GenAIDetect 2025

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)

Proceedings of the Workshop

The 31st International Conference on Computational Linguistics

January 19, 2025
Abu Dhabi, UAE

©2025 International Conference on Computational Linguistics

ISBN 979-8-89176-205-3

Preface

Welcome to the 1st Workshop on GenAI Content Detection (GenAIDetect), co-located with COLING 2025 and hosted in Abu Dhabi, UAE. The GenAIDetect workshop provides a platform to bring together researchers working on all aspects of generative AI content detection across text, image, audio, video, and multimodal data. The aim is to create a space for the entire GenAI content detection community to present and exchange theories, algorithms, software, datasets, and tools.

In its first edition, the workshop offers a rich and diverse full-day program, including keynotes, oral paper, and poster presentation sessions, and a panel discussion. The presented papers cover a broad spectrum of topics, including datasets and benchmarks, watermarking, and various modeling techniques such as graph-based approaches. The workshop also attracted notable contributions across multiple modalities, including image, and text.

In this first edition of the GenAI Content Detection Workshop, we received 20 submissions. Each paper was rigorously peer-reviewed by two to three expert reviewers in the field. Of these submissions, 11 papers were accepted, resulting in a 55% acceptance rate, all of which were selected for oral presentation. Notably, we made no distinction in quality between long and short papers or between oral and poster presentations.

The workshop featured three shared tasks: (1) Binary Multilingual Machine-Generated Text Detection (Human vs. Machine), (2) AI vs. Human – Academic Essay Authenticity Challenge, and (3) Cross-Domain Machine-Generated Text Detection. All tasks were well received, attracting strong participation. For Tasks 1, 2, and 3, we received 17, 7, and 7 system description papers, respectively, resulting in a total of 31 system description papers and 3 overview papers.

The overview paper for the shared tasks was peer-reviewed by at least three expert reviewers, while the system description papers were reviewed by two to three reviewers. The proceedings include research track papers, as well as shared task overview and system description papers.

Finally, we thank all the contributors of papers and the 54 members of the Program Committee for their dedication to providing high-quality reviews in a timely manner. We also extend our gratitude to the COLING 2025 workshop chairs, Katsuhito Sudoh and Mo El-Haj, for organizing the COLING workshop program.

The GenAIDetect Organizers,

Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, George Mikros

Workshop website: https://genai-content-detection.gitlab.io

Organizing Committee

Firoj Alam, Qatar Computing Research Institute, Qatar

Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence, UAE

Nizar Habash, New York University Abu Dhabi, UAE

Iryna Gurevych, Mohamed bin Zayed University of Artificial Intelligence, UAE; Technical University of Darmstadt, Germany

Shammur Chowdhury, Qatar Computing Research Institute, HBKU, Qatar

Artem Shelmanov, Mohamed bin Zayed University of Artificial Intelligence, UAE

Yuxia Wang, Mohamed bin Zayed University of Artificial Intelligence, UAE

Ekaterina Artemova, Toloka AI, Netherlands

Mucahid Kutlu, Qatar University, Qatar

George Mikros, Hamad Bin Khalifa University, Qatar

Program Committee

Hariram Veeramani, UCLA, USA

Md Tanvirul Alam, Rochester Institute of Technology, USA

Sahinur Rahman Laskar, UPES, Dehradun, India

Somnath Banerjee, University of Tartu, Estonia

Noureldin Elmadany, Arab Academy for Science and Technology, Egypt

Md Saiful Islam, University of Alberta, USA

Abdullah I. Alharbi, King Abdulaziz University, Saudi Arabia

Maram Hasanain, Qatar Computing Research Institute, Qatar

Hind AL-Merekhi, Qatar Computing Research Institute, Qatar

Muhammad Tasnim Mohiuddin, Qatar Computing Research Institute, Qatar

Alham Fikri Aji, MBZUAI, UAE

Zhuohan Xie, MBZUAI, UAE

Tarek Mahmoud, MBZUAI, UAE

Jinyan Su, Cornell University, USA

Rui Xing, MBZUAI, UAE

Jiahui Geng, MBZUAI, UAE

Yassine El Kheir, Technical University Berlin, Germany

Giovanni Puccetti, Institute of Information Science and Technology "A. Faedo", Italy

Vladislav Mikhailov, University of Oslo, Norway

Masahiro Kaneko, MBZUAI, UAE

Ryuto Koike, Tokyo Institute of Technology, Japan

Orchid Chetia Phukan, Indraprastha Institute of Information Technology Delhi, India

