VulGen: Workshop on Vulnerabilities in Generative Systems for Information Retrieval

Shuoqi Sun
RMIT University
Naarm / Melbourne, VIC, Australia
shuoqi.sun@student.rmit.edu.au

Sara Allawati
RMIT University
Naarm / Melbourne, VIC, Australia
sara.allawati@student.rmit.edu.au

Laura Dietz
University of New Hampshire
Durham, NH, U.S.A
dietz@cs.unh.edu

Madhurima Khirbat
RMIT University
Naarm / Melbourne, VIC, Australia
madhurima.khirbat@student.rmit.edu.au

Bhaskar Mitra
Independent Researcher
Tiohtià:ke / Montréal, Canada
bhaskar.mitra@acm.org

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Damiano Spina
RMIT University
Naarm / Melbourne, VIC, Australia
damiano.spina@rmit.edu.au

Abstract

Generative systems are rapidly transforming both academic research and industrial practices. These systems are increasingly integrated into information access and information retrieval (IR) tasks and continue to evolve at a substantial pace. Integrating these models into daily workflows exposes critical vulnerabilities, including adversarial attacks, inherent biases, and negative impacts on user behavior, which can lead to suboptimal or even detrimental outcomes. The *VulGen* workshop at SIGIR 2026 brings together the IR community and related disciplines (e.g., cyber security) to map this evolving landscape. Through a full day of structured discussion and engagement, we aim to synthesize the current state of research and identify new avenues for future investigation. Information about VulGen is hosted at: https://vulgen-workshop.github.io/SIGIR2026/.

CCS Concepts

• **Information systems** → **Information retrieval**.

Keywords

Vulnerability; generative system; information security

ACM Reference Format:
Shuoqi Sun, Sara Allawati, Laura Dietz, Madhurima Khirbat, Bhaskar Mitra, Maarten de Rijke, and Damiano Spina. 2026. VulGen: Workshop on Vulnerabilities in Generative Systems for Information Retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3805712.3808654

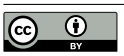
1 Motivation and Theme

Generative systems, which typically include large language models (LLMs), retrieval-augmented generation (RAG), and various integrated tools (e.g., URL interpreter) [5], have quickly gained popularity and reshaped information retrieval (IR) research and system design. Established evidence shows strong effectiveness of generative systems in various IR tasks, such as information seeking (IS) [1], ranking, query argumentation, LLM-as-a-judge, etc. Their applications have been well explored as replacements for classical practices and can be foreseen as future standard baselines. For example, Shah and White [9] propose an agentic ecosystem where LLM-powered agents can potentially solve highly personalized informational tasks and communicate with other agents.

Alongside the pursuit of effectiveness in IR systems, their vulnerabilities have also been researched and remained central concerns (e.g., black-hat search engine optimization (SEO) that may be used to manipulate search engine results). Vulnerabilities in *generative systems* require even more dedicated discussion given their rapidly emerging IR applications, such as LLM use for query-document relevance judgment, RAG-based question answering (Q&A) systems, and personalized agents [12]. As a result, the rise of LLMs and their downstream integrations has introduced novel IR vulnerabilities that necessitate further attention. Instances include LLM narcissism [3], also known as *self-preference bias* [11], where LLMs favor self-generated content and cause bias for query-document relevance judgment task; as well as biases arising from synthetic training data (e.g., generated query and document pairs) [8]. Parallel to academic research, the European Union has adopted the "AI Act," a legal framework that addresses vulnerabilities of generative systems and aims to foster trustworthy, human-centric use.[1]

We propose an exploratory workshop where participants brainstorm together. As generative systems are still evolving, the formal concept and scope of vulnerability have not yet unified. To our knowledge, vulnerability research is typically associated with the topic of "under what circumstances, where a generative system could go wrong," where the "circumstances" and "go wrong" are what we primarily care about in exploration. We interpret three directions: (1) when provided with input from their users, generative systems can be attacked, such as corpus poisoning [10], prompt

[1] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808654

injection [6], and SEO; (2) when influencing their users, generative systems can lead to undesirable patterns, such as misinformation dissemination (i.e., hallucination); (3) when they are building blocks of architectures and pipelines, generative systems can produce biases or unknown consequences, such as preference bias towards LLM-generated content [2, 3], or the consequences of generative content streaming back to the web [4]. However, the concept still remains unclear. As one of the objectives of VulGen workshop, we expect participants to discover real-world cases through breakout discussions, so that we can better understand the scope of vulnerability research.

