

ACL 2026

The 10th Workshop on Online Abuse and Harms (WOAH)

Proceedings of the Workshop

October 29, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-487-3

Introduction

Digital technologies have brought significant benefits to society, transforming how people connect, communicate, and interact. However, these same technologies have also enabled the dissemination and amplification of abusive and harmful content, such as hate speech, harassment, and misinformation. Given the increased volume of content shared online, addressing abuse and harm at scale requires the use of computational tools. Nevertheless, detecting and moderating online abuse is not only a technical problem, but remains a complex task with social, legal, and ethical challenges. The **Workshop on Online Abuse and Harms (WOAH)** is one of the key venues dedicated to addressing these challenges through interdisciplinary research and dialogue.

WOAH invites contributions from a range of fields, including natural language processing, machine learning, computational social science, law, and social science fields. We explicitly encourage interdisciplinary and cross-sectoral submissions, including both technical and non-technical work, as well as research focusing on under-resourced languages and marginalized communities. We also provide space for non-archival submissions and reports from civil society organizations to foster collaboration between academic researchers and practitioners.

The **10th edition of WOA**H will take place on **October 29th, 2026** as part of **EMNLP 2026 in Budapest, Hungary**. The special theme is **Ten Years of WOA**H: **Reflecting on Progress and New Frontiers**. We reflect back on the past decade, as the workshop has developed and responded to shifting online harms, from an early focus on text-based hate speech and harassment to more complex phenomena spanning multiple modalities. The scope of the workshop has expanded in several ways, considering evolving online harms and technologies and their impacts across difficult contexts, communities, and cultures. The development of recommendation systems, shifting platform ecosystems, and generative AI have all expanded the scale, scope, and diversity of harms. In response to these changes, research in WOA

H has addressed topics including misinformation, radicalisation, child sexual exploitation, algorithmic bias, and privacy. The workshop has also charted several methodological turns, moving from monolingual to multilingual and unimodal to multimodal, embracing advances in deep learning while emphasizing interpretability. Despite this progress, there are still fundamental challenges. There is still limited consensus on what constitutes “harm”, how context is used or thresholds are set, or the extent to which these factors vary across cultures. Not only do these issues threaten the validity and generalizability of datasets and models, but they can result in further discrimination and marginalization of affected communities. Research has frequently prioritised detection, high-resource languages, and narrowly defined phenomena over intervention, global perspectives, and systemic or structural harms, with insufficient attention to user agency, platform incentives, lived experience, and participatory approaches. These issues call for critical self-reflection and a continued focus on improvement. As WOA

H continues to be at the forefront of efforts to address online harms, this edition reflects back on the lessons from the past decade, seeking to consolidate our collective knowledge, identify enduring gaps, and better connect research, practice, and policy to guide work on the next generation of online harms.

This year’s program features a selection of high-quality papers, presented through poster sessions and lightning talks, alongside keynote presentations from distinguished researchers and practitioners in the field. In total, we received 77 archival and 7 non-archival submissions, of which we accepted 39 (51%) and 6 (86%), respectively. These works will be showcased through poster sessions and lightning talks, facilitating both in-person and online participation, alongside keynote presentations delivered by Sara Tonelli, Su Lin Blodgett, Sarah T. Roberts and Zeerak Talat.

We thank all our participants and reviewers for their work, and our sponsors for their support. We hope you enjoy this year’s WOA

H and the research published in these proceedings. We hope that WOA

H 2026 serves as a platform for productive discussions, meaningful collaborations, and continued progress in addressing online abuse and harms.

Enjoy the workshop!

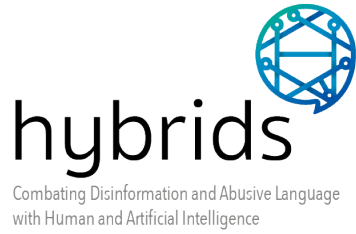
The WOA 2026 Organizers

Agostina Calabrese (Cohere), Thomas Davidson (University of North Carolina at Chapel Hill), Christine De Kock (University of Melbourne), Urja Khurana (Delft University of Technology), Marta Marchiori Manerba (University of Turin), Paloma Piot (Universidade da Coruña), Zeerak Talat (University of Edinburgh)

Sponsors

WOAH is grateful for support from the following sponsors:

