Real-time Credibility Assessment of Webpages through an LLM-powered Bot

Lan Dau*
Jianfei Lyu*
Emilie Zhang*
ldau@wellesley.edu
jl7@wellesley.edu
ez102@wellesley.edu
Wellesley College
Wellesley, MA, USA

Nina Howley
Wellesley College
Wellesley, USA
nh107@wellesley.edu

Dianna Gonzalez
Wellesley College
Wellesley, USA
dg109@wellesley.edu

Orit Shaer
Wellesley College
Wellesley, USA
oshaer@wellesley.edu

Eni Mustafaraj
Wellesley College
Wellesley, USA
eni.mustafaraj@wellesley.edu

Abstract

In today's digital landscape, discerning credible online content has become increasingly challenging as problematic information proliferates. Despite AI advancements, fully autonomous credibility assessment remains a challenging goal due to persistent struggles with context sensitivity and nuanced credibility indicators. Nevertheless, it is worth investigating whether AI systems can effectively support users in the complex task of credibility assessment–and to what degree. 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. Testing with 5 models, we found that CredBot's inter-rater reliability (IRR) between machine and human evaluators can reach as high as 0.7, highlighting the viability of using LLMs as a foundation for building trustworthy, user-aligned credibility tools.

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.*

Keywords

Credibility Assessment, Conversational User Interfaces (CUI), Automated Credibility Checking, Web Literacy

*All authors contributed equally to this research.

UIST '25, 2025, Busan, Korea

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Lan Dau, Jianfei Lyu, Emilie Zhang, Nina Howley, Dianna Gonzalez, Orit Shaer, and Eni Mustafaraj. 2025. Real-time Credibility Assessment of Webpages through an LLM-powered Bot. In *Proceedings of September 28–October 1 (UIST '25).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The topic of online information credibility has become an increasingly important issue considering the global shift to the online information environment, with studies consistently showing that today's adults have trouble identifying misinformation. For example, research has shown that only 26% of adults and 16% of first-year university students could accurately distinguish factual content from opinion or fake news [15] [20]. Furthermore, research shows that typical users often ignore or lack the awareness to use built-in credibility tools like Google's "About this result" [21]. This highlights not only a gap in critical evaluation skills, but also a lack of engagement with readily available tools, leaving users vulnerable to misinformation.

Several existing tools and resources attempt to address this issue by offering credibility assessments of online content, such as Media Bias/Fact Check (MBFC)[1] or NewsGuard[2]. However, these resources depend on expert human evaluators who have evaluated credibility at the domain level, 9,700 domains for MBFC and 10,000 for NewsGuard, a small minority of the 2 million websites that are active at any given time [16], leaving the vast majority of the web unevaluated for susceptible users. Most importantly, these resources only provide domain level credibility scores or labels, instead of evaluating the specific article a user is reading.

Recent research highlights the potential of LLM-powered bots as educational tools for learning tasks like credibility assessment, due to their ability to provide contextual explanations [1]. Furthermore, studies show that visually augmenting credibility indicators in browsers can improve users' accuracy in assessing credibility

[1]https://mediabiasfactcheck.com/methodology/
[2]https://www.newsguardtech.com/ratings/rating-process-criteria/

[18] [12]. These findings motivate our development of CredBot, a conversational agent powered by LLMs designed to deliver instantaneous credibility evaluations and transparent explanations for online articles within the browsing experience, supporting critical and educational engagement without undermining user agency.

2 Research & Credibility Signals

Early research in information credibility has defined credibility as a multidimensional construct, with dimensions such as believability, accuracy, trustworthiness, bias, and completeness of information commonly used by researchers [14]. Researchers have collected user-based criteria by asking participants to list indicators of what makes an article credible [17]. More recent efforts, such as W3C Crowd Sourced Credibility Signals Draft (W3C) [3] also aim to standardize credibility-related observations into structured signals to enable consistent and interoperable assessments by both humans and machines [22].

For CredBot, we utilized the set of credibility signals for news articles developed by Zhang et al., specifically, the *content signals*, which can be determined by only considering the text or content of an article without accessing external sources or article metadata, rather than their *context signals* [24]. Numerous studies have shown that contextual evaluation is inherently subjective, making the assessment of signals such as "Originality" or "Reputation of Citations" challenging—even for humans—due to the lack of objective or universally accepted standards for such context-dependent signals [19] [7]. Below is a table with Zhang's original eight content signals–six of which are used by CredBot [24]. The selection process is detailed in Section 5 Data Preparation.

