large Language Models. arXiv:2310.08780 [cs.CL] https://arxiv.org/abs/2310.08780

[39] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect?. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, Honolulu, Hawaii, USA, 29971–30004.

[40] Michael Schlichtkrull. 2024. Generating Media Background Checks for Automated Source Critical Reasoning. arXiv:2409.00781 [cs.CL] https://arxiv.org/abs/2409.00781

[41] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages.

[42] Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. 2024. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus* 3, 2 (01 2024), pgae035.

[43] Gabriel Simmons. 2023. Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Toronto, Canada, 282–297.

[44] Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. arXiv:2307.03744 [cs.HC] https://arxiv.org/abs/2307.03744

[45] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 14918–14937.

[46] Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang. 2024. The Use of Generative Search Engines for Knowledge Work and Complex Tasks. arXiv:2404.04268 [cs.IR] https://arxiv.org/abs/2404.04268

[47] Daniel Trielli and Nicholas Diakopoulos. 2019. Search as News Curator: The Role of Google in Shaping Attention to News Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15.

[48] Aleksandra Urman and Mykola Makhortykh. 2025. The silence of the LLMs: Cross-lingual analysis of guardrail-related political bias and false information prevalence in ChatGPT, Google Bard (Gemini), and Bing Chat. *Telematics and Informatics* 96 (2025), 102211.

[49] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. arXiv:2310.03214 [cs.CL] https://arxiv.org/abs/2310.03214

[50] Evan M. Williams, Peter Carragher, and Kathleen M. Carley. 2024. Bridging Social Media and Search Engines: Dredge Words and the Detection of Unreliable Domains. arXiv:2406.11423 [cs.SI] https://arxiv.org/abs/2406.11423

[51] Zhijing Wu, Mark Sanderson, B. Barla Cambazoglu, W. Bruce Croft, and Falk Scholer. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1635–1644.

[52] Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When Search Engine Services Meet Large Language Models: Visions and Challenges. *IEEE Transactions on Services Computing* 17, 6 (2024), 4558–4577.

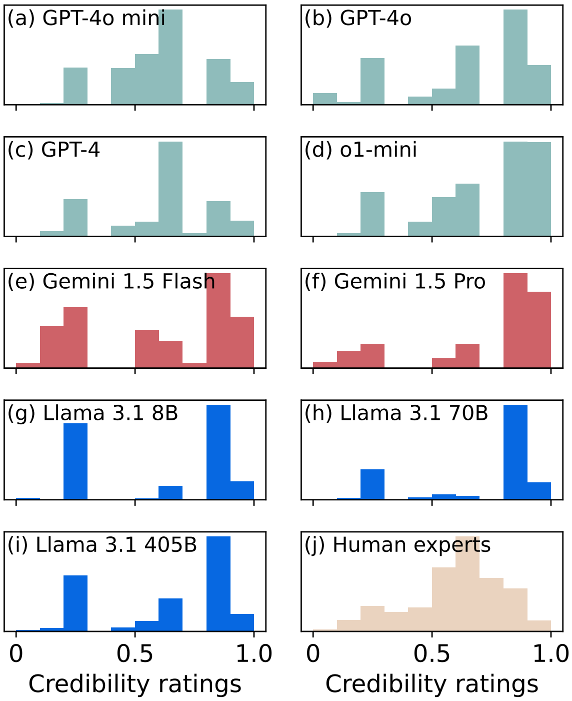

Figure 9: Distribution of domain credibility ratings from different LLMs and human experts.

A Complete Prompts

In the default configuration, the following prompts are used in our experiments.

System prompt:

> **System prompt:**
> You are an assistant to determine the credibility of websites.