Gold Tier



**Funded by
the European Union**

Organizing Committee

Workshop Organizer

Agostina Calabrese, Cohere

Thomas Davidson, University of North Carolina at Chapel Hill

Christine de Kock, University of Melbourne

Urja Khurana, Delft University of Technology

Marta Marchiori Manerba, University of Turin

Paloma Piot, Universidade da Coruña

Zeera Talat, University of Edinburgh

Program Committee

Chairs

Agostina Calabrese, Cohere
Thomas Davidson, University of North Carolina at Chapel Hill
Christine de Kock, University of Melbourne
Urja Khurana, TU Delft
Marta Marchiori Manerba, University of Turin
Paloma Piot, Universidade da Coruña
Zeera Talat, University of Edinburgh

Program Committee

Gavin Abercrombie, Heriot Watt University
Syed Sarfaraz Akhtar, Apple Inc
Iuliia Alieva, University of Stuttgart
Diego Alves, Saarland University
Jisun An, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington
Selenia Anastasi, Università degli Studi di Bologna
Mario Aragon, Universidade de Santiago de Compostela
Sadiah Ashraf, Indiana University Bloomington
Shaina Ashraf, Philipps-University Marburg
Eliseo Bao, IRLab, CITIC, Universidade da Coruña
Valerio Basile, University of Turin
Ava Batchkala, Cohere
Gemma Bel-Enguix, Universidad Nacional Autónoma de México
Caroline Brun, Naver Labs Europe
Ana-Maria Bucur, Università della Svizzera italiana
Tommaso Caselli, Rijksuniversiteit Groningen
Ming-Bin Chen, University of Melbourne
Samuele D'Avenia, University of Turin
Greta Damo, Université Côte d'Azur
Mursal Dawodi, Technical University of Munich
Daryna Dementieva, Technical University of Munich
Eliana Di Palma, University of Turin
Nemanja Djuric, Aurora Innovation
Laura Dozal, University of Arizona
Lia Draetta, University of Turin
Srimonti Dutta, Women in AI (WAI) USA
Maria Espinosa Ruiz, UCM, CUNEF Universidad
Diandra Fabre, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
Marcos Fernandez-Pichel, University of Santiago de Compostela
Chiara Ferrando, University of Turin
Elisabetta Fersini, University of Milano-Bicocca
Jan Fillies, Freie Universität Berlin
Simona Frenda, Heriot-Watt University
Björn Gambäck, Norwegian University of Science and Technology
Fateme Ghavidel, University of Turin (UniTo)

Shlok Gilda, University of Florida
 Voula Giouli, Aristotle University of Thessaloniki / ILSP, ATHENA RC
 Janis Goldzycher, University of Zurich
 Salvatore Greco, King's College London
 Péter Hatvani, PPKE Doctoral School of Linguistics
 Mark Hill, King's College London
 Eduard Hovy, University of Melbourne
 Comfort Ilevbare, Afe Babalola University
 Muhammad Junaid Iqbal, Roma tor vergata
 Malvika Jadhav, University of Florida
 Farane Jalali Farahani, Institute for Artificial Intelligence, University of Stuttgart
 Sangwan Jeon, Université de Lille
 Gaurang Katyal, Roblox
 Abhishek Kaushik, Dkit
 Takyoun Kim, University of Illinois Urbana-Champaign
 Luise Koch, Technical University Munich
 Sandra Kübler, Indiana University
 Roy Ka-Wei Lee, Singapore University of Technology and Design
 Els Lefever, LT3, Ghent University
 Elisa Leonardelli, Foundation Bruno Kessler
 Chaya Liebeskind, Jerusalem College of Technology , Lev Academic Center
 Lucy Lin, Spotify
 Enrico Liscio, Delft University of Technology
 Mareike Lisker, Hochschule für Technik und Wirtschaft (HTW) Berlin
 Marina Litvak, Shamoon College of Engineering
 Soda Maren Lo, University of Turin
 Lawrence Low, National Taipei University of Technology
 Marco Madeddu, University of Turin
 Aaron Maladry, Ghent University
 Sayan Mandal, AMD
 Genoveffa Martone, Fondazione Bruno Kessler, Università Cattolica del Sacro Cuore
 Michele Mastromattei, Humanitas University
 Diana Maynard, University of Sheffield
 Stephen Meisenbacher, Technical University of Munich
 Helena Mihaljevic, Hochschule für Technik und Wirtschaft Berlin
 Manuel Montes, INAOE
 Tatsunori Mori, Yokohama National University
 Véronique Moriceau, IRIT, Université de Toulouse
 Arianna Muti, Bocconi University
 Ricardo Muñoz Sánchez, Språkbanken Text, Göteborgs Universitet
 Nazia Nafis, The University of Sheffield
 Isar Nejadgholi, National Research Council Canada
 Nicolás Benjamín Ocampo, Centrum Wiskunde & Informatica
 Mugdha Pandya, University of Edinburgh
 Yannick Parmentier, Université de Lorraine
 Lucia Passaro, University of Pisa
 Ricardo Penteado, University of São Paulo
 Anxo Perez, IRLab, Computer Science Dept., University of A Coruña
 Michal Ptaszynski, Kitami Institute of Technology
 Jessica Quaye, Harvard University
 Metztlí Ramírez,

