Veracity: An Open-Source AI Fact-Checking System

Taylor Lynn Curtis[1] , **Maximilian Puelma Touzel**[1,3] , **William Garneau**[4] , **Manon Gruaz**[1] , **Mike Pinder**[1] , **Li Wei Wang**[2] , **Sukanya Krishna**[5,6] , **Luda Cohen**[6] , **Jean-François Godbout**[1,3] , **Reihaneh Rabbany**[1,2] and **Kellin Pelrine**[1,2]
[1]Mila - Quebec AI Institute
[2]McGill University
[3]Université de Montréal
[4]Nord AI
[5]Harvard University
[6]Supervised Program for Alignment Research (SPAR)
{taylor.curtis, jean-francois.godbout, reihaneh.rabbany, kellin.pelrine}@mila.quebec

Abstract

The proliferation of misinformation poses a significant threat to society, exacerbated by the capabilities of generative AI. This demo paper introduces Veracity, an open-source AI system designed to empower individuals to combat misinformation through transparent and accessible fact-checking. Veracity leverages the synergy between Large Language Models (LLMs) and web retrieval agents to analyze user-submitted claims and provide grounded veracity assessments with intuitive explanations. Key features include multilingual support, numerical scoring of claim veracity, and an interactive interface inspired by familiar messaging applications. This paper will showcase Veracity's ability to not only detect misinformation but also explain its reasoning, fostering media literacy and promoting a more informed society.

1 Introduction

Experts have rated the dissemination of misinformation and disinformation as the #1 risk the world faces [Torkington, 2024]. This risk has only increased with the proliferation and advancement of generative AI [Bowen *et al.*, 2024; Pelrine *et al.*, 2023b]. Responses to misinformation have up to now been largely centred around platform moderation. As large-scale social media platforms actively eliminate their content moderation teams [Horvath *et al.*, 2025], they pass to the user the personal and social responsibility to assess the reliability of claims and figure out how to make well-grounded decisions in a landscape of uncertain information. In the absence of strong platform-based approaches, solutions that support and empower individuals with tools to validate the information they encounter become essential in dampening the societally corrosive effects of misinformation.

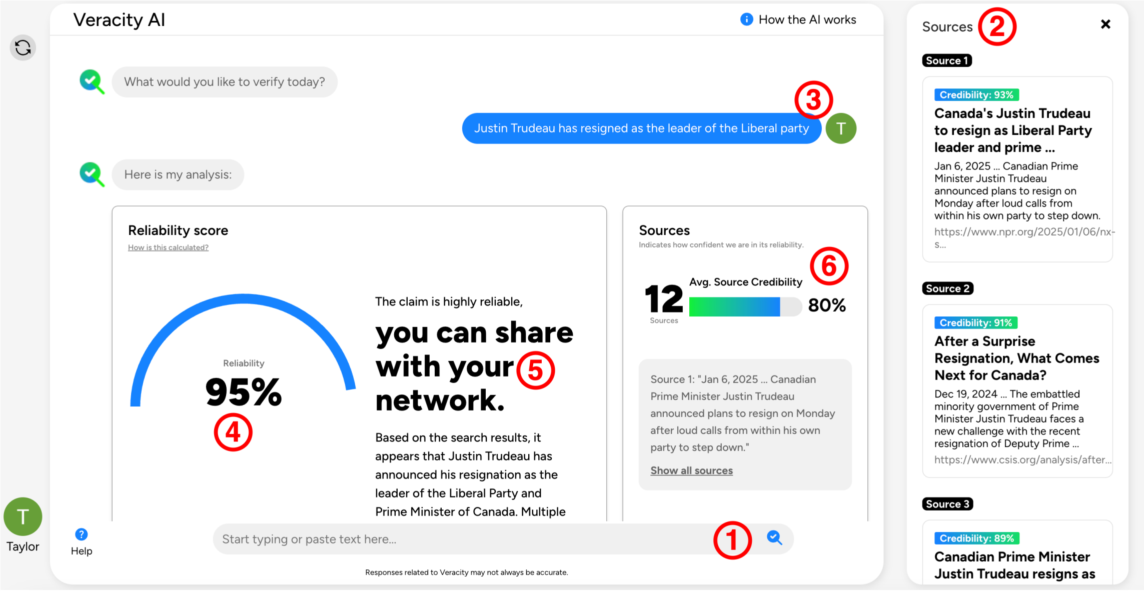
Misinformation is particularly dangerous when it influences public health and democratic processes, as seen in the spread of vaccine-related disinformation and politically motivated claims about censorship, both of which have been shown to exacerbate real-world harm and undermine trust in institutions [Lewandowsky, 2025]. With the rollback of content moderation efforts and increasing concerns over algorithmic bias on social media platforms, independent, reliable fact-checking tools are more necessary than ever.