Credibility Signal	Definition
Title Representativeness	Titles that are misleading or opaque about the topic, claims, or conclusions of the content.
Clickbait Title	Titles that are designed to entice its readers into clicking an accompanying link.
Quotes from Outside Experts	Additional validation from independent experts in the field.
Citation of Organizations and Studies	Organizations or scientific studies that add context or support to the article to enhance its credibility.
Calibration of Confidence	The use of appropriate language to show confidence in claims.
Logical Fallacies	Misleading readers to poor but tempting arguments.
Tone	Aggregate claims or emotionally charged sections, especially the expressions of contempt, outrage, spite, or disgust.
Inference	Conflating correlation with singular causation or general causation.

Table 1: Definitions for each credibility signal

3 CredBot Features

Building on prior research demonstrating the effectiveness of browser augmentations in improving users' credibility judgments [18], CredBot is implemented as a Chrome extension, which enables it to instantly activate on newly opened browser tabs and parse the webpage the user is currently visiting. This way, CredBot is able to provide credibility assessments and explanations seamlessly across millions of websites—including those that expert human evaluators have not yet been able to assess—directly within the user's browsing environment. We chose to operationalize credibility as a categorical variable with labels such as "questionable" rather than as a numerical variable with a score (e.g., 0–100) to encourage users to remain engaged with the content and reduce over-reliance on a single number. Studies have shown that "differences in perceived credibility" can occur based on differences in one's values in credibility; therefore, a one-size-fits-all approach that provides an opaque "credibility score" cannot adapt to individual needs [24]. Furthermore, numeric scores can create an illusion of precision and may lead users to over-trust the system's judgment, potentially discouraging critical thinking or independent evaluation of the site [13] [7] [6] .

3.1 Credibility Assessment and Explanations

At the core of CredBot's functionality is 1) its ability to assess whether or not a web article is questionable with respect to its credibility and 2) its ability to explain the resulting credibility assessment based on credibility signals, as detailed in Section 2. When a user opens a new website, CredBot automatically activates to determine whether the site is questionable in credibility. If the website is flagged as questionable, a red warning banner appears at the top of the extension to alert users to the potential credibility concerns of the site. Users can then click the "Show" button to view CredBot's assessments of the six credibility signals, each accompanied by a rating of either High or Low and explanations that reference the webpage's content to justify the evaluation (Figure 1). These contextualized explanations are not only meant to increase transparency, but also to serve an educational purpose by helping users understand *why* certain features undermine credibility.

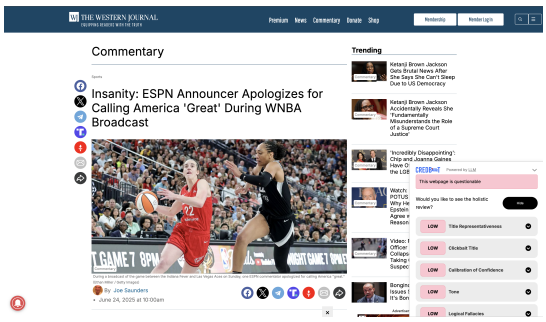

Figure 1: Questionable webpage warning

[3] https://credweb.org/signals-20191126

3.2 Highlighting

Research has found that when a system's inner workings are opaque to the user, it can lead to over-reliance on the output and discourage critical thinking or independent evaluation of the information [5] [2]. To mitigate this concern, on top of the explanations for each of the credibility signals, we provide a highlighting feature that brings users to the specific text on the webpage CredBot used to create the evaluation, and highlights it in yellow. For example, if CredBot gave a Low credibility in "Calibration of Confidence," clicking the highlight icon for the signal will bring the user to the specific place where the author uses overly confident language to state opinion as fact (Figure 2). By explicitly linking evaluations to evidence found in the content itself, CredBot supports the development of users' media literacy and critical thinking skills by training them to recognize and reason about credibility cues independently in future encounters with online information.

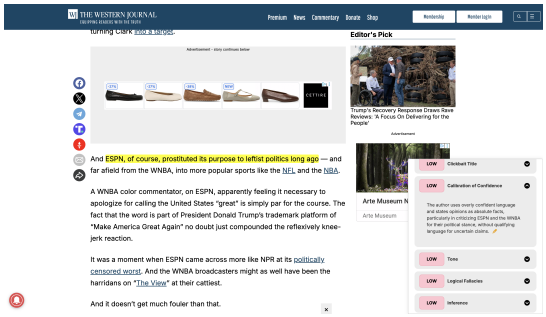

Figure 2: Highlighting evidence

3.3 Chatbot

Lastly, CredBot is designed to allow users to ask questions directly within the interface by typing and entering their questions into the bottom chatbox (Figure 3). AI chatbots have been found to effectively enhance digital dialogue for students in numerous educational settings by providing contextual information and understandable explanations [3] [23] [1].