Georg Rehm, DFKI
Hamidreza Saffari, Politecnico di Milano
Alexandra Schofield, Harvey Mudd College
Indira Sen, RWTH Aachen
Mattia Setzu, University of Pisa
Jeffrey Sorensen, Google Jigsaw
Panagiotis Soustas, University of Bristol
Vivian Stamou, Archimedes AI
Marco Antonio Stranisci, University of Turin
Afrin Sultana, University of Chittagong
Carolina Ternero, University of Washington
Jacob-Junqi Tian, Vector Institute
Zuoyu Tian, Macalester College
Manuel Tonneau, University of Oxford, World Bank
Sergio Ojeda Trueba,
Dimitrios Tsarapatsanis, University of York
Wondimagegnhue Tsegaye Tufa, Vrije Universiteit Amsterdam
Kanishk Verma, ADAPT SFI Centre, Dublin City University
Giulia Villani, Università della Svizzera italiana
Piek Vossen, Vrije Universiteit Amsterdam
Qing Xiao, Carnegie Mellon University
Guanqun Yang, Stevens Institute of Technology
Zachary Yang, McGill | Mila | Ubisoft
Angela Yuan, The University of Melbourne
Nicolas Zampieri, EVERDIAN
Aleš Završnik, Institute of criminology at the Faculty of Law Ljubljana
Fanqi Zeng, University of Oxford
Jiangrui Zheng, Stevens Institute of Technology
Yi Zheng, University of Edinburgh

Table of Contents

<i>Adaptive Causal Coordination Detection for Online Harm Campaigns: A Framework Study with Semi-Supervised Labeling</i>	
Yi Han, Wen Ding, Xuanjing Chen, Zhiguo Tao, Lifan Sun, Zichen Yuan and Yifei Huang	1
<i>RADAR: Retrieval-Augmented Detector with Adversarial Refinement for Adaptive LLM-Generated Fake News Detection</i>	
Song-Duo Ma, Yi-Hung Liu, Hsin-Yu Lin, Pin-Yu Chen, Hong-Yan Huang, Shau-Yung Hsu and Yun-Nung Chen	13
<i>No Better Than Random: Evaluating LIME-Guided Saliency Penalisation for Fine-Grained Sexism Classification</i>	
Olarenwaju Adisa and Chidinma Ndubuisi	27
<i>When Harm Is Polite: Coercion Markers and Low Toxicity in Conversion-Therapy Discourse</i>	
Adhiraj Chhoda	38
<i>Unseen Harm: Measuring Cross-Script Safety Inconsistency with Missed-in-Urdu Scores in LLM Hate Speech Detection.</i>	
Fawzia Kara-Isitt, Sonal Khosla and Stephen Swift	49
<i>RSD framework for the detection of covert online harms</i>	
Henna Paakki	63
<i>Beyond Translation: Cross-Linguistic Expression Strategies of Online Incivility on YouTube: A Comparison of English and Russian</i>	
Mursal Dawodi, Yingping Sun, Albina Fatykhova, Dominic Lammert and Juergen Pfeffer	75
<i>In-Browser Human-AI Co-Writing for Counterspeech: An Exploratory User Study of Agency, Perceived Response Quality, and Time</i>	
Tammy Schmidt and Luise Koch	85
<i>Context-Aware Evidence Minimization for Privacy-Conscious Harmful-Speech Dataset Release</i>	
Batin Örene	103
<i>Leveraging LLMs for Context-Aware Implicit Textual and Multimodal Hate Speech Detection</i>	
Joshua Brook and Ilia Markov	119
<i>Detecting Toxic Communication in Hungarian: Theory, Corpus, and Thread-Conditioned Classification</i>	
Péter Hatvani and Gábor Madarász	139
<i>When Harmful Intent Lies Between the Lines: Pragmatics-Based Automated Red Teaming for LLM Safety Evaluation</i>	
Shuying Ye, Carolina Ternero and Fei Xia	157
<i>Temporal and Geographic Patterns in Child Sexual Abuse Prosecutions: A Decade of Magistrates' Court Records from England and Wales</i>	
Tabassom Sedighi, Timothy McSweeney and Samantha Lundrigan	173
<i>Beyond a Global Norm: Personalizing Toxicity Sensitivity in Language Models Without Retraining</i>	
Rares Diaconescu, Iulia Slanina, Alina Florea, Andrei Trache, Miruna Coroi, Anne Arzberger, Jie Yang and Enrico Liscio	178