A promising solution in this area is an AI Steward that helps people fact-check and filter out manipulative and fake information. In fact, AI can outperform human fact-checkers in both accuracy [Wei *et al.*, 2024; Zhou *et al.*, 2024] and helpfulness [Zhou *et al.*, 2024]. Although there is rapid progress in improving the accuracy of such systems [Tian *et al.*, 2024; Wei *et al.*, 2024; Ram *et al.*, 2024], there is much less research on how to make a high-accuracy system into a helpful and trustworthy one that users can rely on [Augenstein *et al.*, 2024]. Our AI-powered open-source solution, **Veracity**, deploys large language models (LLMs) working with web retrieval agents to provide any member of the public with an efficient and grounded analysis of how factual their input text is. Moreover, through open-sourcing our platform, we hope to bring a test-bed for the research community to design effective fact-checking strategies.

Problem Setting Our society needs tools that support information integrity by defending against rampant misinformation. Individuals currently face the challenge of combatting disinformation largely on their own. Individuals face a lack of 'good' information, and also difficulty in reliably finding information from credible sources to justify whether or not a statement in question is true or false. Tools that help individual users address this challenge exist, but are either proprietary, in which case there are access, transparency and privacy issues, or are limited in their ease of use.

Proposed Solution We propose a fact-checking system solution that uses a Large Language Model (LLM) to summarize relevant text retrieved by a web agent from reliable sources on the internet. The solution was designed to address the following goals related to information integrity:

- Counter misinformation by providing accurate, evidence-based assessments.



What would you like to verify today?

Justin Trudeau has resigned as the leader of the Liberal party

Here is my analysis:

Reliability score

How is this calculated?

95%

The claim is highly reliable,

you can share with your network.

Based on the search results, it appears that Justin Trudeau has announced his resignation as the leader of the Liberal Party and Prime Minister of Canada. Multiple

Sources

Indicates how confident we are in its reliability.

12 Sources

Avg. Source Credibility 80%

Source 1: "Jan 6, 2025 ... Canadian Prime Minister Justin Trudeau announced plans to resign on Monday after loud calls from within his own party to step down."

Show all sources

Start typing or paste text here...

Responses related to Veracity may not always be accurate.

Source 1

Credibility: 93%

Canada's Justin Trudeau to resign as Liberal Party leader and prime ...

Jan 6, 2025 ... Canadian Prime Minister Justin Trudeau announced plans to resign on Monday after loud calls from within his own party to step down.

https://www.npr.org/2025/01/06/nx-s...

Source 2

Credibility: 91%

After a Surprise Resignation, What Comes Next for Canada?

Dec 19, 2024 ... The embattled minority government of Prime Minister Justin Trudeau faces a new challenge with the recent resignation of Deputy Prime ...

https://www.csis.org/analysis/after...

Source 3

Credibility: 89%

Canadian Prime Minister Justin Trudeau resigns as

Figure 1: The main fact-checking page of Veracity

- Foster media literacy by helping users critically evaluate online claims.
- Promote transparency by explaining why a claim is assessed as true or false.
- Ensure broad accessibility, making fact-checking tools open-source and available to anyone.

The system is targeted at:

- The general public, including both tech-savvy users and those less familiar with new technologies.
- Expert users, such as journalists and professional fact-checkers, who will have access to an expert dashboard.

