Information Integrity in the Age of Foundation Models:

A Field Synthesis of the ACM Web Conference 2026 (WWW '26)

Alexander Loth
Frankfurt University of Applied Sciences, Frankfurt am Main, Germany
alexander.loth@stud.fra-uas.de

November 28, 2026 | Conference Synthesis Report No. AL-CSR-2026-01

Abstract

The ACM Web Conference 2026 (WWW '26, rescheduled from April to 29 June–3 July 2026 in Dubai) marked a structural inflection point in research on web information integrity. Rather than treating generative artificial intelligence (GenAI) solely as a generator of synthetic fake news, contributions across the Main Track and Companion workshops converged on an ecosystem-wide view spanning search and platform pollution, multimodal and adversarial detection, retrieval-augmented and multi-agent fact-checking, human verification behavior, cryptographic provenance, and computational reproducibility. This field report synthesizes 33 peer-reviewed WWW '26 contributions (29 external papers and workshop overviews from the Main Track and Companion Proceedings alongside four workshop contributions co-authored by the author across AiOfAi, TrustFM, R2CASS, and the canceled FIIR workshop). Organizing the literature along five themes, the synthesis highlights three cross-cutting shifts: (1) from isolated artifact classification toward ecosystem-level failure modes such as retrieval collapse and adversarial evasion; (2) from opaque single-model detectors toward evidence-grounded, multi-agent, and uncertainty-aware verification pipelines; and (3) from purely automated detection toward provenance infrastructure, human verification habits, and reproducible evaluation protocols. The report closes with a compact synthesis matrix and four open research priorities for the 2027 conference cycle.

Keywords: ACM Web Conference 2026, WWW '26, generative AI, disinformation, misinformation detection, fact-checking, provenance, human perception, reproducibility.

1 Introduction and Scope

The 2026 edition of the ACM Web Conference (WWW '26) took place from 29 June to 3 July 2026 in Dubai, United Arab Emirates, following its rescheduling from April 2026. Across both the Main Proceedings and the Companion Proceedings workshops, a substantial cluster of papers examined how large language models (LLMs), vision-language models (VLMs), and generative diffusion systems reshape the integrity of web information ecosystems. Compared with earlier conference cycles focused predominantly on binary text classification of synthetic news, the WWW '26 corpus reveals a marked broadening of scope: researchers now study systemic feedback loops in search and Q&A platforms, multimodal and cross-domain manipulation, human cognitive verification strategies, cryptographic provenance architectures, and the reproducibility of computational social science itself.

This report synthesizes 33 works from WWW '26 (29 external peer-reviewed papers and workshop overviews plus four workshop papers co-authored by the author) to map the state of the field as presented at the conference. Because the author's own work spanned four distinct

Companion workshops—the 4th Workshop on Adverse Impacts and Collateral Effects of AI Technologies (AiOfAi) [2], the Workshop on Trustworthy Foundation Models for Web Intelligence (TrustFM) [6], the 2nd Workshop on Social Science Meets Web Data: Reproducible and Reusable Computational Approaches (R2CASS) [17], and the Future of Information Integrity Research Workshop (FIIR, whose accepted papers circulated as preprints after the workshop was canceled due to regional conflict)—the synthesis deliberately connects Main Track advances with the thematic discussions emerging from these specialized communities.

Methodological Note. Every cited entry in this report was verified against official Crossref proceedings records (`10.1145/3774904.*` for Main Track and `10.1145/3774905.*` for Companion Proceedings) and arXiv (`2602.03423`), and claims about external papers are strictly restricted to what is documented in their official titles and abstracts (retrieved via Crossref, OpenAlex, and arXiv on 28 November 2026). Full-text empirical figures are reported only where verified directly in the author's own camera-ready papers.