<i>Cross-Ideological Hallmarks of Extremist Rhetoric in Social Media</i>	
Hasan Omarzae and Björn Gambäck	192
<i>How User-AI Mistreatment Occurs and Matters in Conversational Systems?</i>	
Fanqi Zeng, Sadid Hasan and Chaocheng He	204
<i>Mind the Alignment Gap: Why General-Purpose Moderation Fails Children, and How a Child-Centric Taxonomy and Synthetic Data Close It</i>	
Jan Fikeis and Pavel Kordík	219
<i>Evaluating Automatically Inferred Linguistic Perspective as an Auxiliary Signal for Toxicity Detection</i>	
Meghana Gottapu	228
<i>Exploring LLM Performance for sexist hate speech detection on French social media</i>	
Shaden Shabayek, Camilla Penzo, Eunkyung Lee, Farah Benamara, Véronique Moriceau and Gohar Dashyan	234
<i>Who's Behind It? Annotating and Extracting Conspiratorial Actors from German Telegram Posts</i>	
Helena Mihaljevic, Jolanda Beer, Mareike Lisker and Katharina Soemer	249
<i>Beyond Slurs: A Multilingual Analysis of Misogyny Detection Across Dutch, English, and Spanish</i>	
Jirky Dumon, Aaron Maladry and Els Lefever	261
<i>Counterfactual Robustness and the Calibration Trap in Antisemitism Detection</i>	
Giacomo Case, Yasmin Chaqor and Leonardo Cartesegna	273
<i>Annotating Extremist Severity in Brazilian Portuguese: Human-LLM Divergence Across Labels, Evidence, and Interpretation</i>	
Ricardo Penteado and Barbara McGillivray	284
<i>How Reliable Are LLM Annotators for Online Manipulation? Evidence from Phishing and Abuse Corpora</i>	
Senaid Popovic, Damien Riquet, Maxime Meyer, Fabien Lauer and Yannick Parmentier	298
<i>Fifty Tokens of Signal: A Cross-Dataset Audit of Overlap, Shortcuts, and Label Divergence in Online-Abuse Benchmarks</i>	
Dhruv Patel	307
<i>Tailoring In-Context Learning Techniques for Definition-Based Hate Speech Detection in Large Language Models</i>	
Parham Bateni, Pradeep Murukannaiah and Urja Khurana	322
<i>Rule-Based Perturbation, Real Evaluation: How Far Do Fine-Tuned Encoders Go on Chinese Phonetically Cloaked Toxic Language?</i>	
Minggui Duan and Shijie Liu	338
<i>A Multilingual Dataset for Dehumanisation Detection in Russian and Ukrainian Wartime Communication</i>	
Elizaveta Chernenko and Balazs Vedres	346
<i>Introducing the Privacy-HSD Trade-off: Hate Speech Detection, but not at the Cost of Privacy</i>	
Stephen Meisenbacher, Vlad Gabuz, Chirill Donos, Maxim Dnestreanschi, Gabriel Creanga, Andreea-Elena Bodea, Thomas Lampert and Jana Diesner	361
<i>SWARM: A Multilingual Human-Annotated Dataset for Russian Propaganda Detection in Search Engine Results</i>	
Manuel Tonneau, Abhinav Dubey, Farhan Shaikh, Ilaria Vitulano, Martha Stolze, Hale Dedeoglu, Clara Riechert, Ella Kuka, Maryna Sydorova, Mykola Makhortykh and Elizaveta Kuznetsova	374

