Evaluating Credibility on the Web: CredBot and the Role of AI in Real-Time Credibility Assessments

EMILIE ZHANG* and JIANFEI LYU*, Wellesley College, USA
LAN DAU, Wellesley College, USA
MALIKA PARKHOMCHUK, Wellesley College, USA
JENNIFER LONG, Wellesley College, USA
ALEXA HALIM, Wellesley College, USA
ORIT SHAER, Wellesley College, USA
ENI MUSTAFARAJ, Wellesley College, USA

In today's digital landscape, discerning credible online content is increasingly challenging as misinformation proliferates. Despite AI advancements, fully autonomous credibility assessments remain an elusive goal, with current systems often struggling with context sensitivity and nuanced credibility indicators. This paper presents CredBot, a Chrome extension powered by large language models (LLMs) designed to provide real-time, in-browser credibility assessments and serve as an educational tool that encourages users to engage critically with online information. Our study examines whether modern AI can effectively interpret complex credibility signals and function as a scalable educational tool that provides transparent credibility reasoning across diverse content. Initial results indicate almost 80% accuracy in flagging low-quality content, while highlighting areas for improvement in subtle credibility detection. Overall, CredBot illustrates the potential for AI-driven credibility educational tools that have key advantages in scalability and autonomy, reducing users' need to manually check websites for credibility.

CCS Concepts: • **Human-centered computing → Natural language interfaces**; **Web-based interaction**; • **Information systems** → *Page and site ranking*; *Chat*; • **Computing methodologies → Information extraction**; *Natural language generation*.

Additional Key Words and Phrases: Credibility Assessment, Conversational User Interfaces (CUI), Automated Credibility Checking, Web Literacy

ACM Reference Format:
Emilie Zhang, Jianfei Lyu, Lan Dau, Malika Parkhomchuk, Jennifer Long, Alexa Halim, Orit Shaer, and Eni Mustafaraj. 2025. Evaluating Credibility on the Web: CredBot and the Role of AI in Real-Time Credibility Assessments. In *Proceedings of The 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '25).* ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.

Authors' Contact Information: Emilie Zhang, ez102@wellesley.edu; Jianfei Lyu, jl7@wellesley.edu, Wellesley College, Wellesley, Massachusetts, USA; Lan Dau, Wellesley College, Wellesley, USA, ldau@wellesley.edu; Malika Parkhomchuk, Wellesley College, Wellesley, USA; Jennifer Long, Wellesley College, Wellesley, USA; Alexa Halim, Wellesley College, Wellesley, USA; Orit Shaer, Wellesley College, Wellesley, USA, oshaer@wellesley.edu; Eni Mustafaraj, Wellesley College, Wellesley, USA, eni.mustafaraj@wellesley.edu.

1 Introduction

In the seminal 2001 article *The Semantic Web*, Tim Berners-Lee and colleagues envisioned an internet where intelligent agents could navigate and evaluate web content independently, assisting users in complex decision-making by "understanding" the underlying data [2]. Today, advanced AI models can process vast amounts of information, understand natural language, and even analyze contextual cues within web content. This brings us significantly closer to realizing the scenario envisioned in *The Semantic Web*. Yet even now, the challenge of empowering agents to reliably assess the credibility of websites remains largely unsolved. Challenges such as context sensitivity, the complex nature of misinformation, and subtle nuances in credibility indicators mean that it is difficult for current AI to provide consistently accurate assessments [16]. For example, parsing factors like ads density, content transparency, and authorial bias requires a degree of contextual awareness and adaptability that human experts naturally bring to the task — qualities that AI systems have only recently begun to demonstrate. Thus, achieving a reliable, fully autonomous credibility assessment tool for websites remains an open frontier in AI research and development.

For these reasons, we developed CredBot,[1] a conversational user interface (CUI) Chrome extension powered by large language models (LLMs) that automates the credibility-checking process as a latent assistant. CredBot is designed to evaluate web content with a set of credibility signals while remaining unobtrusive, enabling users to read content and view credibility assessments without leaving the page. CredBot also serves as an educational tool by providing detailed credibility analysis across different categories and citing sources referenced in its evaluations. Combining the accessibility and affordability of LLMs and a user-friendly platform, we envision tools like CredBot as practical solutions that can be widely adopted by users, empowering everyday individuals to engage critically with online information in a way that is both efficient and impactful.

With CredBot, we wish to test the limits to which today's AI can fulfill the visions of intelligent agents as outlined in *The Semantic Web*, specifically in the nuanced field of web credibility assessment. More broadly, we seek to determine whether current LLMs have the capabilities to bridge the gap between manual and autonomous credibility assessment.

