Combating Biomedical Misinformation through Multi-modal Claim Detection and Evidence-based Verification

Mariano Barone
mariano.barone@studenti.unina.it
University of Naples Federico II
Naples, Italy

Antonio Romano
antonio.romano5@unina.it
University of Naples Federico II
Naples, Italy

Giuseppe Riccio
giuseppe.riccio3@unina.it
University of Naples Federico II
Naples, Italy

Marco Postiglione
marco.postiglione@northwestern.edu
Northwestern University
Evanston, IL, USA

Vincenzo Moscato
vincenzo.moscato@unina.it
University of Naples Federico II
Naples, Italy

Abstract

Biomedical misinformation — ranging from misleading social media posts to fake news articles and deepfake videos — is increasingly pervasive across digital platforms, posing significant risks to public health and clinical decision-making. We developed CER, a comprehensive fact-checking system designed specifically for biomedical content. CER integrates specialized components for claim detection, scientific evidence retrieval, and veracity assessment, leveraging both transformer models and large language models to process textual and multimedia content. Unlike existing fact-checking systems that focus solely on structured text, CER can analyze claims from diverse sources including videos and web content through automatic transcription and text extraction. The system interfaces with PubMed for evidence retrieval, employing both sparse and dense retrieval methods to gather relevant scientific literature. Our evaluations on standard benchmarks including HealthFC, BioASQ-7, and SciFact demonstrate that CER achieves state-of-the-art performance, with F1-score improvements compared to existing approaches. To ensure reproducibility and transparency, we release the GitHub repository with the source code, within which you can reach an interactive demonstration of the system published on HuggingFace and a video demonstration of the system https://github.com/PRAISELab-PicusLab/CER-Fact-Checking.

CCS Concepts

• **Computing methodologies** → **Natural language generation**;
• **Applied computing** → **Health care information systems**.

Keywords

Fact-Checking, Healthcare, Generative AI, Large Language Models

ACM Reference Format:
Mariano Barone, Antonio Romano, Giuseppe Riccio, Marco Postiglione, and Vincenzo Moscato. 2025. Combating Biomedical Misinformation through Multi-modal Claim Detection and Evidence-based Verification. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730155

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730155

1 Introduction

Biomedical misinformation threatens public health by influencing clinical decisions and eroding trust in medical institutions. Its spread can result in delayed treatments, higher mortality rates, and reliance on unverified practices [2, 4]. Addressing this challenge requires advanced tools for claim verification. Recent AI-driven fact-checking developments have led to tools like factiverse.ai [12], designed to assess online information credibility. Other fact-checking systems, such as Sourcer AI[1] use machine learning models to evaluate the reliability of news articles and social media content. Another fact-checker application is FactCheck Editor [11] that extends automated fact-checking capabilities to text editing workflows, supporting multilingual verification across over 90 languages. However, in biomedical and healthcare fields, fact-checking still relies heavily on manual verification, a time-consuming and complex process due to specialized terminology and the need for empirical evidence interpretation [22]. Moreover, existing automated solutions primarily focus on structured texts such as documents or datasets that follow a predefined format with clear organization, overlooking more diverse sources such as online publications and multimedia content, thereby limiting their effectiveness in countering misinformation across digital platforms. This issue is further exacerbated by the rise of deepfake technologies [18], which enable the creation of fabricated yet highly realistic medical content [10], including altered videos of healthcare professionals, synthetic patient testimonies, and manipulated research presentations [8] . Such deceptive content can spread false medical advice or distort scientific consensus, increasing public confusion and skepticism toward legitimate healthcare recommendations [3]. To overcome these limitations, this work introduces Combining Evidence and Reasoning (CER) a multi-stage LLM-based system integrating claim detection, evidence retrieval, and veracity assessment. Unlike conventional approaches, CER processes both text and video content by extracting audio, generating transcriptions, and identifying claims for verification. This capability extends the reach of automated fact-checking to non-textual sources, enhancing its applicability in real-world

[1] https://sourcerai.co/

Figure 1: Examples showing system functionalities across three input types: individual claims (left), web pages (center), and video content (right). For claims derived from web pages and videos, users can expand the interface—as illustrated in the *Individual Claims* panel—to examine claim justifications and scientific evidence that either supports or refutes each claim.

misinformation scenarios. The system interfaces with PubMed for evidence retrieval, using dense retrieval to collect relevant scientific literature. CER is designed for researchers, fact-checkers, healthcare professionals, and policymakers, providing an effective tool for verifying biomedical claims with transparent, evidence-based explanations. It significantly reduces the efforts required to assess the credibility of claims while improving accuracy through advanced retrieval techniques and LLM-based reasoning [21]. The system's interface presents retrieved evidence alongside claim assessments, offering a clear and interpretable verification process. Experimental evaluations demonstrate that CER achieves state-of-the-art performance on benchmarks such as HealthFC [15], BioASQ-7 [7], and SciFact [16], with F1-score improvements of up to +7.69%. Furthermore, we demonstrate CER's effectiveness as a complementary safeguard in deepfake detection pipelines, addressing a critical gap where traditional detectors struggle to generalize to emerging synthetic media formats amid rapidly evolving generation technologies [19].