Task instruction:

> **Task instruction:**
> Rate the website's credibility: {domain}, on a scale between 0 and 1, where 0 means very low credibility and 1 means very high credibility. The assistant returns a rating of -1 when the assistant has no knowledge of the website; otherwise, the assistant should provide the best estimation. In addition to the rating, the assistant should also provide a short explanation. The output should be in JSON format and follow the following schema. Only return the JSON object, nothing else. {'url': 'example.com', 'rating': 0.5, 'explanation': 'The example website is known to post credible content.'}

Note that we specifically instruct the models to return the results in JSON format for easy parsing. We use the JSON mode for GPT and Gemini models to ensure their LLM responses are in valid JSON format. Unfortunately, the Llama models hosted by together.ai do not support JSON mode. They are usually able to follow the instructions and return the responses in JSON format, but extra cleaning and re-formatting are sometimes required to correctly parse the responses.

When testing the impact of partisan roles, we modify the system prompt:

> **Partisan role assignment:**
> You are an assistant to determine the credibility of websites. You identify as {role} on the US political spectrum.

In the experiments, {role} is replaced with the one of "a Democrat," "an Independent," or "a Republican." The task instruction remains the same.

B Response Examples

In this section, we provide some response examples from the LLMs. Table 1 lists the LLM ratings of reuters.com. Table 2 lists the LLM ratings of aldf.com.

C Additional Results

In Figure 9, we show the distribution of domain credibility ratings from different LLMs and human experts.

Table 1: Ratings of reuters.com by different LLMs.

Model	Rating	Explanation
GPT-4o mini	1	Reuters is a well-established international news organization known for its commitment to factual reporting and journalistic integrity. It has a strong reputation for accuracy and is widely regarded as a reliable source of news.
GPT-4o	1	Reuters is a well-established international news organization known for its accurate and unbiased reporting. It is widely respected in the journalism community and has a long history of providing reliable news.
GPT-4	0.95	Reuters is a well-established and highly reputable global news organization known for its commitment to accurate and unbiased reporting.
o1-mini	0.9	Reuters is a well-established news organization known for reliable and unbiased reporting.
Gemini 1.5 Flash	0.9	Reuters is a well-established and reputable news agency known for its journalistic integrity and fact-checking practices.
Gemini 1.5 Pro	1	Reuters is a highly reputable international news agency known for its journalistic integrity and factual reporting.
Llama 3.1 8B	0.9	Reuters is a well-established and reputable news agency with a long history of providing accurate and unbiased reporting.
Llama 3.1 70B	0.95	Reuters is a well-established and reputable news agency with a long history of providing accurate and unbiased reporting. It is widely considered a trusted source of news and information.
Llama 3.1 405B	0.9	Reuters is a well-established and reputable international news agency with a long history of providing accurate and unbiased reporting.

Table 2: Ratings of aldf.com by different LLMs. Five models believe the website belongs to the Animal Legal Defense Fund (aldf.org).

Model	Rating	Explanation
GPT-4o mini	0.8	The Animal Legal Defense Fund (ALDF) is a well-established organization focused on animal rights and legal advocacy. It is recognized for its credible efforts in promoting animal welfare and has a strong reputation in the field.
GPT-4o	-1	The assistant has no knowledge of the website aldf.com.
GPT-4	-1	I do not have enough information to assess the credibility of aldf.com.
o1-mini	0.7	The website aldf.com is associated with the American Lebanese Democratic Federation, a recognized organization, indicating a moderate to high level of credibility.
Gemini 1.5 Flash	-1	I have no information about this website.
Gemini 1.5 Pro	1	ALDF.org is a highly reputable website of the Animal Legal Defense Fund, a well-known non-profit organization.
Llama 3.1 8B	0.8	The Animal Legal Defense Fund (ALDF) is a well-established and reputable organization that provides accurate and reliable information on animal law and advocacy.
Llama 3.1 70B	0.8	The Animal Legal Defense Fund (ALDF) is a well-established and reputable organization that advocates for animal rights and welfare. Their website provides credible information on animal law and related topics.
Llama 3.1 405B	0.8	The Animal Legal Defense Fund (aldf.org) is a well-established organization with a clear mission and a history of advocating for animal welfare through the legal system. Their website provides credible information on animal law and related topics.