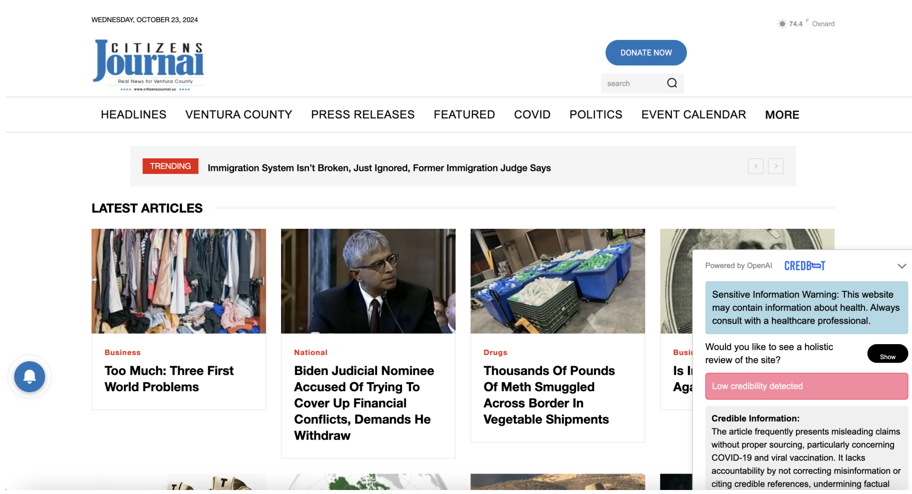

Fig. 1. Example of CredBot evaluating Citizens Journal as low credibility

[1] https://github.com/ritajflyu2/CredBot

2 Related Work

Users are found to be easily swayed by statements that are strategically framed to appear factual and struggle with recognizing native ads — advertisements that seamlessly blend with website content — as promotional material [4, 7, 18, 21]. In 2021, Google introduced features "About this result" and "More about this page" on their Search Engine Results Pages (SERPS) in an attempt to provide background on unfamiliar websites. However, Wang et. al. highlights that these tools are often insufficiently noticed by everyday users, who either overlook these features or lack the motivation to engage with them [17]. Stanford Persuasive Technology Lab's study shows only 14.3% among over 2,500 participants mentions "Accuracy of Information" as a credibility factor [5]. In other words, the majority of users lack the critical evaluation skills necessary to assess online sources and are unmotivated to engage with readily available tools that could assist in credibility assessments, leaving them vulnerable to unreliable information.

Several existing tools and resources attempt to address this issue by offering credibility assessments of online content. For instance, Media Bias/Fact Check (MBFC)[2] rates websites by political bias and factual accuracy to help users assess credibility signals such as Factual Reporting, Bias, Transparency of the Site, and Traffic/Longevity. NewsGuard[3], on the other hand, rates the reliability of news domains based on nine credibility signals. There are also credibility checkers like ScamAdvisor[4] and IFCN (International Fact-Checking Network)[5] that hosts a database of fact-checks. W3C Credible Web Community Group (CWCG)[6] also collects and catalogs a set of crowd sourced credibility signals to support researchers. However, these tools require significant manual resource. They depend on expert evaluators (MBFC, NewsGuard), which allows evaluation of only a subset (8,000 sites for MBFC and 10,000 news and media domains for NewsGuard) of sites across the 2 millon that are active at any given time [11], excluding site types such as blogs, informational resources, forums, or commercial websites, where misinformation is just as prevalent. Moreover, alternatives such as IFCN, ScamAdvisor, MBFC's database either require users to manually input each link or statement for verification or fail to provide automated evaluation system such as W3C CWCG. Given the typical user's low motivation to perform credibility checks on information of the site [17], these manual verification systems are unlikely to gain widespread adoption.

Research in browser-based tools, augmented interfaces, and conversational agents has focused on enhancing user support and credibility assessments across various online environments. Schwarz and Morris (2011) introduced visual augmentations within search results and web pages to support credibility judgments by showing indicators such as expert popularity and overall popularity directly alongside content, which significantly helped users without topic expertise assess content credibility more accurately, illustrating the effectiveness of data-driven browser enhancements as credibility support tools [14]. Additionally, previous research illustrates bots can have as an educational tool that engages contextually to provide information and explanations on learning recommendations [1]. Following these directions, our work builds on these approaches, leveraging a conversational agent within a browser extension to support nuanced credibility evaluations directly within the browsing experience, making it a convenient tool for users seeking trustworthy information online.

[2]https://mediabiasfactcheck.com/methodology/
[3]https://www.newsguardtech.com/ratings/rating-process-criteria/
[4]https://www.scamadviser.com/
[5]https://www.poynter.org/ifcn/
[6]https://credweb.org/signals-20191126#h.4u9corwqzqo3

3 Architecture of CredBot

CredBot is a 3-component system integrated with users browsing experience as a Chrome extension. It is designed to be scalable and user-friendly through seamless usage of LLM in the background and a dedicated user interface (UI).