In summary, VulGen workshop has three primary objectives:

(1) conceptualizing, scoping and increasing awareness of generative IR system vulnerabilities;
(2) fostering an interdisciplinary community of academics and industry practitioners at all career stages to advance the understanding of current vulnerabilities and future research directions for generative systems; and
(3) exploring and discussing state-of-the-art vulnerabilities, and sparking potential new lines of research and collaboration;

We believe that the ACM SIGIR conference is an appropriate venue for the VulGen workshop, which brings together IR researchers who are working on or interested in this topic to explore and discuss it together. With the participants' consent, we will summarize the findings in a SIGIR Forum report.

2 Target Audience

We anticipate attracting researchers dedicated to identifying pressing vulnerabilities of generative systems. Specifically, we target those investigating: (1) the impact of emerging generative technologies on user behavior and information perception, with a focus on adverse effects such as misinformation, over-reliance, and so on; and (2) systemic flaws, including adversarial vulnerabilities and systemic biases. We welcome a diverse audience of researchers and practitioners at all career stages. To foster interdisciplinary dialogue, we encourage participation from fields like cyber security, natural language processing, and computer vision. Outreach will leverage social media (LinkedIn, X, BlueSky, Slack), mailing lists (e.g., SIGIR-List), and a dedicated workshop website.[2]

3 Call for Extended Abstract

We welcome two-page submissions of *Extended Abstract* (excluding references, acknowledgments, and positionality statements) formatted in the ACM SIGIR template. We especially value theoretical, visionary perspectives and practical vulnerability demonstrations. Submission themes include, but are not limited to:

- System Vulnerability Statements: Novel attack schemas and the identification of flaws or biases within current system architectures and applications.
- User-Centric Statements: Research on how vulnerabilities of generative systems influence user behavior and information perception, with an emphasis on negative impact.
- Resources or Practical Demonstrations: Introduction of evaluation benchmarks, platforms, or open-source tools for the community; real-world case studies, including demonstrations of state-of-the-art attacks or exploits.
- Perspective or Theoretical Statements: Identifying new research trajectories, raising awareness of specific vulnerabilities and lacking resources, or proposing novel methodologies for vulnerability research; frameworks and models for characterizing and formalizing vulnerabilities in generative systems.
- Surveys and Reviews: Comprehensive overviews of the current landscape of generative system vulnerabilities.
- Industrial Perspectives: Hypotheses and practices regarding the real-world vulnerabilities of generative systems in production.

Table 1: Representative applications and vulnerabilities.

Application/Vulnerability	Representative Example(s)
Representative Applications	
LLMs for Information Seeking	Question Answering (Q&A) systems and chatbots.
LLM-as-a-Judge	Automated evaluations and data annotations.
LLMs for Ranking	LLM-based rankers and ranking systems.
LLMs for Query Augmentation	Query variant generations and expansions.
LLMs for Simulation	Synthetic dataset generations.
Agentic IR Systems	Autonomous executors and agentic ecosystems.
Personalized LLMs	Personalized assistants.
Generative Media	Video and audio generations.
Representative Vulnerabilities	
Black-Hat SEO	Document and query injections.
Adversarial Attacks	Corpus positioning and backdoor attacks.
Systemic Bias or Consequences	LLM narcissism and inherent output biases.
Misleading or Harmful Output	Misinformation and polarized content generations.
Human-Centered Impacts	Intellectual laziness.

To guide our target audience, we also outline relevant representative applications and vulnerabilities of generative systems alongside the submission themes, which are summarized in Table 1.

