Aletheia: Detect, Discuss, and Stay Informed on Fake News

Dorsaf Sallami , **Esma Aïmeur**
Department of Computer Science and Operations Research (DIRO), Université de Montréal, Canada
dorsaf.sallami@umontreal.ca, aimeur@iro.umontreal.ca

Abstract

In today's digital era, the rapid spread of fake news undermines both social unity and democratic institutions, demanding effective countermeasures. Current browser extensions to counter fake news have significant limitations, such as opaque models, dependency on traditional Machine Learning (ML) techniques, lack of explanatory features, and limited focus on detection without user engagement support. This paper introduces *Aletheia*, a novel browser extension that addresses these shortcomings by leveraging Retrieval Augmented Generation (RAG) and Large Language Models (LLMs) to enhance fake news detection and provide evidence-based explanations. Additionally, *Aletheia* incorporates two key components: a Discussion Hub, enabling users to discuss instances of fake news, and a Stay Informed feature, which displays the latest fact-checks. *Aletheia* surpasses state-of-the-art methods according to experimental results.

1 Introduction

Fake news consists of information intentionally designed to deceive, manipulate, or misinform specific audiences, often spreading rapidly through viral dissemination [Sallami and Aïmeur, 2025]. Its proliferation has significantly undermined social cohesion and democratic processes, raising concerns among various stakeholders [Balakrishnan *et al.*, 2022].

Both researchers and platform operators are investing more in combating fake news, primarily through ML models for detection [Sallami *et al.*, 2023; Amri *et al.*, 2021]. However, these models often lack user-friendly solutions, creating a gap between advanced research and practical tools. Browser extensions have emerged as a promising way to bridge this gap, integrating sophisticated ML methods into users' browsing experiences for real-time detection. Despite their potential, existing extensions face key limitations: they often lack transparency, rely on traditional ML, provide results without explanations, and focus solely on detection without additional features [Sallami and Aïmeur, 2025; Moalla *et al.*, 2025].

To address these limitations, we propose *Aletheia*,[1] a novel browser extension for fake news detection. Our solution advances the field by (1) leveraging RAG and LLMs to detect fake news while providing evidence-based explanations retrieved via RAG-powered Google searches. These explanations enhance transparency and user trust by justifying the model's judgments. Additionally, our extension includes (2) a Discussion Hub, enabling users to post and comment on fake news instances, fostering community engagement and collaborative analysis, and (3) a Stay Informed feature that displays the latest fact-checks from the Google Fact Check API.

2 Related Works

Browser extensions have become vital tools in combating fake news, helping users evaluate online content reliability. Notable examples available in the Google Chrome Extension Store include The Fact Checker [Checker, 2024] and Media Bias Fact Check [Check, 2024]. However, their methodologies remain opaque, as no peer-reviewed research or technical documentation clarifies their underlying models, raising concerns about reproducibility and efficacy.

Alternative approaches typically involve flagging content from known fake news sources using continuously updated lists, as shown by extensions like B.S. Detector [Detector, 2024], which poses practical challenges due to the frequent need for list updates. Similarly, tools like FactIt [Velasco *et al.*, 2023] rely on traditional machine learning techniques such as Logistic Regression, predating LLM advances. These methods struggle to adapt to the evolving complexity of fake news, yielding marginal improvements in detection accuracy [Kuntur *et al.*, 2024]. More advanced solutions, including Check-It [Paschalides *et al.*, 2021], TrustyTweet [Hartwig and Reuter, 2019], BRENDA [Botnevik *et al.*, 2020], and ShareAware [von der Weth *et al.*, 2020], leverage deep neural network algorithms for fake news detection. Despite these advancements, our review reveals that no existing browser extension currently employs LLMs for this purpose,

[1]The name *Aletheia* ($\alpha\lambda\eta\vartheta\varepsilon\iota\alpha$) is an ancient Greek word that is commonly translated as "truth" [Woleński, 2004]. The name reflects the tool's purpose to expose fake news, aligning with its philosophical roots in truth-seeking.

even though LLMs surpass these models in detection performance [Kuntur *et al.*, 2024].