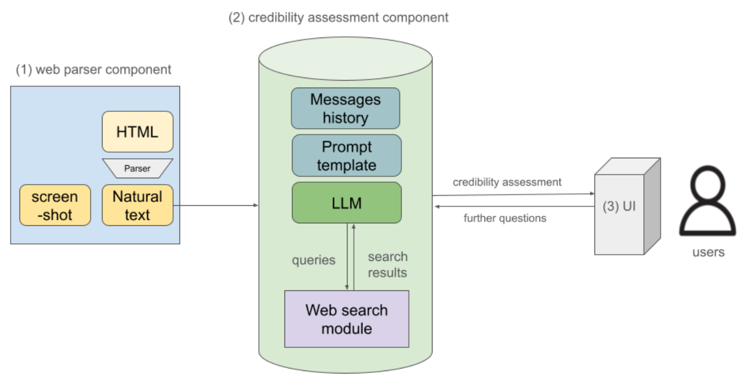

Fig. 2. CredBot system design overview. The system first takes a screenshot and scrapes the website's raw HTML data upon users' visit and parses this into a natural language text format. Next, the system reviews the screenshot and website text and flags users about the presence of health information. Then, the system performs fact checking through a web search module, which get presented along with our set of criteria to elicit a credibility assessment from the LLM. Users can then make further inquiries through the Chatbot component.

An overview of CredBot's system design is shown in Figure 1. Here is a description of its main components:

(1) The first component consists of a custom built web parser to scrape the homepage raw HTML and process it to a human readable text format. This allows us to significantly reduce the number of tokens passed to the LLM. We also take a screenshot of the website to capture the placement of articles and ads.
(2) The credibility assessment module consists of medical information detection and credibility assessment. We first prompt the LLM to review the website text for presence of medical information and flag the users if they are found. Then we prompt the LLM to extract claims in the text for generation of search queries, through which results provide context for the LLM to assess the website based on our set of criteria. We will go into more details in Section 4 on what these criteria are, how we create them, and our prompt for the LLM to follow the credibility assessment flow.
(3) The final LLM assessment of site credibility is presented on the UI, along with explanation for its classification. URLs referenced to support the explanation are also included so that users can verify the explanation. The UI also presents a Chatbot interface that allows users further questions to the LLM.

4 Methodology

4.1 Choice of LLM

We choose to use GPT4o-mini as CredBot's LLM for its structured output capability, 128k context window, and low cost, which enables the system to economically synthesize credibility assessment over our multi-step pipeline.

4.2 Credibility Signals

We develop our own set of credibility signals by cross-referencing existing signals from NewsGuard, MBFC, and the W3C crowdsourced credibility signal draft [10, 12, 19]. While finding commonalities between the 3 sources and grouping the criteria into main categories, one important task is making the signals interpretable by LLMs without needing human judgment, as certain criteria are not well defined enough to leverage LLMs' judgment. For example, the criteria with most weight in the Credibility category in News Guard, "Does not repeatedly publish false or egregiously misleading content" [12], is difficult for GPT4o-mini to assess accurately due to its knowledge cut-off. Determining a site's credibility reputation would be highly controversial since LLMs like GPT4o-mini can reflect biases from the data they are trained on. Recognizing the potential for biases inherent to LLMs trained on mixed-source data [6], we provided additional few-shot examples to improve GPT4o-mini's accuracy. We divided the credibility signals into 4 main categories, each with a set of criteria (Table 1), the full details of each criterion are included in the Appendix.

Category 1: Credible Information	Category 2: Language	Categoory 3: Transparency	Category 4: Ads and Sponsorship
Uses Credible Sources: to be credible, the site should gather and present their information responsibly.	Bias: to be credible, the site should not take information out of context to further their viewpoint.	The site is more credible if it discloses its ownership and its source of financing	Clear Labels: to be credible, a site should disclose its commercial relationships and clearly label its ad content.
No Misinformation: to be credible, a site should be consistently factual and rely on credible information and sources.	Avoids deceptive/sensational titles: to be credible, a site should try not to use deceptively sensational titles that attempt to trick the reader into reading the article.	The site is more credible if the site provides the names of its content creators, along with either contact or biographical information. Know that this information may not be located on the home page.	Ad volume: to be credible, a site should have a reasonable ad volume with high ad quality. Generally speaking, larger ad volume indicates less credibility – however, also consider the quality of the ads in conjunction with ad volume.
	Language formality: to be credible, the site should generally have correct spelling, correct grammar, and use formal language.	The site is more credible if it has contact information readily available	
	Distinction between news and opinion: to be credible, the site should responsibly differentiate between news and opinion.	The site is more credible if it has an about page	
	Title Quality: to be credible, a site should have a title that is representative of the content of the article, and not misleading about its contents.	The site is more credible if it discloses possible conflicts of interest.	

Table 1. Four main categories of Credibility Signals.

4.3 Weight Determination: Determining the Category Importance

When determining the importance of each category for creating accurate credibility assessments, we selected 12 websites, with 4 from each of MBFC's High, Medium, and Low credibility categories. Two researchers separately labeled scores for the four individual categories (credible information, language, transparency, ads and sponsorship) for each website. We then fit a classification decision tree to MBFC's label using the categories as features. We observed that Category 1 is the only statistically significant feature; such that it corresponds to the final credibility evaluation.