2 CER: Overview

In this section, we present our proposed methodology to classify biomedical *claims*[2]. Our architecture, as illustrated in Figure 2, is divided into four main components: (1) Claim Detection, (2) Scientific Evidence Retrieval, (3) LLM Reasoning and (4) Veracity Prediction. The remainder of this section provides an in-depth description of each component.

[2]In this work, a *claim* refers to any statement or question whose truthfulness or factual accuracy can be evaluated based on external evidence. This broad definition encompasses declarative sentences as well as interrogative forms.

⌨️ *Inputs.* CER does not only handle plain text data, but also *web pages* and *videos*. When dealing with web pages, it extracts their HTML and identifies potential claims with an LLM-based method; the claim verification pipeline is then applied to all the claims. Similarly, audios are automatically extracted and transcribed when operating with videos, allowing the system to detect and validate claims from spoken content.

🔍 *Claim Detection.* The claim detection phase leverages a LLM to identify statements worthy of verification within biomedical texts extracted from online sources, like articles and videos. The process begins with a text preprocessing step, including segmentation and tokenization, to prepare the input for LLM analysis. Claims are identified by classifying each sentence or paragraph based on its informational relevance, pertinence, and potential impact on clinical or public health decisions, using zero-shot or few-shot prompting techniques. When the input is a *video*, the system first extracts the audio, then transcribes the content, and uses the LLM to spot verifiable claims. Transcription is performed using *Whisper small-v3*, a robust multilingual speech recognition model capable of handling medical terminology and varied audio conditions. For *web pages* (URLs), it analyzes the HTML code, extracting and processing the text to pinpoint pertinent claims. By incorporating LLMs in the claim detection step, CER guarantees scalability and flexibility in handling biomedical content from a variety of sources, such as textual documents, web pages, and multimedia. We use *meta/llama-3.1-405b-instruct* for claim detection tasks, given its strong performance in instruction-following scenarios.

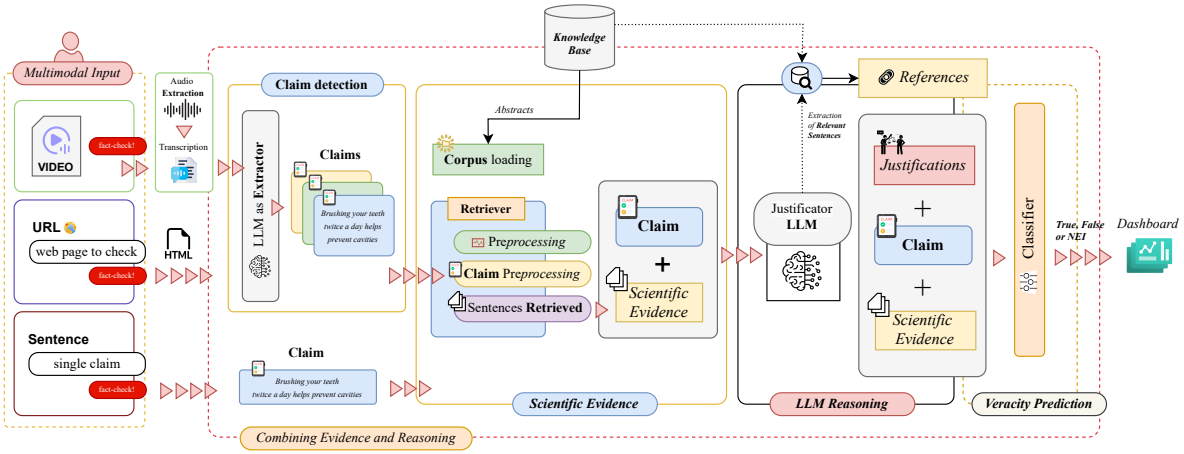

Figure 2: CER workflow. The system processes multimodal inputs, including text, URLs, and video content, by extracting and transcribing audio when necessary. Claims are detected using an LLM-based extractor and preprocessed before being matched against a scientific corpus indexed with a retrieval system. Relevant evidence passages are retrieved and combined with the original claim as input to an LLM, which generates detailed justifications. The system also incorporates a knowledge base for additional reference extraction. Finally, a binary classifier assesses the claim's veracity based on the claim text, retrieved scientific evidence, and generated justifications, producing a final classification (true/false/nei). The results are visualized through a dashboard for interpretability and decision support