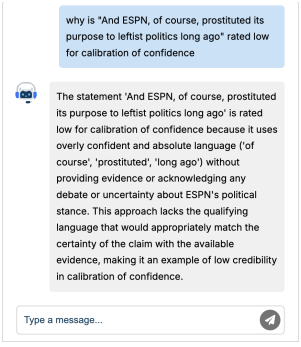

Figure 3: Chatbot

This gives users a chance to inquire about specific aspects of the credibility evaluations within the article's context, request detailed definitions on the credibility criteria used, as well as any other questions they may have, and get real-time responses. This encourages users to actively engage with the reasoning behind each signal, turning passive consumption into an interactive learning experience.

4 Architecture

4.1 System Design

Shown in Figure 4, CredBot's system contains a separate frontend and backend architecture, which work together to deliver an article's credibility assessment to the user. The frontend scrapes HTML data from the webpage before making retrieval requests to the backend and delivering information to the UI. The backend works using three components which are outlined below.

(1) Endpoint router: handles incoming requests from the frontend and triggers the appropriate module to ensure the correct response structure
(2) Web parser: extracts textual information from a given article, significantly reducing the number of tokens passed to the LLM .
(3) LLM Gateway: picks the appropriate prompt template and calls the LLM while storing all context in Messages History

4.2 System Prompt

The prompts are designed using a template with a persona description, a widely adopted method in prompt engineering [8]. The prompt emphasizes maintaining a neutral political stance, and instructs the LLM to evaluate the credibility of each signal as either High or Low based on the provided website text and credibility signal definitions.

Each credibility signal is defined in three parts: first is a brief, question-based definition of the signal in the format of "what to look for", followed by detailed criteria for both High and Low credibility ratings. The detailed criteria are specific, observable features from W3C. For the full prompt, see Appendix 1.

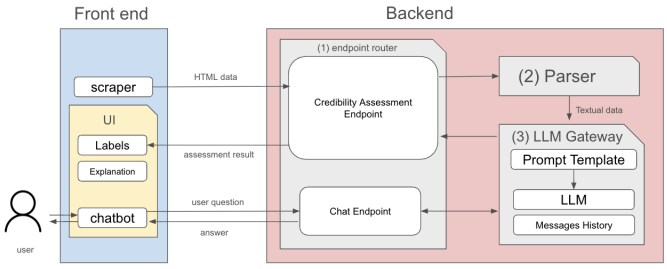

Figure 4: Overview of CredBot system design

5 Data Preparation

The dataset of articles used for testing and training CredBot is derived from domains rated by MBFC. An even number of articles from five biases (left, center, right, fake-news, and conspiracy-pseudoscience) and three credibility levels (high, medium and low) are collected and parsed. However, because the credibility levels assigned by MBFC are at the domain level, four human labelers from the Wellesley Credibility Lab research team evaluated a subset of 51 articles by individually scoring them on the 8 credibility signals.

To ensure unbiased evaluation, each article is randomly assigned to a pair of two labelers, with article assignments made by the alternate pair to prevent selection bias. All scoring is conducted blindly. During the labeling process, labelers report significant subjectivity challenges with "Citation of Organizations and Studies",

finding it difficult to consistently evaluate the quality and relevance of sources, as this often requires domain expertise to assess source credibility. For "Quotes from Outside Expert," labelers struggled to determine whether quoted material was appropriately contextualized or represented, leading to inconsistent scoring patterns. These signals were removed from the final signal set to maintain annotation reliability and ensure consistent ground truth labels.

Inter-rater reliability (IRR) is measured using Cohen's kappa, yielding scores of 0.89 and 0.83 for the two labeler pairs respectively, indicating substantial to near-perfect agreement according to standard interpretation guidelines. These high reliability scores validate the quality of human annotations as ground truth labels. Disagreements between labeler pairs are resolved through consensus discussions to produce the final dataset used for model evaluation.

6 Signal Validation

We tested CredBot's architecture with five LLM models: GPT-4o, GPT-4o mini, o3-mini, Qwen3, and DeepSeek-V3. These models were chosen to test a range of cost, accessibility and efficiency. To assess each LLM's agreement with the ground truth, we calculated the IRR score using Cohen's kappa. The results in Table 2 show moderate agreement for GPT-4o-mini, and substantial agreement for GPT4o, o3-mini, Gwen3 and DeepSeek-V3. These agreement rates validate that LLMs can correctly interpret credibility signals within the same analytical frameworks used by labelers, and highlight open source models' potential as capable alternatives to commercially available LLMs.