4.4 Web Search

We improve the accuracy of Category 1 by implementing a web search function to verify the credibility of its contents. Since websites can use credible sources to generate misinformation by taking information out of context, we decide to utilize web search only for fact checking claims. We first provide the context of the research, then instruct LLM to summarize and select 5 key statements that the site makes and generate succinct queries to verify the statements. We then collect the top 3 results for each query using the DuckDuckGo search API [3]. These results are then included in our prompt for the LLM to deliver the assessment for Category 1. We include a list of links that it referenced in its assessment under the initial credibility message, enabling users to easily verify each source.

Here we describe the instructions to LLM to generate queries for web searches[7]. We prohibit the LLM from referencing MBFC since sites assessed by MBFC comprise our benchmark set. Additionally, we found that LLM can generate broad queries asking if the site contains misinformation in general, so we restrict it to only query about the statements that the site makes.

> You are a professional assistant with neutral political stand meant to help the average user determine the credibility of the news website they are currently on by determining whether its sources are credible and their claims have misinformation.
> Using the website text passed to you, summarize the statements that the site/article makes and select 5 key statements you want to verify.
> Generate 5 succinct queries, maximum 10 words, to help you verify the statements you selected from the site. Do not use Media Bias Fact Check (MBFC) page of the site itself as a source to verify the queries.
> Do not generate broad queries that ask if the site contains misinformation, keep it specific about the 5 key statements from the site.

4.5 Final Credibility Signals and Usage

Our final version only uses the criteria in category 1 (Credible Information) along with the web search, to determine the page's credibility (Table 1). While the other three categories are not factored into CredBot's primary credibility evaluation, they serve an educational role that promotes a more critical and informed perspective in users when evaluating online information. We included these categories to guide users on additional factors to consider when assessing a site's credibility.

Below is the prompt to evaluate the site's overall credibility using category 1: Credible Information[8]. This process mirrors our method for assessing credibility in web search results. We first give the context of the result, then prompt the LLM to use the screenshot of the site, the site's text, and the web search results to analyze the site based on 2 criteria:

[7] https://github.com/ritajflyu2/CredBot/blob/main/Prompts.md
[8] 7

Use Credible Sources and No Misinformation. Recognizing the LLMs have a left leaning bias [20], we emphasize the importance of maintaining a neutral viewpoint in credibility and clarify that a site can still be credible if it have certain bias as long as it doesn't use misinformation to support its bias.

> You are a professional assistant meant to help the average user determine the credibility of the news website they are currently on.
> Using the screenshot of the website, along with website text and additional research results on claims made on the site, reference the sources from your search results, perform logical reasoning analysis to determine whether this website is high in credibility, medium in credibility, or low in credibility, using the following criteria:
> Uses Credible Sources: to be credible, the site should gather and present their information responsibly. Examples that support this criteria and would be beneficial to the site's credibility include: The site generally references reputable sources for its information; The site generally publishes claims that can be fact-checked by gathering information from credible sources or direct evidence; the site points out false claims if such quotes are used in articles to ensure readers are aware of its falsity; when using attributions of non-original content, attributions are given and are accurate.
> No Misinformation: to be credible, a site should be consistently factual and rely on credible information and sources. Examples of misinformation include : 1. Does not use facts. 2. Deliberately chooses to use false information/misinformation. 3. Spreads fake news, conspiracy theories, or propaganda as fact. Instead, the site should generally report on events factually and present information in context.
> Since your system has a left-leaning bias, it's important stand in a neutral viewpoint, and to note that a credible site can have a particular bias and political affiliation, but it cannot use misinformation, fake news, or scientifically inaccurate conspiracies to support its bias.

5 CredBot

CredBot is a Chrome extension powered by GPT designed to assist users in evaluating web credibility. The following subsections provide an in-depth look at each core feature, illustrating how CredBot enables users to navigate online content with increased awareness and critical thinking.

5.1 Sensitive System Warning

CredBot includes a sensitive information alert system that identifies when a web page contains potentially sensitive medical or health-related content (Fig 4).

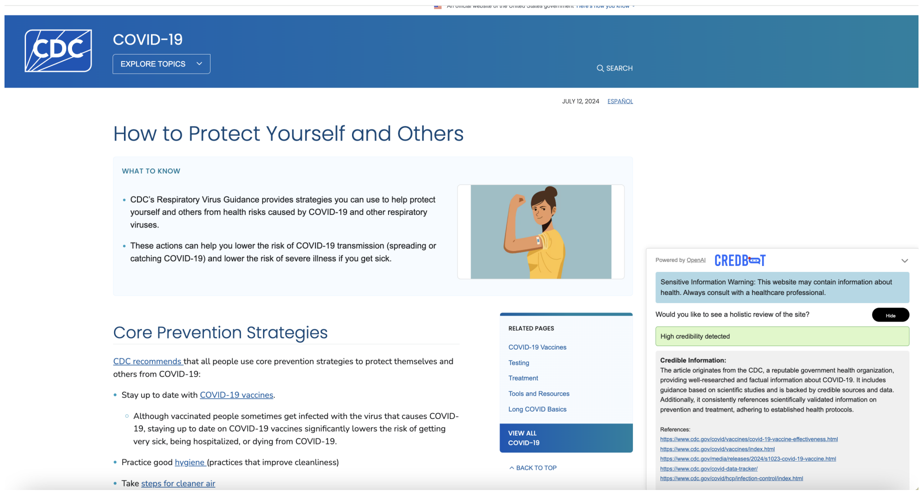

Fig. 3. Example of CredBot operating on a CDC website and high credibility banner