2 Theme 1: Ecosystem Pollution, Retrieval Collapse, and Collateral Effects

A central shift visible at WWW '26 is the recognition that generative AI alters web infrastructure even in the absence of targeted political disinformation campaigns. When synthetic content is produced at near-zero marginal cost, it enters public search indices, community Q&A platforms, and training corpora, degrading the reliability of downstream retrieval and human knowledge sharing.

Yu et al. [32] formalize this structural risk as *Retrieval Collapse*, a two-stage ecosystem failure in which AI-generated content first dominates web search results and erodes source diversity, and subsequently feeds lower-quality synthetic evidence back into search engines and Retrieval-Augmented Generation (RAG) pipelines. Complementing this retrieval perspective, Shin and Son [21] examine how the surge of LLM-generated answers affects human knowledge exchange on a major online Q&A platform (Naver Knowledge iN), studying how synthetic responses interact with community-authored contributions. On the adversarial side of search infrastructure, Chen et al. [4] present a systematic study of the resilience of LLM-enhanced search engines against black-hat search engine optimization (SEO) manipulation, demonstrating that combining web-scale retrieval with AI summarization creates new attack surfaces rather than automatically insulating users from manipulated ranking signals.

Within the AiOfAi workshop [2], which explicitly framed the dual-use paradox of AI technologies that simultaneously threaten and support online truth and security, Loth et al. [14] characterize this broader transition as *industrialized deception*: the shift from manual fabrication toward automated, ecosystem-level erosion of trust. Rather than treating countermeasures solely at the model level, Loth et al. [14] argue for moving from descriptive surveys to reproducible experimental pipelines—connecting controlled stimulus generation (RogueGPT) with human perception evaluation (JudgeGPT)—to study how detection, inoculation, and dual-use risks coevolve.

3 Theme 2: Multimodal, Cross-Domain, and Adversarial Detection

Despite widespread skepticism about whether purely automated detectors can keep pace with frontier generators, detection remained the largest technical cluster at WWW '26. However, the emphasis shifted decisively away from static, in-domain text classifiers toward *multimodal*, *unseen-domain*, *adversarially robust*, and *proactive* approaches.

Multimodal and Video Misinformation. Several Main Track and workshop contributions argued that real-world deception rarely resides in text alone. Cui et al. [8] propose a tri-axis formulation separating image veracity, text veracity, and cross-modal consistency into an eight-way taxonomy and benchmark to expose unseen combinations of multimodal manipulation. For short-form video, Zeng et al. [33] address diverse micro-video misinformation spanning multimodal manipulation, AI-generated content, cognitive bias, and out-of-context reuse, focusing on fine-grained attribution to support debunking. Similarly, Chen et al. [3] introduce TrueLens, a video fake news detection framework that consolidates external world knowledge with internal multimodal cues through dual-level evidence gathering. In the Companion Proceedings, Chen et al. [5] study out-of-context image-text mismatches when heavy external web evidence is unavailable, while Thapa et al. [24] evaluate interpretable vision-language models for concept-grounded detection of vaccine misinformation in visual formats such as memes and infographics.

Deepfake Generalization and Structural Hallucination. For synthetic audiovisual media, generalization across generators and semantic domains remained a primary concern. Kang et al. [11] propose HAUTE, combining physiological facial Action Unit dynamics with temporal-contextual embeddings to detect sophisticated deepfake videos beyond surface visual artifacts, while Liu et al. [12] disentangle semantic and forgery subspaces via orthogonal decomposition to improve cross-domain generalization in multimodal deepfake detection. Extending detection to structured web data, Noël and Seidou [18] introduce ReGA, a zero-overhead graph-alignment method for detecting structural hallucinations in knowledge-graph RAG without relying on costly generative judges.