🔖 *Scientific Evidence Retrieval.* The process begins with a retrieval module interfacing with PubMed, a comprehensive repository of peer-reviewed biomedical literature. Focusing on abstracts, which offer dense and accessible research summaries, evidence is extracted using Dense Retrieval (multi-qa-MiniLM-L6-cos-v1) with a FAISS index. Dense Retrieval employs SBERT embeddings to capture semantic nuances. Future extensions may incorporate BioBERT-based retrieval for better domain coverage. Preprocessing includes normalization, tokenization, and stopword removal. Claims are processed as queries, and the top 20 results are ranked for relevance. For each claim, up to three evidence pieces are selected and formatted as a concatenated structure with a separator token ([SEP]), balancing information coverage and computational efficiency. These structured claim-evidence pairs are then prepared for reasoning.

✨ *LLM Reasoning.* The reasoning phase employs large language models (LLMs) as assistants to assess claim veracity and generate justifications, mitigating the risk of hallucinations. We use *meta-llama/Meta-Llama-3.1-405B-Instruct*, an instruction-tuned language model by Meta, chosen for its strong reasoning capabilities and coherent, context-aware justifications. Its robustness across biomedical domains makes it ideal for tasks requiring factual accuracy and explanatory output. Given a structured prompt with claim and evidence, the model returns a binary veracity assessment and a detailed justification. The outputs are passed to the subsequent veracity prediction stage, ensuring the reasoning aligns with retrieved evidence. This design ensures transparency and accuracy in evaluating biomedical claims. Notably, while the LLM provides a preliminary binary veracity judgment, this output is not directly used in the final classification. Instead, the justification is combined with the original claim and fed into the classification model (e.g., DeBERTa or BERT) to determine the final label.

🧠 *Veracity Prediction.* The final stage integrates the outputs of the LLM reasoning module into a classification framework to assign one of three labels: "true," "false," or "nei (not enough information)." Two approaches are explored: zero-shot classification and fine-tuning. The zero-shot approach leverages pre-trained models to classify claims without task-specific training, providing rapid deployment but limited domain precision. Fine-tuning adapts the model using labeled domain-specific data, optimizing accuracy and relevance at the cost of additional computational resources. This stage ensures robust and reliable fact-checking tailored to biomedical applications.

3 Experiments

This section presents an empirical evaluation of the CER architecture across multiple datasets and configurations. Our evaluation focuses on comparing our proposed approach with both fact-checking and deepfake detection baselines.

3.1 Experimental Setup

3.1.1 Datasets. We evaluate CER on three established fact-checking benchmark datasets and a novel video dataset. For text-based evaluation, we use HealthFC [15], which contains 750 health-related claims verified by medical professionals and labeled as true, false, or NEI (Not Enough Information); BioASQ-7b [7], comprising 745 biomedical claims labeled as true or false; and SciFact [16], consisting of 1.4k scientific claims labeled as "support", "confute", or "NEI", with accompanying expert-written rationales for validation. To assess CER's effectiveness on deepfake content, we created a balanced dataset of 40 healthcare videos: 20 deepfake videos collected