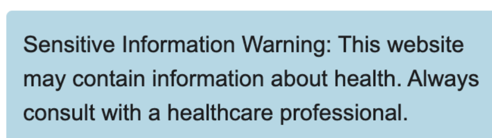

Fig. 4. Sensitive Information Warning

By incorporating this feature, CredBot aims to become a more comprehensive tool that not only guides users in evaluating website credibility, but also assists users in recognizing when they are exposed to potentially sensitive or misleading health information.

5.2 Credibility Evaluation

At the core of CredBot's functionality is its ability to assess and convey the credibility of a website through a color-coded banner system. CredBot's credibility evaluation appears as a colored banner: green for high (Fig 3), yellow for medium (Fig5), and red for low credibility (Fig 6) along with a brief evaluation message.

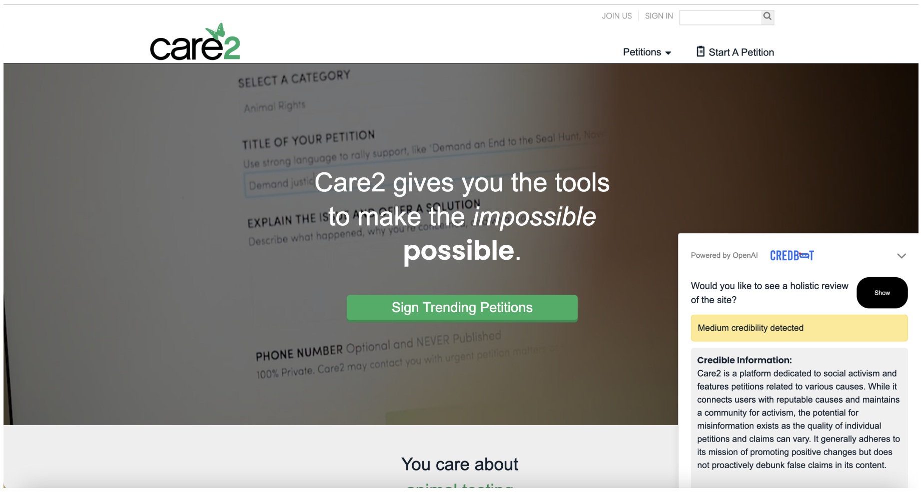

Fig. 5. Example of medium credibility banner

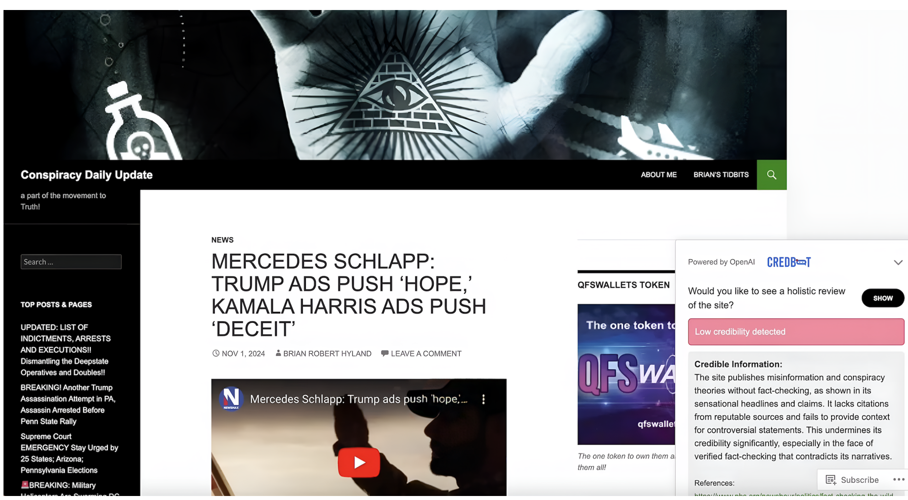

Fig. 6. Example of low credibility banner

This evaluation is based on the "credible information" category, which assesses 1) the reliability of sources and 2) the presence of misinformation. CredBot also provides a list of URLs that it referenced to make the credibility evaluation (Fig 7).