Linguistic Fingerprints, Domain Shift, and Adversarial Evasion. On the textual side, WWW '26 papers explicitly confronted the brittleness of surface-level detectors. Wang et al. [27] analyze prompt-induced linguistic fingerprints by comparing paired LLM-generated real and fake news reconstructed under malicious prompts, while Ghasiya and Sasahara [9] examine empirically derived linguistic differences between real news and LLM-generated fake news to inform detection. Two further Main Track studies highlight critical robustness gaps: Tahmasebi et al. [23] investigate the vulnerability of LLM-based fake news detectors to adversarial sentiment attacks that exploit emotion cues, and Yang et al. [31] tackle cross-domain fake news detection on previously unseen domains through LLM-based domain-aware user modeling. Moving upstream from reactive classification to early warning, Comito et al. [7] operationalize polarization and confirmation bias into a semi-supervised pipeline for proactive misinformation topic forecasting on social platforms.

4 Theme 3: Evidence-Grounded and Multi-Agent Fact-Checking

A third prominent cluster at WWW '26 reframed verification from isolated classification into structured, multi-step evidence gathering and reasoning. Across both general social media and specialized scientific domains, researchers highlighted two recurring bottlenecks in automated fact-checking pipelines: noisy web retrieval and uncalibrated model reasoning.

To counter the "information-drowning" problem—where abundant truthful background context obscures sparse deceptive signals—Wang et al. [28] propose PAMAS, a self-adaptive multi-agent system that aggregates multiple perspectives for misinformation detection. Yang et al. [29] combine denoising evidence retrieval with LLM-based debate verification to mitigate noisy web snippets and single-model verification biases. In the multimodal setting, Yang et al. [30] introduce RAMA, a retrieval-augmented multi-agent framework that transforms ambiguous multimedia claims into targeted web search queries. Addressing specialized domains such as biomedicine and materials science, Vinod and Erk [26] propose an uncertainty-aware, web-conditioned scientific

fact-checking pipeline built around atomic predicate-argument decomposition and calibrated uncertainty gating.

Taken together, these four papers point toward a clear architectural convergence at WWW '26: verification systems are increasingly structured as explicit decomposition, retrieval-denoising, and uncertainty-calibrated multi-agent workflows rather than single-pass prompt classifiers.

5 Theme 4: Human Perception, Verification Behavior, and Synthetic Respondents

While Themes 2 and 3 advance automated verification, another set of WWW '26 papers examined how humans actually judge, verify, and interact with synthetic and misleading information—and whether LLMs can validly stand in for human respondents in web research.

At the TrustFM workshop [6], Loth et al. [13] analyze 918 human evaluations across five foundation models collected via JudgeGPT and RogueGPT, decoupling perceived authenticity (legitimate vs. fake) from source attribution (human vs. machine). Framing detection performance with Structural Causal Models, they report negligible correlation with political orientation ($r = -0.10$) compared with prior familiarity with fake news ($r = 0.35$), and identify a "fluency trap" in which highly fluent GPT-4 outputs bypass human source-monitoring mechanisms. This points toward pre-bunking interventions centered on cognitive source monitoring rather than demographic targeting.

Two external WWW '26 studies directly reinforce the behavioral dimension of human verification. In the Main Track, Artemenko et al. [1] conduct a large-scale online experiment comparing lateral reading (consulting multiple external sources in the browser) against vertical reading (spending more time on fewer sources) for health misinformation, showing that lateral navigation improves detection accuracy even under confirmation bias. In the Companion Proceedings, Pandey et al. [19] investigate how the effectiveness of AI chatbots in correcting misinformed beliefs varies across topic domains, cautioning against domain-agnostic conversational interventions.

Finally, a methodological warning emerged inside the Companion workshops regarding the use of LLMs as proxies for human subjects: Taday Morocho et al. [22] evaluate persona-conditioned open-weight LLMs against U.S. World Values Survey microdata to test whether multi-attribute persona prompting improves reliability or introduces systematic distortions. Their inquiry resonates directly with wider concerns at WWW '26 that synthetic respondents cannot replace empirical human perception studies without rigorous validity checks.