Selection Process. The selection process will be single-blind. The Program Committee (see Section 6) will evaluate submissions based on: (1) relevance to the workshop themes; (2) significance of statements; and (3) the potential to stimulate meaningful discussion. Authors of accepted submissions will be invited to present at the workshop. We also encourage researchers without accepted submissions to attend and participate, subject to venue capacity. Accepted extended abstracts will be *non-archival*. PDFs of the accepted submissions will be hosted on the workshop's webpage. We encourage authors to upload their accepted works to arXiv, and we will include links to these preprints on the webpage.

[2]VulGen workshop website: https://vulgen-workshop.github.io/SIGIR2026/

Table 2: VulGen workshop schedule.

Time	Events
9:00 – 9:15	Opening Remarks
9:15 – 10:30	Lightning Presentation
10:30 – 11:00	*Coffee Break*
11:00 – 12:00	Poster and Demonstrations
12:00 – 12:30	Breakout Group Formation
12:30 – 14:00	*Lunch - Breakout Discussion*
14:00 – 14:30	Keynote by Christopher Leckie
14:30 – 15:00	Breakout Discussion
15:00 – 16:00	Discussion Reporting
16:00 – 16:30	*Coffee Break*
16:30 – 17:15	Panel Session
17:15 – 17:30	Closing Remarks

4 Workshop Format

We propose a *full-day* workshop incorporating multiple modalities to provide a diverse and engaging program. The tentative schedule is summarized in Table 2, with detailed descriptions provided below.

4.1 Presentations and Showcases

Accepted submissions will be featured in a Lightning Presentation session. Each presentation will last between five and ten minutes, with the exact duration and the inclusion of individual Q&A segments contingent upon the total number of accepted papers.

Presenters are also encouraged to submit a poster or demo for the showcase session to engage more freely. To foster a broader community, we are open to coordinating joint poster and demo sessions with other concurrent workshops.

4.2 Breakout Discussion

The workshop will facilitate a structured *Breakout Group Discussion*. Participants will first identify topics of shared interest (e.g., ranking, LLM-as-a-judge, user-centric topics) to form thematically aligned groups. Groups will then discover, discuss, and characterize novel vulnerabilities using a standardized case-study format (context, target IR task, and vulnerability description), drawing upon provided reference materials [e.g., 3, 7]. Finally, groups will synthesize their discussions and present their findings in a plenary session.

4.3 Keynote and Panel Session

Keynote. Professor Christopher Leckie from The University of Melbourne will deliver a interdisciplinary keynote, who has broad expertise in developing AI for domains such as cyber security.[3]

Title: *Emerging Security Threats from Generative AI in Information Seeking Environments*

Abstract: Progress in Generative AI (GenAI) is enabling new ways for organisations to support the needs of diverse user communities in information seeking environments. For example, voice-enabled chatbot interfaces are making it possible for users to access information and services within an organisation in a flexible and unstructured manner. However, the flexibility and openness of such GenAI interfaces are also creating a new type of attack surface that malicious actors can exploit to attack organisations. The risk of such attacks is compounded by the pressure on organisations to rapidly provide AI-based interfaces for customer access at a time when their testing and assurance processes for these interfaces are still relatively immature. In this talk we will highlight some of the emerging trends for malicious misuse of GenAI in information seeking environments, with a specific focus on threats involving voice-enabled chatbots in conversational settings. We will also present our current research initiative, which is the development of a GenAI test range that can enable research on testing, detecting and defending against these rapidly emerging threats.

Panel Session. We will host a joint panel session with another SIGIR 2026 workshop: *Justice, Emancipation, Democracy, and Information Access (JEDI): The SIGIR Workshop on Resisting Corporate and Authoritarian Capture of Information Access Platforms.*[4] While JEDI explores the impacts of IR systems through critical sociopolitical lens, both VulGen and JEDI share a core commitment to ensure that IR research serves broader societal interests. To explore these intersecting goals, we will invite two or three panelists from both organizing committees. This collaborative format will provide participants with diverse perspectives from two dedicated research communities and attract a broader audience across SIGIR.