Another key limitation is the lack of explainability. While tools like COVID-FakeExplainer [Warman and Kabir, 2023] use SHAP (SHapley Additive exPlanations) [Lundberg and Lee, 2017] to justify predictions, these technical details often confuse non-expert users. This gap between algorithmic output and user understanding undermines trust and calls for more intuitive, human-centric interfaces [Epstein *et al.*, 2022].

Moreover, existing extensions primarily focus on detection, lacking interactive features. Our work bridges this gap by introducing a browser extension that integrates LLM-driven detection with two key features: (1) a discussion hub for user engagement on disputed content and (2) real-time fact-checking updates. By combining automated analysis with community-driven discourse, our solution goes beyond detection to promote digital literacy and critical thinking.

3 System Design

Aletheia adopts a client-server architecture comprising a browser extension frontend and a Python Flask backend. The overall system architecture is depicted in Figure 1.

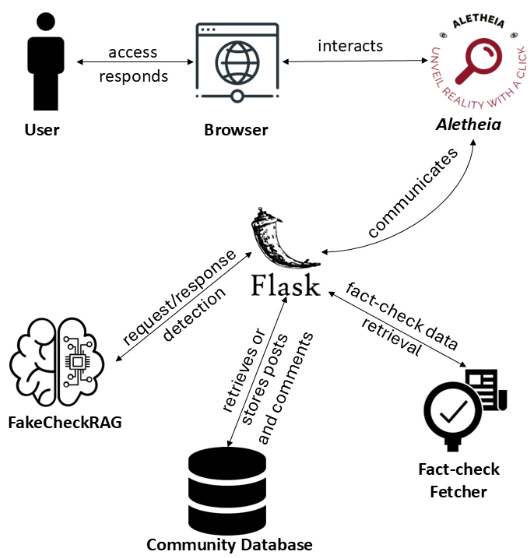

Figure 1: System architecture.

3.1 Frontend: Browser Extension

The frontend of *Aletheia* is a browser extension compatible with Google Chrome. When a user activates the fact-checking feature by clicking the extension icon, JavaScript modules extract relevant information from the current web page and send a query to the backend server. Upon receiving the server's response, another JavaScript module processes and displays the fact-checking results to the user.

Figure 2 illustrates the interface of *Aletheia*. When the user launches *Aletheia*, a popup appears, allowing them to select the desired functionality. Users can utilize the 'Verify It' component by entering news content and clicking the 'Detect' button to view the results, as shown in Figure 2(a). Afterwards, they can click on 'Show Explanation' to understand the reasoning behind the detection, as depicted in Figure 2(b). In addition to detection, users can engage in discussions through the 'Discussion Hub' component, facilitating conversations about fake news, as illustrated in Figure 2(c). Lastly, users can stay up to date with the latest fact-checks via the 'Stay Informed' component, as shown in Figure 2(d).

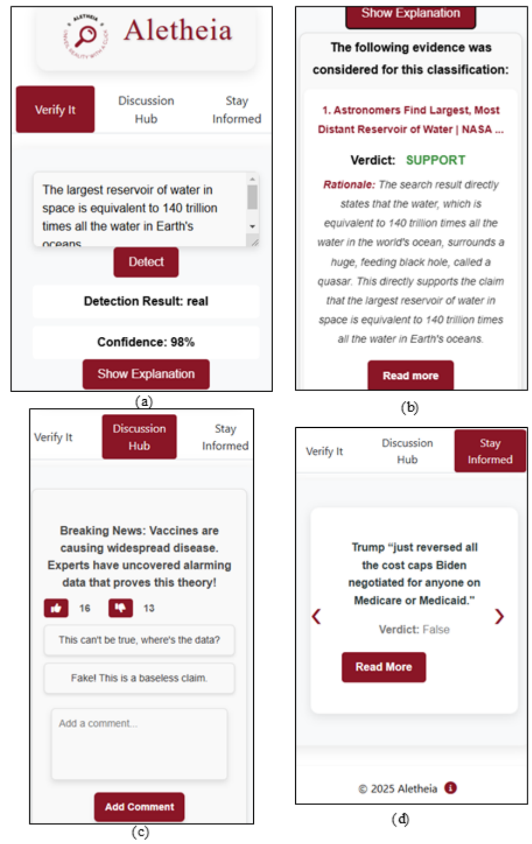

Figure 2: Demonstration Snapshots.