6 Theme 5: Provenance Infrastructure, Watermarking, and Reproducible Evaluation

The fifth theme across WWW '26 addresses what comes after acknowledging that neither human intuition nor reactive post-hoc detectors are sufficient on their own: building cryptographic provenance, watermarking, and reproducible evaluation standards as shared infrastructure.

On the media provenance and watermarking front, Huang et al. [10] propose AWMA-MoE, an attention-guided watermark adapter using a mixture-of-experts design for latent diffusion models aimed at balancing watermark robustness against image quality degradation. Moving from generation-time watermarking to edge verification, Vela-Koentarjo and Singh [25] propose a blockchain-assisted decentralized provenance architecture for tamper-evident registration of security camera video streams at the edge. Complementing these server- and edge-side approaches at the point of end-user consumption, Loth et al. [16]—accepted at the FIIR workshop prior to its regional-conflict cancellation and released as a preprint—present Origin Lens, a privacy-first mobile framework (Rust/Flutter) performing on-device cryptographic image provenance

Table 1: Thematic synthesis of the 33 reviewed WWW '26 contributions across Main Track and Companion workshops.

Theme	Key Shift at WWW '26	Representative Works	Open Gap
1. Ecosystem Pollution	From single fake articles to search/RAG collapse and Q&A degradation	[2, 4, 14, 21, 32]	Longitudinal web-scale telemetry
2. Detection	From in-domain text classifiers to multimodal, unseen-domain, and adversarial robustness	[3, 5, 7–9, 11, 12, 18, 23, 24, 27, 31, 33]	Cross-generator transfer under realistic evasion
3. Fact-Checking	From single-pass prompts to denoising retrieval, multi-agent debate, and uncertainty gating	[26, 28–30]	Latency, cost, and real-world calibration
4. Human Factors	From demographic stereotypes to lateral reading, source monitoring, and proxy audits	[1, 6, 13, 19, 22]	Linking perception to real platform behavior
5. Provenance & Reproducibility	From ad-hoc detectors to watermarks, on-device/edge provenance, and reproducible pipelines	[10, 15–17, 20, 25]	Adoption incentives and metadata stripping

verification and metadata-based AI detection aligned with EU AI Act and DSA transparency goals.

At the level of research governance and methodology, the R2CASS workshop [17] centered on computational reproducibility when studying volatile web and social media data. Within R2CASS, Loth et al. [15] report the first wave of an expert perception survey ($N = 21$ across AI research, policy, journalism, and industry), documenting skepticism toward standalone technical detectors alongside strong preference for provenance standards, regulatory frameworks, and standardized reproducibility checklists. Directly alongside this call, Shah et al. [20] investigate whether prompt-based and agent-based LLM workflows can automate the repair and reproduction of published computational social science analyses that fail due to missing dependencies, brittle paths, or analytical logic gaps—underscoring that reproducibility itself remains an unsolved engineering bottleneck in web research.

7 Cross-Cutting Synthesis and Open Research Agenda

Table 1 maps the 33 synthesized WWW '26 works across the five thematic clusters, highlighting the dominant shift in problem framing and the primary remaining limitation in each area.

Across these five themes, three overarching conclusions emerge from WWW '26:

1. **Artifact-level accuracy is no longer sufficient.** Ecosystem studies on retrieval collapse [32], black-hat SEO against LLM search [4], and adversarial sentiment attacks [23] show that gains on static benchmarks can coexist with systemic degradation on the open web.

2. **Verification is converging on layered socio-technical stacks.** No single intervention dominates: automated multi-agent fact-checking [28–30], watermarking and cryptographic provenance [10, 16, 25], and human behavioral strategies such as lateral reading and source monitoring [1, 13] address complementary failure points.

3. **Measurement validity and reproducibility lag behind capability.** Both expert surveys [15] and empirical reproducibility audits [20]—as well as reliability evaluations of synthetic survey respondents [22]—indicate that web integrity research needs shared benchmarks, transparent reporting of operational choices, and reproducible code/data artifacts before policy claims can rest on automated evaluations.