Category	Fact-Checker	HealthFC			BioASQ			SciFact		
		Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
Baselines	All True (baseline)	8.97	33.33	14.14	82.39	1.0	90.34	13.71	33.33	19.43
	All False (baseline)	5.55	33.33	9.52	17.60	1.0	29.94	7.16	33.33	11.78
	All NEI (baseline)	18.79	33.33	24.04	–	–	–	12.47	33.33	18.15
Online platforms	Factiverse.AI [13]	53.47	30.19	38.59	91.31	84.01	87.51	50.25	46.80	48.46
	Perplexity.AI[8]	56.40	47.20	51.39	83.20	80.13	81.64	52.75	50.00	51.34
LLM Predictions	Mixtral LLM prediction[10]	40.10	41.24	40.61	72.43	70.69	71.54	45.89	40.60	43.02
	GPT4o mini	40.52	39.88	39.06	53.85	52.71	52.79	39.01	37.64	38.31
Literature baselines	Vladika and Matthes [14]	52.60	44.50	40.60	60.90	64.40	61.70	46.00	42.30	44.07
	Bekoulis et al. [1]	41.72	49.36	45.21	47.42	52.36	49.76	54.12	51.20	52.62
	Zaheer et al. [20]	26.67	35.95	30.62	69.62	33.57	45.29	38.70	35.20	36.87
	Lan et al. [5]	21.40	39.36	27.72	74.60	50.37	60.13	34.45	31.85	33.10
	Vladika et al. [15]	68.24	66.84	67.53	-	-	-	-	-	-
CER (*zero-shot*)	CER (DeBERTA v3-large)	58.26	51.48	54.61	93.10	90.48	91.77	55.72	51.90	53.74
CER (*fine-tuned*)	CER (DeBERTA v3-large)	64.55	70.14	67.22	**93.70**	**96.75**	**95.20**	**61.44**	**60.72**	**61.14**
	CER (PubMedBERT)	67.55	**72.43**	**69.90**	89.76	90.02	89.88	57.70	56.13	56.91
	CER (BERT)	**69.92**	69.33	69.62	81.28	83.89	82.56	58.10	55.89	56.97

Table 1: Comparison with state-of-the-art baselines. This table presents the macro precision, recall, and F1 scores for different baseline models evaluated on the experimented datasets. Best results are highlighted in bold. '–' indicates instances where a technique could not be applied.

Method	*Real* class			*Fake* class		
	P	R	F1	P	R	F1
Face X-Ray	1.00	1.00	1.00	1.00	0.90	0.95
ResNet-34	1.00	0.89	0.94	1.00	0.80	0.89
F3Net	1.00	0.89	0.94	1.00	0.45	0.62
CER	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Comparison with deepfake detection baselines. This table presents the precision (P), recall (R), and F1 scores for different baseline models and the two classes (i.e. real, fake).

from public sources and 20 authentic videos containing verified healthcare information[3].

3.1.2 Baselines. We compared CER with open- and closed-source fact-checkers. Factiverse.AI and Perplexity.AI are closed-source platforms that classify evidence as "Supporting", "Mixed", or "Disputing" using sources like Google, Bing, Wikipedia, and FactSearch. Vladika and Matthes [14] propose a voting-based architecture similar to ours, while Bekoulis et al. [1] focus on Wikipedia document and sentence retrieval for claim validation. Other methods include BigBird [20], ALBERT [5], and Vladika et al. [15]'s pipeline for medical claim fact-checking. Simple baselines ("All True", "All False", "All NEI") were also included. Additionally, we compared our proposed method with several deepfake detection baselines: Face X-Ray [6], ResNet-34 [17], and F3Net [9].

3.2 Results

We evaluated CER against state-of-the-art fact-checking baselines using both general-domain models (BERT, DeBERTa-v3) and a domain-specific biomedical model (PubMedBERT) in zero-shot and fine-tuned configurations. As shown in Table 1, fine-tuned CER consistently outperforms all baselines across datasets, achieving F1-scores of 69.90% on HealthFC, 95.20% on BioASQ, and 61.14% on SciFact. The model maintains competitive performance even in zero-shot settings, reaching 54.61% on HealthFC, 91.77% on BioASQ, and 53.74% on SciFact. While PubMedBERT leverages domain-specific biomedical pretraining, it achieves lower performance on both SciFact (56.91%) and BioASQ (89.88%). This performance gap highlights the advantages of combining general-domain knowledge with task-specific fine-tuning. Similarly, although DeBERTa-v3 demonstrates strong results on BioASQ (95.20%) and SciFact (61.14%), it falls short of CER's performance on HealthFC and BioASQ. Table 2 presents a comparison between CER and established deepfake detection baselines. Our classification strategy designated a video as "fake" if our system identified at least one false claim within its content. Interestingly, this approach achieved perfect accuracy across the entire benchmark dataset, surpassing the performance of all evaluated deepfake detectors. While these results are preliminary, they highlight CER's potential as a complementary component in deepfake detection pipelines. This is especially significant given that traditional deepfake detectors often struggle to generalize to novel forms of synthetic media, due to the rapid evolution of deepfake generation technologies [19].

[3]Given the limited size of our benchmark dataset, the deepfake detection results should be considered preliminary.

4 Conclusion

We introduced CER (Combining Evidence and Reasoning), a biomedical fact-checking system with a user-friendly demo. It verifies medical claims, web pages, and videos by integrating scientific evidence retrieval with LLM-based reasoning. This structured approach balances factual accuracy and interpretability, offering explainable responses with accurate sources.