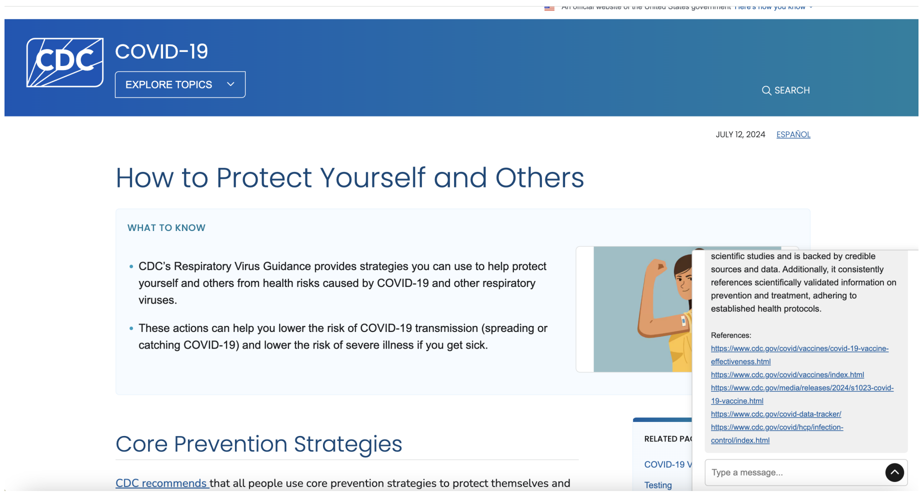

Fig. 7. CredBot: example of source references

When users seek more detailed insights by pressing the "show" button, CredBot displays additional information related to the other three categories, language, transparency, and ads and sponsorship (Fig 8), as well as a disclaimer (Fig 9).

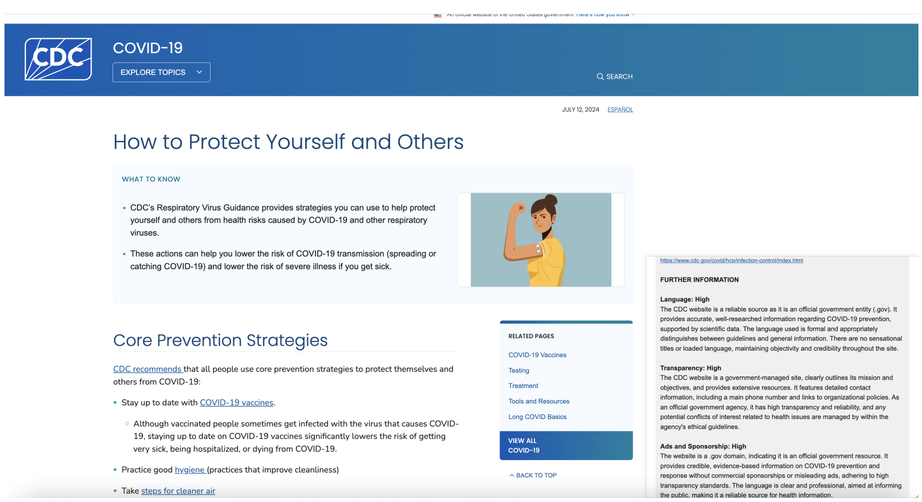

Fig. 8. CredBot: example of additional information

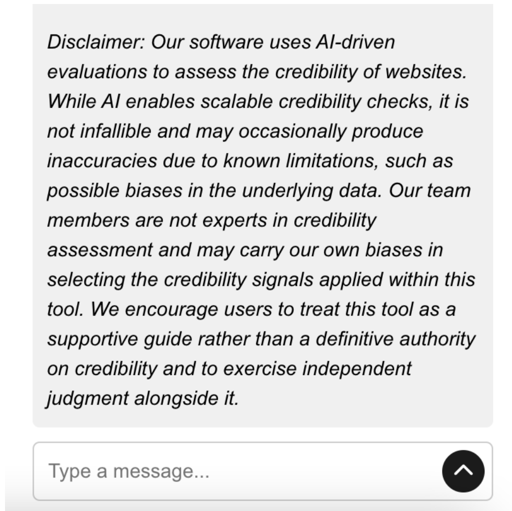

Fig. 9. CredBot Disclaimer

5.3 Chatbot User Interactions

CredBot is designed to allow users to ask questions directly within the interface, as shown in Fig 10 and Fig 11. CredBot leverages a CUI powered by GPT4o-mini, enabling it to give real-time responses to a wide variety of questions related to user's current site. Users can inquire about specific aspects of the credibility evaluation or request further details on the criteria used, as well as any other question they have.