Four Priorities for 2027. Looking ahead to the 2027 conference cycle, four priorities stand out: (i) evaluating detectors and fact-checkers under realistic cross-generator, cross-domain, and adversarial-evasion regimes rather than isolated benchmarks; (ii) pairing provenance standards (such as C2PA and robust watermarks) with usability studies of how ordinary users interpret provenance indicators during lateral reading; (iii) auditing when LLM-based agents and synthetic respondents preserve or distort human variance; and (iv) treating computational reproducibility packages as a standard requirement in web integrity workshops and tracks.

Note on Authorship and Scope

This field synthesis was compiled as an independent post-conference technical report documenting the information-integrity research landscape at The ACM Web Conference 2026 (WWW '26), including the author's four accepted workshop contributions across AiOfAi, TrustFM, R2CASS, and the canceled FIIR workshop. All external citations were verified against official ACM/Crossref proceedings metadata and abstracts.

References

[1] Elena Artemenko, Olessia Koltsova, Maksim Terpilovskii, and Natalia Khazova. Fact-checking strategies in web: Lateral navigation in the browser boosts health misinformation detection even under confirmation bias. In *Proceedings of the ACM Web Conference 2026*, pages 8633–8636. ACM, 2026. doi: 10.1145/3774904.3792927.

[2] Esma Aïmeur, Rim Ben Salem, Dorsaf Sallami, Julita Vassileva, and Nora Boulahia-Cuppens. Aiofai: 4th workshop on adverse impacts and collateral effects of artificial intelligence technologies. In *Companion Proceedings of the ACM Web Conference 2026*, pages 279–283. ACM, 2026. doi: 10.1145/3774905.3795468.

[3] Junyi Chen, Qian Liu, Jing Sun, and Yi Zhang. Truelens: Video fake news detection with dual level evidence gathering and consolidation. In *Proceedings of the ACM Web Conference 2026*, pages 7178–7188. ACM, 2026. doi: 10.1145/3774904.3792362.

[4] Pei Chen, Geng Hong, Xinyi Wu, Mengying Wu, Zixuan Zhu, Mingxuan Liu, Baojun Liu, Mi Zhang, and Min Yang. Unveiling the resilience of llm-enhanced search engines against black-hat seo manipulation. In *Proceedings of the ACM Web Conference 2026*, pages 2962–2973. ACM, 2026. doi: 10.1145/3774904.3792399.

[5] Siyuan Chen, Usman Naseem, and Mehwish Nasim. No evidence, no problem: When less is more for out of context multimodal misinformation detection. In *Companion Proceedings of the ACM Web Conference 2026*, pages 909–916. ACM, 2026. doi: 10.1145/3774905.3795462.

[6] Haoang Chi, Qi Wang, Jiantong Jiang, Jiangchao Yao, Feng Liu, and Bo Han. Trustworthy foundation models for web intelligence: Causal perspectives and challenges. In *Companion Proceedings of the ACM Web Conference 2026*, pages 1123–1124. ACM, 2026. doi: 10.1145/3774905.3795831.

[7] Carmela Comito, Annalisa Socievole, and Ester Zumpano. Proactive misinformation forecasting via incremental topic detection and sentiment polarization. In *Companion Proceedings of the ACM Web Conference 2026*, pages 391–398. ACM, 2026. doi: 10.1145/3774905.3794646.

[8] Shuhan Cui, Ruimin Chu, Hanrui Wang, Patrick H. Chen, Ching-Chun Chang, and Isao Echizen. An 8-way taxonomy for multimodal disinformation and detection benchmark. In *Proceedings of the ACM Web Conference 2026*, pages 8545–8548. ACM, 2026. doi: 10.1145/3774904.3792906.