5 Workshop Organizers

All workshop organizers are planning to attend SIGIR'26 in person.

- **Shuoqi Sun** is a PhD Candidate at RMIT University and affiliated with ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) working on effectiveness and vulnerabilities of LLM-based retrieval systems. He is a member of the RMIT-ADMS team who won the 1st prize at the SIGIR 2025 LiveRAG Challenge and 1st prize at the dynamic evaluation task of the NeurIPS 2025 MMU-RAG Challenge.
- **Sara Allawati** is a PhD Candidate at RMIT University and affiliated with ADM+S, working on interactive information retrieval in the era of GenAI. Sara is part of the ADM+S executive committee and Emerging Professionals Committee at the Australian Computer Society. In both roles, she takes part in organizing various community-building events.
- **Prof.Dr. Laura Dietz** is an Associate Professor of Computer Science at the University of New Hampshire, USA. She has authored several seminal papers on the use of LLMs for automatic evaluation, including analyses of their vulnerabilities and risks. She has organized multiple workshops and tutorials, including *Utilizing Knowledge Graphs for Information Retrieval* at SIGIR 2017 and WSDM 2017, and *Neuro-Symbolic Representations for Information Retrieval* at ECIR 2023 and SIGIR 2023.
- **Madhurima Khirbat** is a PhD candidate at RMIT University and affiliated with ADM+S, working on the evaluation of LLM-based recommender systems. She is interested in responsible AI and moderates the Recommender Systems and Responsible AI (R2AI) group at RMIT University.
- **Dr. Bhaskar Mitra** is an IR researcher based in Tiohtià:ke / Montréal, Canada. His research focuses on AI-mediated online information access and questions of social justice and emancipation in the context of these sociotechnical systems. He is currently serving as the ACM SIGIR Secretary. He co-organized several

[3] https://findanexpert.unimelb.edu.au/profile/6335-christopher-leckie

[4] https://jedi.inertial.science/sigir2026/

workshops (Neu-IR @ SIGIR 2016–2017, HIPstIR 2019, Search Futures @ ECIR2024, LLM4Eval @ SIGIR 2024–2025), shared evaluation tasks (TREC Deep Learning Track 2019–2023, TREC Tip-of-the-Tongue Track 2023–2024, and MS MARCO ranking leaderboards), and tutorials (WSDM 2017–2018, SIGIR 2017, ECIR 2018, and AFIRM 2019–2020).

- **Prof.Dr. Maarten de Rijke** is a Distinguished University Professor of AI and IR at the University of Amsterdam. He is also the Scientific Director of the national Innovation Center for AI in The Netherlands, where he leads a large-scale program on trustworthy AI-based systems. He is also affiliated with the ADM+S program. He is a serial workshop and tutorial co-organizer, with more than 50 events at all of the main IR conferences to his name.
- **Dr. Damiano Spina** is a Senior Lecturer at the School of Computing Technologies, RMIT University, Australia and an Associate Investigator at ADM+S. His research focuses on interactive information retrieval and evaluation of information access systems, including RAG. He has co-organized workshops in international conferences – including SIGIR, CHIIR, and UbiComp – and shared tasks for evaluation campaigns at CLEF and IberLEF. He is a Local Arrangements Co-Chair for SIGIR'26.

6 Program Committee Members

The Program Committee (PC) is formed by members from both industry and academia:

- Nalin Arachchilage, RMIT University, Australia.
- Pablo Castells, Amazon, USA.
- Charles L. A. Clarke, University of Waterloo, Canada.
- Juwon Kim, Amazon, USA.
- Dario Di Palma, Politecnico di Bari, Italy.
- Udita Patel, Amazon, USA.
- Yongli Ren, RMIT University, Australia.
- Mark Sanderson, RMIT University, Australia.
- Paul Thomas, Microsoft, Australia.