Our Contribution This paper describes the design and functionality of an open-source, claim-focused fact-checking system that is designed to enhance transparency in model decision-making. We detail its application domain, technical architecture, AI techniques, and interactive elements. Unlike traditional black-box models, our applications allows users to submit claims and receive structured responses that provide clear analysis on how reasoning was done to reach the veracity decisions. With a strong emphasis on open research, our system is built to be fully accessible to anyone, so anyone can download the application and run it locally. This is important for ensuring reproducibility and collaboration/feedback from the community. Key features include multilingual support, a numerical scoring for claim veracity, and we also demonstrate how this tool addresses misinformation by developing an intuitive, transparent platform for claim verification.

2 System Description

System Overview The main functionality of the system can be seen in Figure 1. This is the system's main page, where the user is taken immediately upon logging in. The behaviour of each part of the interface is described by the numerical mappings shown in Figure 1:

1. **Claim submission box**: This box is where the user can type or copy/paste the claim they want the AI to verify.
2. **Sources panel**: When a claim is submitted, the LLM will (if it decides it is necessary) use a web agent to retrieve sources; all of the sources used will be displayed here.
3. **Claim under analysis:** After a user submits a claim, it is displayed on the screen.
4. **Reliability score:** This is the score generated by the LLM that reflects the reliability of the claim, where 0% maps to completely unreliable or false and 100% maps to completely reliable or true.
5. **Textual instruction per reliability score & LLM explanation**: The user is shown an actionable message that interprets the model's veracity score and a share recommendation (the score must be greater than 60% for a positive recommendation). Below this is the LLM reasoning that explains its reliability score.
6. **Source summary**: This includes aggregate information about the sources used to determine the reliability of the claim, including the number of sources and the average credibility ranking of the sources.

Technology Stack The system is divided into separate frontend and backend tech stacks, with the frontend being served by HTTPS requests to an application programming interface (API). The frontend and backend exist separately, except for the API that forms a contract between the two.

Frontend The web display, or visualization of the application, was implemented using Next.js [Vercel, 2025] and deployed using the Vercel deployment pipeline within the package. The frontend also uses Sass, Typescript, and Chart.js [Chart.js Contributors, 2025]. For complete documentation on the frontend technology stack, please see the frontend project wiki [link].

Backend The backend, encompassing the application logic and the persistence (i.e. database) layers is deployed using the Google Cloud Platform (GCP). The application logic or API is deployed on Kubernetes, and the database is deployed on Cloud SQL [Cloud, 2025b; Cloud, 2025a]. Beyond deployment, the API is designed using FastAPI [Ramírez, 2025], the database is implemented in PostgreSQL [Group, 2025], and the object mapping between the API and the database is managed by SQLAlchemy [Bayer, 2025]. For full documentation on this tech stack, please see the backend project wiki [link].

3 AI Techniques and Innovations

Core AI Methods This system uses AI to power its fact-checking methodology, specifically LLM technology. Despite the challenges of misinformation detection, including the tendency of misinformation to contain a mix of both true and false information, LLMs have been shown to be effective tools for detecting misinformation online [Pelrine *et al.*, 2023a; Chen and Shu, 2024]. However, LLMs alone may not be enough. Many studies have shown the benefits of retrieving information from online sources to improve the performance of fact-checking and misinformation detection [Bekoulis *et al.*, 2021; Kondamudi *et al.*, 2023; Zhou and Zafarani, 2020].

To achieve this goal, the system is an implementation of the LLM/web search engine teaming proposed by Tian et al. [2024] in their *Web Retrieval Agents for Evidence-Based Misinformation Detection*. The interactions between the LLM, web search engine, and user are described by Figure 2.