3.2 Backend: Server

The backend server hosts a RESTful API that connects with the browser extension. When a user submits a claim, the extension sends the relevant data to the server, which then orchestrates communication with various backend components via dedicated API endpoints. This setup ensures comprehensive responses from each module:

Fact-check Fetcher

The Fact-check Fetcher automates the retrieval of fact-checking data from the Google Fact Check API.[2] It sends requests to the API, obtains raw data, and cleans it to ensure usability. Key information such as claim text, verdict, source URL, and review publication date is extracted from each fact-checked claim. Only claims reviewed in the last 30 days are kept, ensuring up-to-date fact checks.

[2] https://toolbox.google.com/factcheck/apis

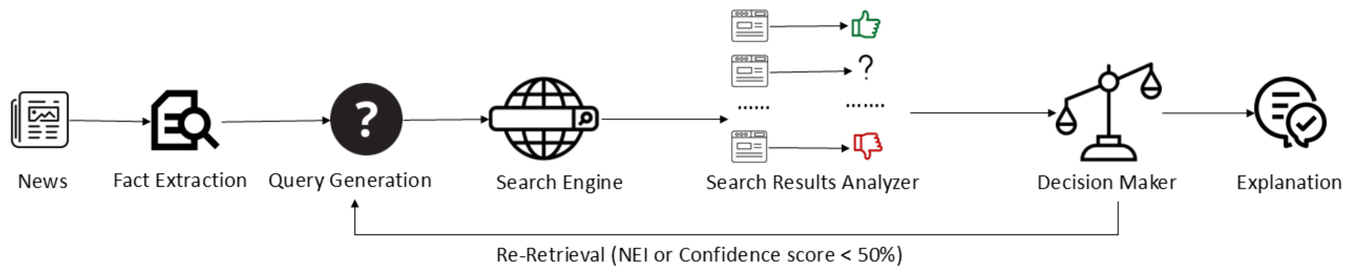

Figure 3: Overview of the FakeCheckRAG.

Community Database

The Community Database, implemented using PostgreSQL, manages posts, comments, and voting through a structured data schema. Designed for scalability and robustness, this database can handle large datasets efficiently while providing optimized storage solutions. Object-relational mapping is utilized to streamline communication between the server and the database, enhancing data management processes.

FakeCheckRAG

FakeCheckRAG, illustrated in Figure 3, begins by extracting the primary fact from a news article and formulating a search query. Using the Google Search API,[3] the system retrieves approximately ten relevant web links as evidence. To ensure source credibility, a filtering mechanism excludes any results from a predefined list of 1,044 known fake news websites [Papadogiannakis *et al.*, 2023]. Each search result is then analyzed by an LLM to determine whether it supports, contradicts, or is unrelated to the extracted fact. The aggregated evidence is used to classify the news claim as *Real*, *Fake*, or *NEI* (Not Enough Information). To ensure reliability, each classification is assigned a confidence score (0–100%), mitigating inconsistencies [Xiong *et al.*, 2023] and hallucinations [Ye and Durrett, 2022]. If the collected evidence is deemed sufficient, the system provides a final prediction along with an explanatory text based on the evidence's adequacy. If a research condition is met, the system initiates an updated search to gather additional information. This re-retrieval process ensures continuous evidence accumulation, merging initial data into an established evidence pool and generating new queries to enhance the accuracy and reliability of the truthfulness assessment.

4 Model Performance

To evaluate FakeCheckRAG's performance, we conduct experiments using the PolitiFact dataset [Shu *et al.*, 2020]. We then compare it against eleven baselines: (1) **Classical evidence-based methods**: DeClarE [Popat *et al.*, 2018], HAN [Ma *et al.*, 2019], EHIAN [Wu *et al.*, 2021], MAC [Vo and Lee, 2021], GET [Xu *et al.*, 2022], MUSER [Liao *et al.*, 2023], and ReRead [Hu *et al.*, 2023]. (2) **LLM-based methods**: GPT-3.5-turbo [OpenAI, 2022], Vicuna-7B [Chiang *et al.*, 2023], WEBGLM [Liu *et al.*, 2023], ProgramFC [Pan *et al.*, 2023], and STEEL [Li *et al.*, 2024].