[9] Piyush Ghasiya and Kazutoshi Sasahara. From generation to detection: Leveraging empirically derived linguistic hints for llm-based fake news detection. In *Companion Proceedings of the ACM Web Conference 2026*, pages 473–483. ACM, 2026. doi: 10.1145/3774905.3796489.

[10] Yicheng Huang, Xinyu Xiao, Jian Zhang, Shuhan Qi, Yulin Wu, and Xuan Wang. Awma-moe: Attention-guided watermark adapter with moe for latent diffusion models. In *Proceedings of the ACM Web Conference 2026*, pages 8533–8536. ACM, 2026. doi: 10.1145/3774904.3792903.

[11] Chaewon Kang, Youjin Lee, and Jinyoung Han. Haute: Harmonizing action units with temporal-contextual embeddings for deepfake detection. In *Proceedings of the ACM Web Conference 2026*, pages 8597–8600. ACM, 2026. doi: 10.1145/3774904.3792919.

[12] Qian Liu, Zongxiong Chen, Sonja Schimmler, and Manfred Hauswirth. Mixture of orthogonal decomposition for semantically generalisable multimodal deepfake detection with deepfake manifold disentanglement. In *Companion Proceedings of the ACM Web Conference 2026*, pages 959–964. ACM, 2026. doi: 10.1145/3774905.3796193.

[13] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. Eroding the truth-default: A causal analysis of human susceptibility to foundation model hallucinations and disinformation in the wild. In *Companion Proceedings of the ACM Web Conference 2026*, pages 1125–1132. ACM, 2026. doi: 10.1145/3774905.3795832.

[14] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. Industrialized deception: The collateral effects of llm-generated misinformation on digital ecosystems. In *Companion Proceedings of the ACM Web Conference 2026*, pages 295–302. ACM, 2026. doi: 10.1145/3774905.3795471.

[15] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. The verification crisis: Expert perceptions of genai disinformation and the case for reproducible provenance. In *Companion Proceedings of the ACM Web Conference 2026*, pages 980–988. ACM, 2026. doi: 10.1145/3774905.3795484.

[16] Alexander Loth, Dominique Conceicao Rosario, Peter Ebinger, Martin Kappes, and Marc-Oliver Pahl. Origin lens: A privacy-first mobile framework for cryptographic image provenance and AI detection, 2026. Accepted at the Future of Information Integrity Research Workshop (FIIR, WWW '26 Companion; workshop canceled due to regional conflict). Available at https://arxiv.org/abs/2602.03423.

[17] Fakhri Momeni, Arnim Bleier, Danilo Dessì, and Muhammad Taimoor Khan. The 2nd workshop on social science meets web data: Reproducible and reusable computational approaches (r2cass). In *Companion Proceedings of the ACM Web Conference 2026*, pages 967–969. ACM, 2026. doi: 10.1145/3774905.3795479.

[18] Valentin Noël and Elimane Yassine Seidou. Rega: Zero-overhead graph alignment for structural hallucination detection without generation. In *Companion Proceedings of the ACM Web Conference 2026*, pages 617–625. ACM, 2026. doi: 10.1145/3774905.3794657.

[19] Mihir Pandey, Shail Shankar, and Suyash Mishra. Exploring the domain sensitivity of ai chatbots in correcting misinformation. In *Companion Proceedings of the ACM Web Conference 2026*, pages 492–495. ACM, 2026. doi: 10.1145/3774905.3796491.

[20] Syed Mehtab Hussain Shah, Frank Hopfgartner, and Arnim Bleier. Automating computational reproducibility in social science: Comparing prompt-based and agent-based approaches. In *Companion Proceedings of the ACM Web Conference 2026*, pages 989–998. ACM, 2026. doi: 10.1145/3774905.3795485.