Acknowledgements

This work was conducted with the financial support of (1) the PNRR MUR project PE0000013-FAIR and (2) the Italian ministry of economic development, via the ICARUS (Intelligent Contract Automation for Rethinking User Services) project (CUP: B69J23000270005).

References

[1] Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A Review on Fact Extraction and Verification. *ACM Comput. Surv.* 55, 1, Article 12 (Nov. 2021), 35 pages. doi:10.1145/3485127

[2] Michael A Gisondi, Rachel Barber, Jemery Samuel Faust, Ali Raja, Matthew C Strehlow, Lauren M Westafer, and Michael Gottlieb. 2022. A deadly infodemic: social media and the power of COVID-19 misinformation. e35552 pages.

[3] Fred Matanel Grabovski, Lior Yasur, Guy Amit, Yuval Elovici, and Yisroel Mirsky. 2024. Back-in-Time Diffusion: Unsupervised Detection of Medical Deepfakes. *CoRR* abs/2407.15169 (2024). doi:10.48550/ARXIV.2407.15169 arXiv:2407.15169

[4] Skyler B Johnson, Matthew Parsons, Tanya Dorff, Meena S Moran, John H Ward, Stacey A Cohen, Wallace Akerley, Jessica Bauman, Joleen Hubbard, Daniel E Spratt, et al. 2022. Cancer misinformation and harmful information on Facebook and other social media: a brief report. *JNCI: Journal of the National Cancer Institute* 114, 7 (2022), 1036–1039.

[5] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *CoRR* abs/1909.11942 (2019), –. arXiv:1909.11942 http://arxiv.org/abs/1909.11942

[6] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5001–5010.

[7] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2020. Results of the seventh edition of the BioASQ Challenge. *CoRR* abs/2006.09174 (2020), –. arXiv:2006.09174 https://arxiv.org/abs/2006.09174

[8] Shankargouda Patil. 2024. Artificial Intelligence Face Swapping: Promise and Peril in Health Care. *Mayo Clinic Proceedings: Digital Health* 2 (03 2024), 159. doi:10.1016/j.mcpdig.2024.01.009

[9] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*. Springer, 86–103.

[10] Jamshir Qureshi and Samina Khan. 2024. Artificial Intelligence (AI) Deepfakes in Healthcare Systems: A Double-Edged Sword? Balancing Opportunities and Navigating Risks. *Preprints* (February 2024). doi:10.20944/preprints202402.0176.v1

[11] Vinay Setty. 2024. FactCheck Editor: Multilingual Text Editor with End-to-End fact-checking. arXiv:2404.19482 [cs.CL] https://arxiv.org/abs/2404.19482

[12] Vinay Setty. 2024. Surprising Efficacy of Fine-Tuned Transformers for Fact-Checking over Larger Language Models. arXiv:2402.12147 [cs.CL] https://arxiv.org/abs/2402.12147

[13] Vinay Setty. 2024. Surprising Efficacy of Fine-Tuned Transformers for Fact-Checking over Larger Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, Washington DC, USA, 2842–2846. doi:10.1145/3626772.3661361

[14] Juraj Vladika and Florian Matthes. 2024. Improving Health Question Answering with Reliable and Time-Aware Evidence Retrieval. *CoRR* abs/2404.08359 (2024), –. doi:10.48550/ARXIV.2404.08359 arXiv:2404.08359

[15] Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italy, 8095–8107. https://aclanthology.org/2024.lrec-main.709

[16] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7534–7550. doi:10.18653/V1/2020.EMNLP-MAIN.609

[17] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8695–8704.

[18] Mika Westerlund. 2019. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review* 9 (11/2019 2019), 40–53. doi:10.22215/timreview/1282

[19] Kelu Yao, Jin Wang, Boyu Diao, and Chao Li. 2023. Towards Understanding the Generalization of Deepfake Detectors from a Game-Theoretical View. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2031–2041. doi:10.1109/ICCV51070.2023.00194

[20] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). NeurIPS, Online, –. https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html

[21] Xuan Zhang and Wei Gao. 2024. Reinforcement Retrieval Leveraging Fine-grained Feedback for Fact Checking News Claims with Black-Box LLM. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italy, 13861–13873. https://aclanthology.org/2024.lrec-main.1209

[22] Bu Zhong. 2023. Going beyond fact-checking to fight health misinformation: A multi-level analysis of the Twitter response to health news stories. *Int. J. Inf. Manag.* 70 (2023), 102626. doi:10.1016/J.IJINFOMGT.2023.102626