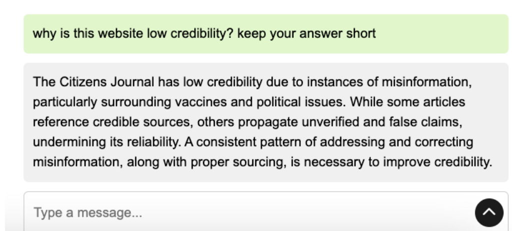

Fig. 10. CredBot: chatbot user interaction

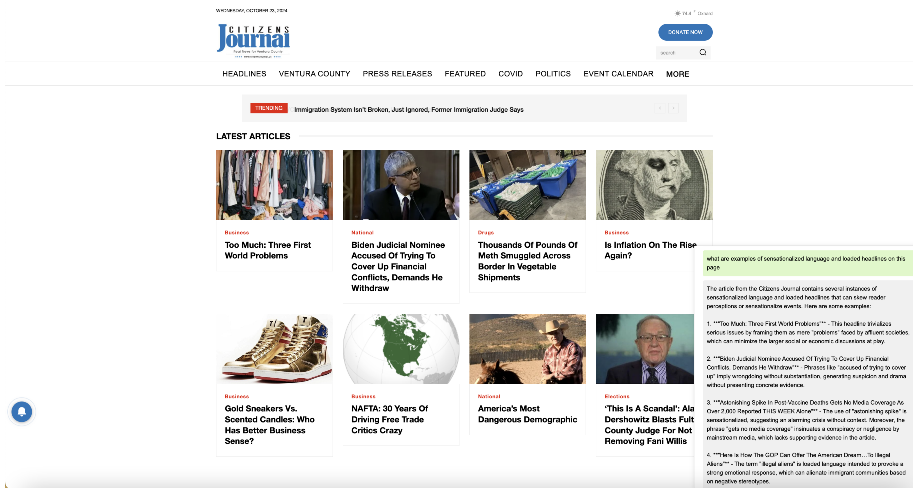

Fig. 11. CredBot: chatbot user interaction 2

6 Experiments & Results

6.1 Experiment Setup

For benchmarking, we randomly selected 45 websites across 7 types of websites designated by Media Bias Fact Check: pro-science, political left leaning, political center leaning, political right leaning, conspiracy-pseudoscience, fake-news, impostor. Details on the benchmarking dataset can be found in Appendix. The mean accuracy for websites in the dataset are recorded across 7 runs.

6.2 Results

Credibility Level	Accuracy
High	73.33
Medium	68.57
Low	77.14

Table 2. Accuracy performance on benchmark dataset across High, Medium, Low credibility websites.

The accuracy performance across the credibility results are shown in Table 2, and accuracy performance across types of website is shown in Table 3. We observed that our system scores higher at predicting credibility for pro-science, politically left leaning, conspiracy-pseudoscience, and fake-news. This suggests that CredBot does better with websites that display stronger bias and misinformation that can be debunked through fact checking.

7 Limitations and Future Work

7.1 Limitations

CredBot struggles with 'impostor sites' that closely mimic the design and content of real news sources but lack legitimacy, existing primarily to mislead users without disclosing their true affiliations or intentions. For instance, "Dartmouth

Type of website	Accuracy
Pro-science	100
Left leaning	76.19
Center leaning	54.76
Right leaning	70.32
Conspiracy-pseudoscience	100
Fake-news	83.33
Impostor	50.00

Table 3. Accuracy performance on benchmark dataset across 7 types of websites

Times" is rated "High" by CredBot because it detected factual reporting on topics like local gas prices and references to credible sources — criteria central to CredBot's credibility evaluation. However, Media Bias/Fact Check rates it as "Low Credibility" because it primarily relies on algorithm-generated content — a strategy that is particularly deceptive to LLMs [9]. CredBot's lower accuracy with center-leaning sites suggests challenges in assessing websites with an ambivalent political stance (Table 3). Testing also reveals CredBot's tendency to reflect academic and institutional values, often rating right-leaning sources as more biased, aligning with prior studies [13].

7.2 Future Work

Our next step is to further refine our credibility signals and weighting system. Our team's labeling process inevitably reflects our own biases; in future work, we aim to conduct further research on credibility signals and weighting, ideally involving a diverse group of experts. Leite et al. (2024) found 12 out of the 19 signals they proposed exhibit a statistically significant association across all datasets [8]. We plan to reference their findings to expand our future set of signals. Additionally, we plan to add features allowing users to adjust the weight of each category, making CredBot a more interactive and educational tool. Currently, CredBot's evaluation message can tend to be generic, using concepts such as "sensational title" and "loaded words" that may be unfamiliar to the average user. Our next step is to provide more explanations for such terms, and to include direct quotes and specific examples from the referenced sites. This approach will give users concrete evidence of where these cases appear, creating more trust in CredBot's assessments.

We plan to conduct user studies of CredBot to evaluate its usability, effectiveness, and user trust in its credibility assessments, targeting a diverse group of users with varying educational backgrounds and ages.

Furthermore, while we were finalizing CredBot's software and manuscript, Tian et al. (2024) published *Web Retrieval Agents for Evidence-Based Misinformation Detection*, which introduces a comprehensive approach to integrating web search agents with LLMs for automated fact-checking and misinformation detection. We intend to incorporate elements of their approach to improve web retrieval for CredBot's credibility assessments in future iterations of the tool[15].

8 Significance

Our work with CredBot demonstrates LLMs' potential to support credibility assessment on a broad scale, bridging key gaps in current methods and expanding the reach of credibility tools. As a tool capable of assessing the credibility of any website, CredBot addresses key scalability limitations found in existing credibility checkers, which rely on human evaluators that, despite significant effort, can only cover a fraction of the web, leaving the vast majority of sites unevaluated. Furthermore, CredBot's use of LLMs offers a cost-effective alternative to expert-driven assessments, such as NewsGuard, which, while highly thorough, are resource-intensive. Furthermore, CredBot provides an accessible solution that has the potential to serve as both a practical tool and an educational resource, fostering users' web literacy and critical evaluation skills in online environment where misinformation is prevalent. As one of the first AI-powered credibility assessment tools implemented as a web extension, CredBot contributes to ongoing research in AI-based credibility solutions. It serves as a testing ground for exploring the strengths and limitations of language models in this complex task, paving the way for future advancements in AI-driven credibility systems.