Model	Human Agreement	IRR
GPT-4o-mini	70.26%	0.55
GPT-4o	78.10%	0.67
o3-mini	81.70%	0.67
Qwen3	81.70%	0.68
DeepSeek-V3	82.68%	0.70

Table 2: Percent agreement and IRR for each tested model between human labels and model labels

Qualitative examination of the LLMs' explanations reveals that they generally produce quality explanations. The models are able to identify signals under given definitions and provide systematic, evidence-based reasoning by referencing specific textual details. For example, o3-mini assigns High credibility for Clickbait Title in an article titled "Mason County Commission honors local track and field champions," with the explanation "The title clearly and informatively states the subject and event without vague or provocative language", aligning with labeler's evaluation of the article as factual local news. In another article, Deepseek-V3 assigns Low credibility for signal Calibration of Confidence with the explanation "The article contains statements with overly confident language ("undoubtedly," "definitely") regarding complex issues that are still debated, without sufficient qualifying language for uncertain claims," which correctly identifies signifiers also noted by our labelers.

However, review of disagreements reveals several weaknesses, highlighted in Table 3.

7 Conclusion and Future Work

CredBot currently focuses on traditional web media rather than social media, due to its unique challenges such as user-generated content, limited source information, and rapid content turnover. These differences demand distinct credibility signals that can vary widely not only between social and web media, but also between different social media platforms [4] [10] [11] [9]. Similarly, the content signals developed by Zhang et al. [24] are better tailored to news articles than to opinion pieces; for example, requirements like citing sources for quotes are not directly applicable to opinion-based journalism. In our future work, we plan to develop a distinct set of credibility signals tailored to opinion-based articles, and explore credibility assessment in the context of social media.

While we limited CredBot's evaluation criteria to textual content in this iteration, we plan to work on incorporating visual content, as well as to refine the highlighting feature and to test CredBot on more models such as Claude and Gemini. We also plan to incorporate a fact-checking feature that provides reputable sources to encourage independent research by simulating the verification process for users. This iteration of CredBot incorporates user interface (UI) decisions and features, such as CredBot's size, based on feedback from a preliminary user study. However, we aim to run user studies to evaluate CredBot's usability, perceived trustworthiness, and interface design, as well as implement a built-in feedback feature to create a user-driven feedback loop.

CredBot demonstrates the potential of LLMs to perform article-level credibility assessments in a scalable and cost-effective manner. By automating evaluations typically performed by human experts and supporting compatibility with open-source models, CredBot offers a practical alternative designed to assist and educate users rather than replace human judgment. As one of the first LLM-powered browser extensions for credibility assessment, CredBot achieves up to 0.7 in IRR with our human evaluators, laying groundwork for future research on AI-driven solutions for online credibility and illustrating the potential of modern LLMs to interpret complex credibility signals effectively.

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. We are grateful to Jenny Long and Alexa Halim for their support in the very first prototype of CredBot in the fall of 2023, as well as Jenny Long's design for CredBot's logo. We are grateful for Malika Parkhomchuk's efforts on incorporating visual content into CredBot's credibility criteria.

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

[2] Muhammad Aljukhadar, Sylvain Senecal, and Charles-Etienne Daoust. 2022. The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Information Management* 59, 8 (2022), 103694.

Signal	Title Representativeness	Clickbait Title	Tone
Summary	Conflation: model conflates topical accuracy with title representativeness	Leniency: model assigns credible label despite acknowledging title's dramatic nature	Conflation: models fail to discern quotes from author's language
Input	"Stop Netanyahu Before He Gets Us All Killed"	"Gullible Europe has signed the death warrant for its own car industry – The Government has made its mind up that it's not going to abandon the all-electric dogma"	"'Harvard University brazenly refused to provide the required information requested and ignored a follow up request from the Department's Office of General Council,' DHS said."
LLM Label	High	High	Low
LLM Explanation	The title accurately reflects the article's content, which discusses Netanyahu's actions and their potential global consequences, supported by historical evidence and analysis.	The title clearly establishes the topic: the demise of Europe's car industry due to government decisions on all-electric policies. While dramatic, it specifies who is affected and what is happening, avoiding vague clickbait phrasing.	The article uses exaggerated claims and emotional language, particularly in quotes from DHS, which describe Harvard in extremely negative terms and use polarizing language.
Human Label	Low	Low	High
Human Explanation	Title provides no indication of what specific policies, actions, or consequences the article will address.	Given the strong language and intention of provoking controversy, the title is considered clickbait.	Quotes that the article cites are dramatic due to the subject's sensitive nature, but the tone of reporting the event is consistently objective.