[3] https://developers.google.com/custom-search/v1/overview

Method	Real			Fake		
	F1	P	R	F1	P	R
DeClarE	0.65	0.68	0.67	0.65	0.61	0.66
HAN	0.67	0.67	0.68	0.64	0.65	0.63
EHIAN	0.67	0.68	0.65	0.65	0.62	0.62
MAC	0.70	0.69	0.70	0.65	0.65	0.64
GET	0.72	0.71	0.77	0.66	0.72	0.66
MUSER	0.75	0.73	0.78	0.70	0.72	0.68
ReRead	0.71	0.71	0.75	0.68	0.71	0.69
GPT-3.5-turbo	0.57	0.55	0.56	0.55	0.56	0.57
Vicuna-7B	0.52	0.53	0.52	0.51	0.52	0.51
WEBGLM	0.60	0.61	0.63	0.61	0.66	0.62
ProgramFC	0.73	0.72	0.74	0.63	0.62	0.64
STEEL	0.78	0.74	0.78	0.72	0.74	0.72
FakeCheckRAG$_{3.5}$	0.82	0.71	**0.99**	0.74	**0.98**	0.59
FakeCheckRAG$_4$	**0.85**	**0.83**	0.86	**0.83**	0.84	**0.83**

Table 1: Comparison of our model's performance against baselines.

FakeCheckRAG is evaluated using two backbone models: GPT-4 (FakeCheckRAG$_4$) and GPT-3.5-turbo (FakeCheckRAG$_{3.5}$). As shown in Table 1, FakeCheckRAG outperforms both classical evidence-based methods, like ReRead and MUSER, and modern LLM-based approaches, such as STEEL and WEBGLM. FakeCheckRAG$_{3.5}$ achieves an F1 score of 0.82 for real content and 0.74 for fake content. FakeCheckRAG$_4$ further improves performance, achieving F1 scores of 0.85 for real content and 0.83 for fake content. These results highlight FakeCheckRAG, particularly with GPT-4, as a new benchmark for accurate and reliable fake news detection.

5 Conclusion

In this paper, we presented *Aletheia*, a novel browser extension that leverages RAG and LLMs for accurate fake news detection with evidence-based explanations. Its Discussion Hub fosters engagement, while the Stay Informed feature delivers real-time fact-checks. Experimental results show *Aletheia* outperforms different baselines. Despite its advantages, *Aletheia* faces challenges, including dependency on external APIs. Future work will address these limitations by reducing reliance on third-party APIs and expanding support for multiple languages.

References

[Amri *et al.*, 2021] Sabrine Amri, Dorsaf Sallami, and Esma Aïmeur. Exmulf: An explainable multimodal content-based fake news detection system. In *International Symposium on Foundations and Practice of Security*, pages 177–187. Springer, 2021.

[Balakrishnan *et al.*, 2022] Vimala Balakrishnan, Ng Wei Zhen, Soo Mun Chong, Gan Joo Han, and Tan Jiat Lee. Infodemic and fake news–a comprehensive overview of its global magnitude during the covid-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction*, page 103144, 2022.

[Botnevik *et al.*, 2020] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. Brenda: Browser extension for fake news detection. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 2117–2120, 2020.

[Check, 2024] Media Bias Fact Check. https://shorturl.at/oGBd9, 2024. Accessed: February 8, 2025.

[Checker, 2024] The Fact Checker. https://shorturl.at/rstpg, 2024. Accessed: February 8, 2025.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[Detector, 2024] B.S. Detector. https://shorturl.at/Kyvpk, 2024. Accessed: February 8, 2025.

[Epstein *et al.*, 2022] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. Do explanations increase the effectiveness of ai-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 183–193, 2022.