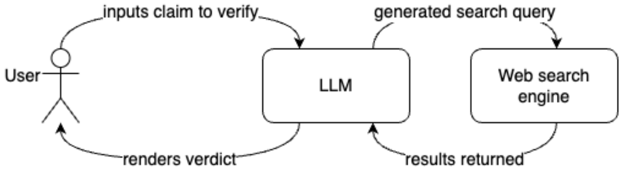


Figure 2: The interactions between LLM, search engine, and user

Innovations This system represents a unique innovation and application of AI in the fact-checking space. In addition to the teaming of web search agents and LLM reasoning as described above, this system has a couple of important innovations or distinctions from other AI fact-checking systems. In particular, this is due to a few important features:

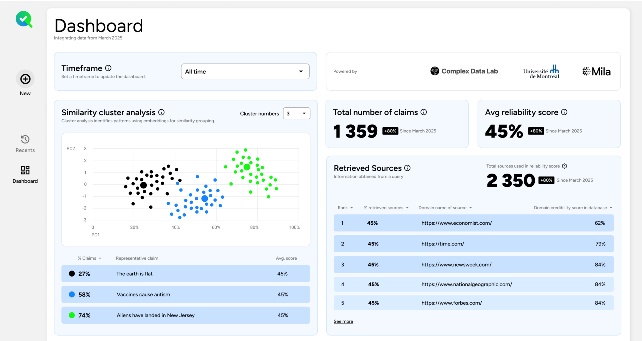
- **Sources display**: Not only does the system use a search engine to select relevant sources, the LLM is shown these sources to help it render its verdict. The user is also shown sources, and their documented credibility [Lin *et al.*, 2023].
- **Score-based analysis**: This is the first tool of its kind to present a reliability score that represents the factuality of a user's claim and to ask the LLM to justify this score.

4 Interactive Elements

The system was designed to invoke a feeling of familiarity and trust from all users, while prioritizing the interactivity of the system. The modalities of interaction, as well as the central interface, were inspired by standard messaging applications (WhatsApp, Messenger, etc.). In addition to the main interaction (a user submits a claim and reviews the result), the extra interactions are outlined in this section.

Collection of User Feedback The system is designed to enable continuous improvement. This is done by user feedback. The feedback mechanism is user-driven and is specific to the system's analysis of a particular claim. The user can select a rating between 1 and 5 stars to reflect how well the model analyzed their claim. Following this selection, the user can select a series of 'tags' or small textual snippets that reflect different functionalities of the system, such as the sources. They may also submit an optional comment.

Expert Dashboard The system is also unique in that it is not just designed for users to fact-check relevant claims. Registered users who identify themselves as experts, and are approved by the system administration, will have access to a fact-checking expert dashboard. This dashboard is designed to display aggregate information from the application to these users. For example, it displays a clustering graph of the most common trends in claims submitted to the system.


Dashboard
Integrating data from March 2025

Timeframe ⓘ
Set a timeframe to update the dashboard.
All time

Powered by Complex Data Lab Université de Montréal Mila

Similarity cluster analysis ⓘ
Cluster analysis identifies patterns using embeddings for similarity grouping.
Cluster numbers 3

% Claims	Representative claim	Avg. score
● 27%	The earth is flat	45%
● 58%	Vaccines cause autism	45%
● 74%	Aliens have landed in New Jersey	45%

Total number of claims ⓘ
1 359 +80% Since March 2025

Avg reliability score ⓘ
45% +80% Since March 2025

Retrieved Sources ⓘ
Information obtained from a query
Total sources used in reliability score ⓘ
2 350 +80% Since March 2025

Rank	% retrieved sources	Domain name of source	Domain credibility score in database
1	45%	https://www.economist.com/	62%
2	45%	https://time.com/	79%
3	45%	https://www.newsweek.com/	84%
4	45%	https://www.nationalgeographic.com/	84%
5	45%	https://www.forbes.com/	84%

See more

Figure 3: Snippet of the expert dashboard

5 Conclusion

This demo has showcased Veracity, an open-source AI system that combines LLMs and web retrieval agents to provide transparent and accessible fact-checking. While AI systems employing LLMs and web retrieval for fact-checking exist, open-source versions are not readily available. Veracity aims to fill this gap by providing a production ready factuality assessment application, with intuitive explanations, i) empowering individuals to critically evaluate information and contribute to a more informed society; ii) empowering the research community to expand the system's capabilities and build the next generation of AI-powered fact-checking systems. Future work includes improving the handling of complex claims, improving user interaction features, broadening language and context support, and more advanced credibility measurement techniques. Veracity's open-source[1] nature encourages community involvement and further development to address the ongoing challenge of misinformation.