Koel Dutta Chowdhury, Saarland University, Germany

Saied Alshahrani, Clarkson University, USA

Fahad Shamshad, MBZUAI, UAE
Md Messal Monem Miah, Texas A&M University, USA
Shah Nawaz, Johannes Kepler University Linz, Austria
Dr. Bishwa Ranjan Das, Interscience Institute of Management and Technology, India
Nada Ayman GabAllah, Coventry University, Egypt Branch, Egypt
Kamel Gaanoun, Institut National de Statistiques et d'Economie Appliquée, Morocco
Shailja Thakur, IBM Research, USA
Amr Keleg, University of Edinburgh, UK
Akim Tsvigun, University of Amsterdam, Netherlands
Muhammad Amin Nadim, University of Pegaso, Italy
Marc Franco-Salvador, Symanto Research, Germany
José Ángel González, Symanto Research, Germany
Areg Mikael Sarvazyan, Symanto Research, Germany
Walter Daelemans, University of Antwerp, Belgium
Piotr Przybyła, Universitat Pompeu Fabra, Italy
Adaku Uchendu, MIT Lincoln Laboratory, USA
Sheily Panwar, CUC-Ulster University (Doha), Qatar
Xianjun Yang, UCSB, USA
Fatima Haouari, Qatar University
Fahim Shakil Tamim, IUBAT, Bangladesh
Zach Johnson, Microsoft, USA
Abubakarr Jaye, Microsoft, USA
Jason Lucas, Penn State University, USA
Francisco Rangel, Genaios, Germany
Erum Haris, University of Leeds, UK
Mohammad Ruhul Amin, Fordham University, USA
Ram Mohan Rao Kadiyala, University of Maryland, USA
Nadia Ghezaiel, University of Hail, Saudi Arabia
Md Towhidul Absar Chowdhury, Rochester Institute of Technology, USA
Dominik Macko, Kempelen Institute of Intelligent Technologies (KInIT), Slovakia
Gleb Kuzmin, Laboratory for Analysis and Controllable Text Generation Technologies RAS, Russia
Lyudmila Rvanova, Laboratory for Analysis and Controllable Text Generation Technologies RAS, Russia

Keynote: Detectability of Language Model Generated Content: Myths, Challenges, and Opportunities

Rakesh M. Verma,

University of Houston, Houston, TX, USA

rmverma2@central.uh.edu

Abstract: Large language models (LLMs) have captured the imagination of researchers and users worldwide, and led to a capacity race. We start with some myths on the detectability of LLM generated content. We then consider LLMs through the alignment-utility tradeoff perspective and reveal some surprising consequences of aligning LLM models for different purposes, such as 1) ability to mimic humans on natural language tasks and 2) safety..Throughout the talk, we shall highlight the challenges and opportunities for research with respect to detecting LLM generated content for the different purposes of alignment.

Background: The rapid development of Large language models (LLMs) has captured the imagination of researchers and users worldwide and led to a capacity race with bigger models coming out every couple of months. Concomitantly, there is growing concern about their potential for misuse in several different fields. Maintaining the integrity of digital communication platforms such as Reddit has become increasingly vital due to these advancements in LLM abilities. While the threat of misuse has always been around, the scale at which they can now disseminate disinformation, hate speech, and create spear phishing attacks has grown alarmingly. This development poses significant challenges distinguishing between human and machine-generated content, especially on social media, where generative AI can serious consequences.

In this keynote, we will present some myths on the detectability of LLM generated content. We will also present the strengths of LLMs and their limitations. Notable strengths include the ability to leverage world knowledge that has been documented in their training dataset and new knowledge through retrieval augmented generation (RAG). Notable limitations include staying on task, which has been called different terms such as hallucination, etc., and vulnerability to different kinds of attacks.

In the second part of the talk, we consider LLMs through the alignment-utility perspective and reveal some surprising consequences of aligning LLM models for different purposes, such as 1) ability to mimic humans on natural language tasks and 2) safety. We will examine the challenges of creating datasets to examine the generative capabilities of LLMs and reflect upon our recent effort to create a dataset of tweets involving both censored and uncensored models. The talk will include brief results of our recent experiments with this datasets and four prominent LLMs including Llama and GPT4o,

We will conclude the talk with challenges and opportunities for research with respect to detecting LLM generated content including for the different purposes of alignment.

Acknowledgments: This research is joint with my PhD students at the University of Houston, especially Bryan Tuck and Fatima Z. Qachfar. Research partly supported by NSF grants 2210198 and 2244279 and ARO grants W911NF-20-1-0254 and W911NF-23-1-0191. Verma is the founder of Everest Cyber Security and Analytics, Inc.

Bio: Rakesh Verma is a Professor of Computer Science at the University of Houston, where he teaches a course on security analytics. He has made research contributions in equational logic programming, algorithm design and analysis, computer security, and data science. He is the author of Cybersecurity Analytics (CRC Press, 2019) (Verma and Marchette, 2019); a cybersecurity section associate editor of the Frontiers in Big Data Journal; and co-organizer of the ACM Annual International Workshop on Security and Privacy Analytics (IWSPA) since 2015.