[Hartwig and Reuter, 2019] Katrin Hartwig and Christian Reuter. Trustytweet: an indicator-based browser-plugin to assist users in dealing with fake news on twitter. 2019.

[Hu *et al.*, 2023] Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip Yu. Read it twice: Towards faithfully interpretable fact verification by revisiting evidence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2319–2323, 2023.

[Kuntur *et al.*, 2024] Soveatin Kuntur, Anna Wróblewska, Marcin Paprzycki, and Maria Ganzha. Under the influence: A survey of large language models in fake news detection. *IEEE Transactions on Artificial Intelligence*, 2024.

[Li *et al.*, 2024] Guanghua Li, Wensheng Lu, Wei Zhang, Defu Lian, Kezhong Lu, Rui Mao, Kai Shu, and Hao Liao. Re-search for the truth: Multi-round retrieval-augmented large language models are strong fake news detectors. *arXiv e-prints*, pages arXiv–2403, 2024.

[Liao *et al.*, 2023] Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4461–4472, 2023.

[Liu *et al.*, 2023] Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4549–4560, 2023.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[Ma *et al.*, 2019] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics, 2019.

[Moalla *et al.*, 2025] Hounaida Moalla, Hana Abid, Dorsaf Sallami, Esma Aïmeur, and Bassem Ben Hamed. Exploring the power of dual deep learning for fake news detection. *Informatica*, 48(4), 2025.

[OpenAI, 2022] OpenAI. https://platform.openai.com/docs/models/gpt-3-5, 2022. Accessed: February 8, 2025.

[Pan *et al.*, 2023] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*, 2023.

[Papadogiannakis *et al.*, 2023] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. Who funds misinformation? a systematic analysis of the ad-related profit routines of fake news sites. In *Proceedings of the ACM Web Conference 2023*, pages 2765–2776, 2023.

[Paschalides *et al.*, 2021] Demetris Paschalides, Chrysovalantis Christodoulou, Kalia Orphanou, Rafael Andreou, Alexandros Kornilakis, George Pallis, Marios D Dikaiakos, and Evangelos Markatos. Check-it: A plugin for detecting fake news on the web. *Online Social Networks and Media*, 25:100156, 2021.

[Popat *et al.*, 2018] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*, 2018.

[Sallami and Aïmeur, 2025] Dorsaf Sallami and Esma Aïmeur. Exploring beyond detection: a review on fake news prevention and mitigation techniques. *Journal of Computational Social Science*, 8(1):1–38, 2025.

[Sallami *et al.*, 2023] Dorsaf Sallami, Rim Ben Salem, and Esma Aïmeur. Trust-based recommender system for fake news mitigation. In *Adjunct Proceedings of the 31st ACM*

Conference on User Modeling, Adaptation and Personalization, pages 104–109, 2023.

[Shu *et al.*, 2020] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

[Velasco *et al.*, 2023] Abigail T Velasco, Allen Roi C Cortez, John Meynard B Camay, Ian Michael C Giba, and Marlon A Diloy. Factit: A fact-checking browser extension. In *2023 IEEE 12th International Conference on Educational and Information Technology (ICEIT)*, pages 342–347. IEEE, 2023.

[Vo and Lee, 2021] Nguyen Vo and Kyumin Lee. Hierarchical multi-head attentive network for evidence-aware fake news detection. *arXiv preprint arXiv:2102.02680*, 2021.

[von der Weth *et al.*, 2020] Christian von der Weth, Jithin Vachery, and Mohan Kankanhalli. Nudging users to slow down the spread of fake news in social media. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2020.

[Warman and Kabir, 2023] Dylan Warman and Muhammad Ashad Kabir. Covidfakeexplainer: An explainable machine learning based web application for detecting covid-19 fake news. In *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 01–06. IEEE, 2023.

[Woleński, 2004] Jan Woleński. Aletheia in greek thought until aristotle. *Annals of Pure and Applied Logic*, 127(1-3):339–360, 2004.

[Wu *et al.*, 2021] Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1388–1394, 2021.

[Xiong *et al.*, 2023] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

[Xu *et al.*, 2022] Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022*, pages 2501–2510, 2022.

[Ye and Durrett, 2022] Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392, 2022.