[1] Links: frontend GitHub repository, backend GitHub repository

Acknowledgements

This project has been funded through the Canada CIFAR AI Chair program, Digital Citizen Contribution Program, and Mila's Technology Maturation Grant.

Author Contributions

Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine contributed equal advising to the project.

References

[Augenstein *et al.*, 2024] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863, Aug 2024.

[Bayer, 2025] Mike Bayer. Sqlalchemy: The database toolkit for python, 2025.

[Bekoulis *et al.*, 2021] Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. A review on fact extraction and verification. *ACM Computing Surveys (CSUR)*, 55(1):1–35, 2021.

[Bowen *et al.*, 2024] Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Data poisoning in llms: Jailbreak-tuning and scaling laws. *arXiv preprint arXiv:2408.02946*, 2024.

[Chart.js Contributors, 2025] Chart.js Contributors. Chart.js: Simple yet flexible javascript charting library, 2025.

[Chen and Shu, 2024] Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368, 2024.

[Cloud, 2025a] Google Cloud. Cloud sql documentation, 2025.

[Cloud, 2025b] Google Cloud. What is kubernetes?, 2025.

[Group, 2025] PostgreSQL Global Development Group. Postgresql: The world's most advanced open source relational database, 2025.

[Horvath *et al.*, 2025] Bruna Horvath, Jason Abbruzzese, and Ben Goggin. Meta is ending its fact-checking program in favor of a 'community notes' system similar to X's. https://www.nbcnews.com/tech/social-media/meta-ends-fact-checking-program-community-notes\protect\penalty\z@-x-rcna186468, January 2025. [Accessed 03-02-2025].

[Kondamudi *et al.*, 2023] Medeswara Rao Kondamudi, Somya Ranjan Sahoo, Lokesh Chouhan, and Nandakishor Yadav. A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. *Journal of King Saud University-Computer and Information Sciences*, 35(6):101571, 2023.

[Lewandowsky, 2025] Stephan Lewandowsky. Free speech, fact checking, and the right to accurate information. *Science*, 387(6734), 2025.

[Lin *et al.*, 2023] Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. High level of correspondence across different news domain quality rating sets. *PNAS Nexus*, 2(9):pgad286, 09 2023.

[Pelrine *et al.*, 2023a] Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*, 2023.

[Pelrine *et al.*, 2023b] Kellin Pelrine, Mohammad Taufeeque, Michał Zajac, Euan McLean, and Adam Gleave. Exploiting novel gpt-4 apis. *arXiv preprint arXiv:2312.14302*, 2023.

[Ram *et al.*, 2024] Ashwin Ram, Yigit Ege Bayiz, Arash Amini, Mustafa Munir, and Radu Marculescu. Credirag: Network-augmented credibility-based retrieval for misinformation detection in reddit. *arXiv preprint arXiv:2410.12061*, 2024.

[Ramírez, 2025] Sebastián Ramírez. Fastapi, 2025.

[Tian *et al.*, 2024] Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-Francois Godbout, Reihaneh Rabbany, and Kellin Pelrine. Web retrieval agents for evidence-based misinformation detection. *arXiv preprint arXiv:2409.00009*, 2024.

[Torkington, 2024] Simon Torkington. These are the 3 biggest emerging risks the world is facing. https://www.weforum.org/stories/2024/01/ai-disinformation-global-risks/, January 2024. [Accessed 02-02-2025].

[Vercel, 2025] Vercel. Next.js, 2025.

[Wei *et al.*, 2024] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*, 2024.

[Zhou and Zafarani, 2020] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[Zhou *et al.*, 2024] Xinyi Zhou, Ashish Sharma, Amy X Zhang, and Tim Althoff. Correcting misinformation on social media with a large language model. arxiv [preprint](2024), 2024.