Table of Contents

Workshop Program

Sunday 19th January, 2025

09:00 – 09:10 Opening Remarks

09:10 – 09:50 Keynote: Detectability of Language Model Generated Content: Myths, Challenges, and Opportunities, Rakesh Verma

09:50 - 10:35 Session 1

09:50 – 10:05 GPT-4 is Judged More Human than Humans in Displaced and Inverted Turing Tests
Ishika M. Rathi, Sydney Taylor, Benjamin Bergen and Cameron Jones

10:05 – 10:20 SilverSpeak: Evading AI-Generated Text Detectors using Homoglyphs
Aldan Creo and Shushanta Pudasaini

10:20 – 10:35 Mirror Minds: An Empirical Study on Detecting LLM-Generated Text via LLMs
Josh Baradia, Shubham Gupta and Suman Kundu

10:35 – 11:00 Coffee Break

11:00 - 12:00 Session 2

11:00 – 11:15 Human vs. AI: A Novel Benchmark and a Comparative Study on the Detection of Generated Images and the Impact of Prompts
Philipp Moeßner and Heike Adel

11:15 – 11:30 Benchmarking AI Text Detection: Assessing Detectors Against New Datasets, Evasion Tactics, and Enhanced LLMs
Shushanta Pudasaini, Luis Miralles, David Lillis and Marisa Llorens Salvador

11:30 – 11:45 Cross-table Synthetic Tabular Data Detection
G. Charbel N. Kindji, Lina M. Rojas Barahona, Elisa Fromont and Tanguy Urvoy

11:45 – 12:05 GenAI Content Detection Task 1: English and Multilingual Machine-Generated Text Detection: AI vs. Human
Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych and Preslav Nakov

12:05 – 13:30 Lunch Break

13:30 - 15:25 Session 3

13:30 – 13:50 GenAI Content Detection Task 2: AI vs. Human – Academic Essay Authenticity Challenge
Shammur Absar Chowdhury, Hind Almerekhi, Mucahid Kutlu, Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin, George Mikros and Firoj Alam

13:50 – 14:10 GenAI Content Detection Task 3: Cross-Domain Machine Generated Text Detection Challenge
Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki and Chris Callison-Burch

14:10 – 14:25 DAMAGE: Detecting Adversarially Modified AI Generated Text
Elyas Masrour, Bradley N. Emi and Max Spero

14:25 – 14:40 Text Graph Neural Networks for Detecting AI-Generated Content
Andric Valdez and Helena Gomez-Adorno

14:40 – 14:55 I Know You Did Not Write That! A Sampling Based Watermarking Method for Identifying Machine Generated Text
Kaan Efe Keleş, Ömer Kaan Gürbüz and Mucahid Kutlu

14:55 – 15:10 The Consistent Lack of Variance of Psychological Factors Expressed by LLMs and Spambots
Vasudha Varadarajan, Salvatore Giorgi, Siddharth Mangalik, Nikita Soni, Dave M. Markowitz and H. Andrew Schwartz

15:10 – 15:25 Your Large Language Models are Leaving Fingerprints
Hope Elizabeth McGovern, Rickard Stureborg, Yoshi Suhara and Dimitris Alikaniotis

15:30 – 16:30 Break + Poster Session

16:30 - 17:30 Session 4

16:30 - 16:40 SzegedAI at GenAI Detection Task 1: Beyond Binary - Soft-Voting Multi-Class Classification for Binary Machine-Generated Text Detection Across Diverse Language Models *Mihaly Kiss and Gábor Berend*

16:40 - 16:50 Advacheck at GenAI Detection Task 1: AI Detection Powered by Domain-Aware Multi-Tasking *German Gritsai, Anastasia Voznuyk, Ildar Khabutdinov and Andrey Grabovoy*

16:50 - 17:00 IntegrityAI at GenAI Detection Task 2: Detecting Machine-Generated Academic Essays in English and Arabic Using ELECTRA and Stylometry *Mohammad ALSmadi*

17:00 - 17:10 CNLP-NITS-PP at GenAI Detection Task 3: Cross-Domain Machine-Generated Text Detection Using DistilBERT Techniques *Sai Teja Lekkala, Annepaka Yadagiri, Mangadoddi Srikar Vardhan and Partha Pakray*

17:10 - 17:20 BBN-U.Oregon's ALERT system at GenAI Content Detection Task 3: Robust Authorship Style Representations for Cross-Domain Machine-Generated Text Detection *Hemanth Kandula, Chak Fai Li, Haoling Qiu, Damianos Karakos, Hieu Man, Thien Huu Nguyen and Brian Ulicny*

17:20 - 17:30 Leidos at GenAI Detection Task 3: A Weight-Balanced Transformer Approach for AI Generated Text Detection Across Domains *Abishek R. Edikala, Gregorios A. Katsios, Noelie Creaghe and Ning Yu*

17:30 – 17:45 Best Paper Award + Closing Ceremony