Acknowledgments

We are grateful to professor Orit Shaer and Wellesley Cred Lab's Principal Investigator, Professor Eni Mustafaraj for their invaluable support of this research in numerous ways.

References

[1] Hasan Abu-Rasheed, Mohamad Hussam Abdulsalam, Christian Weber, and Madjid Fathi. 2024. Supporting Student Decisions on Learning Recommendations: An LLM-Based Chatbot with Knowledge Graph Contextualization for Conversational Explainability and Mentoring. *arXiv preprint arXiv:2401.08517* (2024). https://doi.org/10.48550/arXiv.2401.08517 Submitted on 16 Jan 2024 (v1), last revised 24 Jan 2024 (this version, v3).

[2] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American* (May 2001). https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf

[3] DuckDuckGo. 2024. DuckDuckGo Search API. https://api.duckduckgo.com/ Accessed: 2024-10-30.

[4] BJ Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How Do Users Evaluate the Credibility of Web Sites? A Study with Over 2,500 Participants. In *Proceedings of the 2003 Conference on Designing for User Experiences (DUX '03)*. ACM. http://doi.acm.org/10.1145/997078.997097

[5] B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. 2003. How Do Users Evaluate the Credibility of Web Sites?: A Study with Over 2,500 Participants. In *Proceedings of the 2003 Conference on Designing for User Experiences*. Association for Computing Machinery, 1–15. https://doi.org/10.1145/997078.997097

[6] Jaime Gallegos and Others. 2024. Bias and Fairness in Large Language Models: A Survey. *MIT Press Computational Linguistics* 50, 3 (2024), 1097–1132. https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A

[7] David A. Hyman, David Franklyn, Calla Yee, and Mohammad Rahmati. 2017. Going Native: Can Consumers Recognize Native Advertising? Does It Matter? *Yale Journal of Law & Technology* 19 (2017), 77–111.

[8] João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. Weakly Supervised Veracity Classification with LLM-Predicted Credibility Signals. In *Proceedings of the Conference on Weakly Supervised Learning*. The University of Sheffield, Sheffield, United Kingdom.

[9] Media Bias/Fact Check. 2024. Dartmouth Times Bias. https://mediabiasfactcheck.com/dartmouth-times-bias/

[10] Media Bias/Fact Check (MBFC). 2024. Media Bias/Fact Check (MBFC). https://mediabiasfactcheck.com/

[11] Netcraft. 2024. September 2024 Web Server Survey. https://www.netcraft.com/blog/september-2024-web-server-survey/

[12] NewsGuard. 2024. NewsGuard Rating Process and Criteria. https://www.newsguardtech.com/ratings/rating-process-criteria/

[13] Shibani Santurkar, Emre Durmus, Faisal Ladhak, Christina 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)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 29971–30004.

[14] Julia Schwarz and Meredith Morris. 2011. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1245–1254. https://doi.org/10.1145/1978942.1979127

[15] Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Web Retrieval Agents for Evidence-Based Misinformation Detection. In *Proceedings of the COLM 2024 Conference*. COLM. Published as a conference paper at COLM 2024.

[16] Alexander Wan, Eric Wallace, and Dan Klein. 2024. What Evidence Do Language Models Find Convincing? *arXiv preprint arXiv:2402.11782* (2024). https://arxiv.org/abs/2402.11782

[17] Ace Wang, Liz Maylin De Jesus Sanchez, Anya Wintner, Yuanxin Zhuo, and Eni Mustafaraj. 2023. Assessing Google Search's New Features in Supporting Credibility Judgments of Unknown Websites. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval*.

[18] Sam Wineburg, Sarah McGrew, Teresa Ortega, and Joel Breakstone. 2019. *Evaluating Information: The Cornerstone of Civic Online Reasoning*. Technical Report. Stanford Digital Repository. https://purl.stanford.edu/fv751yt5934

[19] World Wide Web Consortium (W3C). 2019. Credibility Signals from CredWeb. https://credweb.org/signals-20191126#h.ltfs3lopw59m

[20] Kai-Cheng Yang and Filippo Menczer. 2024. Accuracy and Political Bias of News Source Credibility Ratings by Large Language Models. *arXiv preprint arXiv:2304.00228* (2024). https://arxiv.org/html/2304.00228v2

[21] Eric Zeng, Kohno Tadayoshi, and Franziska Roesner. 2020. Bad News: Clickbait and Deceptive Ads on News and Misinformation Websites. In *Proceedings of the Workshop on Technology and Consumer Protection (ConPro '20)*.

A Additional Information

Full List of Credibility Signal[Link to source]

45 Benchmark Sites [Link to source]

Demo Video [Link to source]

Received ; revised ; accepted