[21] Dongwon Shin and Sooel Son. LLMs killed Q&A stars? analyzing the impact of LLM-generated answers on an online Q&A platform. In *Proceedings of the ACM Web Conference 2026*, pages 4929–4939. ACM, 2026. doi: 10.1145/3774904.3792739.

[22] Erika Elizabeth Taday Morocho, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. Assessing the reliability of persona-conditioned llms as synthetic survey respondents. In *Companion Proceedings of the ACM Web Conference 2026*, pages 320–329. ACM, 2026. doi: 10.1145/3774905.3795477.

[23] Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. Robust fake news detection using large language models under adversarial sentiment attacks. In *Proceedings of the ACM Web Conference 2026*, pages 1717–1726. ACM, 2026. doi: 10.1145/3774904.3792606.

[24] Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM Web Conference 2026*, pages 830–838. ACM, 2026. doi: 10.1145/3774905.3795453.

[25] Marco Vela-Koentarjo and Madhusudan Singh. Decentralized video provenance at the edge: Blockchain-assisted integrity for security camera streams. In *Companion Proceedings of the ACM Web Conference 2026*, pages 1551–1555. ACM, 2026. doi: 10.1145/3774905.3794702.

[26] Ashwin Vinod and Katrin Erk. Uncertainty-aware web-conditioned scientific fact-checking. In *Proceedings of the ACM Web Conference 2026*, pages 8729–8732. ACM, 2026. doi: 10.1145/3774904.3792951.

[27] Chi Wang, Min Gao, Zongwei Wang, Junwei Yin, Kai Shu, and Chenghua Lin. Prompt-induced linguistic fingerprints for llm-generated fake news detection. In *Proceedings of the ACM Web Conference 2026*, pages 7633–7644. ACM, 2026. doi: 10.1145/3774904.3792723.

[28] Zongwei Wang, Min Gao, Junliang Yu, Tong Chen, and Chenghua Lin. Pamas: Self-adaptive multi-agent system with perspective aggregation for misinformation detection. In *Proceedings of the ACM Web Conference 2026*, pages 6977–6988. ACM, 2026. doi: 10.1145/3774904.3792166.

[29] Jun Yang, Yuhan Bai, Dandan Song, Zhijing Wu, and Yuhang Tian. A fact-checking framework with denoising evidence retrieval and llm-based debate verification. In *Proceedings of the ACM Web Conference 2026*, pages 2094–2104. ACM, 2026. doi: 10.1145/3774904.3792285.

[30] Shuo Yang, Zijian Yu, Zhenzhe Ying, Yuqin Dai, Guoqing Wang, Jun Lan, Jinfeng Xu, Jinze Li, and Edith C.H. Ngai. Rama: Retrieval-augmented multi-agent framework for misinformation detection in multimodal fact-checking. In *Companion Proceedings of the ACM Web Conference 2026*, pages 416–424. ACM, 2026. doi: 10.1145/3774905.3796483.

[31] Xuankai Yang, Yan Wang, Jiajie Zhu, Pengfei Ding, Hongyang Liu, Xiuzhen Zhang, and Huan Liu. Cross-domain fake news detection on unseen domains via llm-based domain-aware user modeling. In *Proceedings of the ACM Web Conference 2026*, pages 9494–9505. ACM, 2026. doi: 10.1145/3774904.3792990.

[32] Hongyeon Yu, Dongchan Kim, and Young-Bum Kim. Retrieval collapses when ai pollutes the web. In *Proceedings of the ACM Web Conference 2026*, pages 8745–8748. ACM, 2026. doi: 10.1145/3774904.3792955.

[33] Zhi Zeng, Yifei Yang, Jiaying Wu, Xulang Zhang, Xiangzheng Kong, Herun Wan, Zihan Ma, and Minnan Luo. From manipulation to mistrust: Explaining diverse micro-video misinformation for robust debunking in the wild. In *Proceedings of the ACM Web Conference 2026*, pages 7621–7632. ACM, 2026. doi: 10.1145/3774904.3792721.