7 Related Workshops

We surveyed major conferences in IR, recommender systems, machine learning, natural language processing, and cyber security. Table 3 summarizes workshops related to ours. Our analysis reveals that none of surveyed workshops focuses on generative system vulnerabilities in IR context and their implications on system design, user experience, and unintended bias in modern information environments. We believe that having the first edition of the VulGen workshop at SIGIR'26 will bridge this gap, sparking new research directions and collaboration within the international IR community.

Table 3: Related workshops (2023–2025). We use abbreviated names (where applicable) for conferences and workshops.

Venue	Related Workshops
Information Retrieval (IR)	
SIGIR'25	Robust IR; LLM4Eval
SIGIR'24	LLM-IGS-II
CIKM'25	Human-Centric AI; MMGenSR
CIKM'24	GTA3; Trustworthy and Responsible AI for IKMS
CIKM'23	LLMIT; Personalized Generative AI
SIGIR-AP'25	R3AG
Recommender Systems	
RecSys'25	FAccTRec; EARL
RecSys'24	ROEGEN
Machine Learning (ML) / Natural Language Processing (NLP) / Cyber Security	
ACL'25	TrustNLP; LLMSEC
NeurIPS'25	Lock-LLM
NeurIPS'24	Towards Safe and Trustworthy Agents
NeurIPS'23	Multi-Agent Security: Security as Key to AI Safety
CCS'25	AISec; HAIPS

Acknowledgments

The authors acknowledge the peoples of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin Nation on whose unceded lands ACM SIGIR 2026 was hosted. We pay our respects to their Elders past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their continuing connection to land, sea, sky, and community.

Shuoqi Sun, Sara Allawati, Madhurima Khirbat, Maarten de Rijke and Damiano Spina are partially supported by the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005). Maarten de Rijke was also supported by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.-1389.20.183, and KICH3.LTP.20.006, and the European Union under grant agreement No. 101201510 (UNITE). Opinions expressed are those of the author(s) only and do not necessarily reflect those of their respective employers, funders and/or granting authorities.

References

[1] Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How People Use ChatGPT*. Working Paper. National Bureau of Economic Research.

[2] Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. Neural Retrievers are Biased Towards LLM-Generated Content. In *Proc. KDD'24*.

[3] Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. In *Proc. ICTIR'25*.

[4] Michele Garetto, Alessandro Cornacchia, Franco Galante, Emilio Leonardi, Alessandro Nordio, and Alberto Tarable. 2025. Information Retrieval in the Age of Generative AI: The RGB Model. In *Proc. SIGIR'25*.

[5] Mohanna Hoveyda, Harrie Oosterhuis, Arjen P. de Vries, Maarten de Rijke, and Faegheh Hasibi. 2025. Adaptive Orchestration of Modular Generative Information Access Systems. In *Proc. SIGIR'25*.

[6] Yang Jiao, Xiaodong Wang, and Kai Yang. 2025. PR-Attack: Coordinated Prompt-RAG Attacks on Retrieval-Augmented Generation in Large Language Models via Bilevel Optimization. In *Proc. SIGIR'25*.

[7] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2025. Robust Neural Information Retrieval: An Adversarial and Out-of-Distribution Perspective. *ACM Trans. Inf. Syst.* 44, 1 (2025).

[8] Hossein A. Rahmani, Varsha Ramineni, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. 2025. Towards Understanding Bias in Synthetic Data for Evaluation. In *Proc. CIKM'25*.

[9] Chirag Shah and Ryen W. White. 2025. From To-Do to Ta-Da: Transforming Task-Focused IR with Generative AI. In *Proc. SIGIR'25*.

[10] Jinyan Su, Preslav Nakov, and Claire Cardie. 2025. Corpus Poisoning via Approximate Greedy Gradient Descent. In *Findings of ACL'25*.

[11] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-Preference Bias in LLM-as-a-Judge. In *NeurIPS 2024 Workshop on SafeGenAi*.

[12] Ryen W. White and Chirag Shah (Eds.). 2025. *Information Access in the Era of Generative AI*. The Information Retrieval Series, Vol. 51. Springer Nature Switzerland, Cham.