Table 3: Examples of LLMs weaknesses in assessing credibility signals

[3] M. S. Alwazzan. 2024. Investigating the Effectiveness of Artificial Intelligence Chatbots in Enhancing Digital Dialogue Skills for Students. *European Journal of Educational Research* 13, 2 (2024), 573–584. https://www.eu-jer.com/investigating-the-effectiveness-of-artificial-intelligence-chatbots-in-enhancing-digital-dialogue-skills-for-students

[4] Monika Choudhary, Satyendra Singh Chouhan, and Santosh Singh Rathore. 2024. Beyond Text: Multimodal Credibility Assessment Approaches for Online User-Generated Content. *ACM Trans. Intell. Syst. Technol.* 15, 5 (2024). https://doi.org/10.1145/3673236

[5] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[6] Jeffrey A. Friedman, Jennifer S. Lerner, and Richard Zeckhauser. 2017. Behavioral Consequences of Probabilistic Precision: Experimental Evidence from National Security Professionals. *International Organization* 71, 4 (2017), 803–826.

[7] Thomas R. Guskey. 2013. The Case Against Percentage Grades. *Educational Leadership* 71, 1 (2013), 68–72.

[8] Chenyu Hou, Gaoxia Zhu, Juan Zheng, Lishan Zhang, Xiaoshan Huang, Tianlong Zhong, Shan Li, Hanxiang Du, and Chin Lee Ker. 2024. Prompt-based and Fine-tuned GPT Models for Context-Dependent and -Independent Deductive Coding in Social Annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)*. ACM, 518–526. https://doi.org/10.1145/3636555.3636910

[9] Eva L. Jenkins et al. 2020. Assessing the Credibility and Authenticity of Social Media Content for Applications in Health Communication: Scoping Review. *Journal of Medical Internet Research* 22, 7 (2020). https://doi.org/10.2196/17296

[10] H. Keshavarz. 2021. Evaluating Credibility of Social Media Information: Current Challenges, Research Directions and Practical Criteria. *Information Discovery and Delivery* 49, 4 (2021), 269–279. https://doi.org/10.1108/IDD-03-2020-0033

[11] Raynard S. Kington et al. 2021. Identifying Credible Sources of Health Information in Social Media: Principles and Attributes. *NAM Perspectives* 2021 (2021), 10.31478/202107a. https://doi.org/10.31478/202107a Published 16 Jul. 2021.

[12] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287590

[13] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. ACM, 29–38. https://doi.org/10.1145/3287560.3287590 Published 29 January 2019.

[14] Miriam J. Metzger, Andrew J. Flanagin, Keren Eyal, Daisy R. Lemus, and Ronald M. McCann. 2003. Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. *Annals of the International Communication Association* 27, 1 (2003), 293–335. https://doi.org/10.1080/23808985.2003.11679029

[15] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Nami Sumida. 2018. *Distinguishing Between Factual and Opinion Statements in the News*. Report. Pew Research Center. https://www.pewresearch.org/journalism/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news/

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

[17] Soo Young Rieh and David R. Danielson. 2007. Credibility: A Multidisciplinary Framework. *Annual Review of Information Science and Technology* 41, 1 (2007), 307–364. https://doi.org/10.1002/aris.2007.1440410114

[18] 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

[19] Daniel Starch and Edward C. Elliott. 1912. Reliability of the Grading of High-School Work in English. *The School Review* 20, 7 (1912), 442–457. http://www.jstor.org/stable/1076706 Accessed 24 Jan. 2025.

[20] Albie van Zyl, Marita Turpin, and Machdel Matthee. 2020. How Can Critical Thinking Be Used to Assess the Credibility of Online Information? In *Responsible Design, Implementation and Use of Information and Communication Technology*. Springer, 199–210. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7134292/

[21] 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*.

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

[23] Rui Wu, Zhi Liu, Xiaomin Fan, and Haoran Xie. 2023. Do AI chatbots improve students' learning outcomes? A meta-analysis. *British Journal of Educational Technology* 54, 1 (2023), 130–154.

[24] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer 8. Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion Proceedings of the Web Conference 2018*. ACM, Lyon, France, 603–612. https://doi.org/10.1145/3184558.3188731

A Additional Information

(1) System Prompt: [Link to source]

Received ; revised ; accepted