<i>Who Let the Dogs Out? Auditing Dogwhistle Coverage in Hate Speech Benchmarks Across Target Groups</i>	
Reva Hirave and Ryan Oet	392
<i>GENDIS: Gender-Related Discourse and Stance in Multilingual YouTube Comments from India, Kenya, and Nigeria</i>	
Luise Koch, Manuel Tonneau, Niyati Malhotra, Sunny Rai, Ana Maria Munoz Boudet, Victor Orozco-Olvera, Samuel Fraiberger and Sharath Chandra Guntuku	411
<i>MMJail: Compounded Multilingual and Multi-Turn Vulnerabilities in LLM Bias Resistance</i>	
Anisha Rao, Takyoung Kim and Dilek Hakkani-Tur	434
<i>Toxicity to Harm: Dataset Archaeology of a Decade of Online Abuse Research</i>	
Srimonti Dutta and Akshata Moharir	450
<i>Does LLM Instability Reflect Annotator Disagreement? A Study on Hate Speech Detection</i>	
Sneha Prashanth, Aradhana Mahati, Maksym Ziemlewski, Nicolas Loaiza and Urja Khurana	466
<i>You Never Can Tell: A Survey of Preferences on LLM-Generated Counterspeech in an Austrian Context</i>	
Pia Pachinger and Arianna Muti	477
<i>Propaganda Forensics: Recovering the Generation Pipeline of an AI-Driven Influence Campaign</i>	
Benjamin Icard, Elouan Vuichard, Louis Lefebvre, Lila Sainero, Thomas Girault, Alice Breton, Tanguy Launay, Gauvain Bourgne, Morgane Casanova, Guillaume Gadek, Victor Klötzer, Michel Le Nouy, Guillaume Gravier, Jean-Gabriel Ganascia and Paul Égré	490
<i>HurtNet: a Multilingual Dictionary of Hurtful Words</i>	
Maria Alexandra Roussopoulou, Eliana Di Palma, Andrea Di Fabio, Adel Mahmoud Wizani, Vivian Stamou, Anaïs Ollagnier, Petya Osenova, Valerio Basile and Stella Markantonatou	507

Program

Thursday, October 29, 2026

09:00 - 09:20 *Opening Remarks*

09:20 - 10:05 *Invited Talk 1 - Sara Tonelli*

10:05 - 10:30 *Lightning Talks for Remote Attendants*

Down the Rabbit Hole: Red Teaming Chatbot Robustness Against Conspiracy Theories

Richard Sieg

No Better Than Random: Evaluating LIME-Guided Saliency Penalisation for Fine-Grained Sexism Classification

Olarenwaju Adisa and Chidinma Ndubuisi

Context-Aware Evidence Minimization for Privacy-Conscious Harmful-Speech Dataset Release

Batin Örene

Exploring LLM Performance for sexist hate speech detection on French social media

Shaden Shabayek, Camilla Penzo, Eunkyung Lee, Farah Benamara, Véronique Moriceau and Gohar Dashyan

Fifty Tokens of Signal: A Cross-Dataset Audit of Overlap, Shortcuts, and Label Divergence in Online-Abuse Benchmarks

Dhruv Patel

Tailoring In-Context Learning Techniques for Definition-Based Hate Speech Detection in Large Language Models

Parham Bateni, Pradeep Murukannaiah and Urja Khurana

Does LLM Instability Reflect Annotator Disagreement? A Study on Hate Speech Detection

Sneha Prashanth, Aradhana Mahati, Maksym Ziemlewski, Nicolas Loaiza and Urja Khurana

10:30 - 11:00 *Coffee Break*

11:00 - 11:45 *Invited Talk 2 - Su Lin Blodgett*

11:45 - 12:30 *Invited Talk 3 - Zeerak Talat*

Thursday, October 29, 2026 (continued)

12:30 - 13:30 *Lunch Break*

13:30 - 13:45 *Best paper Award*

Who Let the Dogs Out? Auditing Dogwhistle Coverage in Hate Speech Benchmarks Across Target Groups

Reva Hirave and Ryan Oet

13:45 - 14:15 *Lightning Talks for Remote Attendants*

Adaptive Causal Coordination Detection for Online Harm Campaigns: A Framework Study with Semi-Supervised Labeling

Yi Han, Wen Ding, Xuanjing Chen, Zhiguo Tao, Lifan Sun, Zichen Yuan and Yifei Huang

When Harm Is Polite: Coercion Markers and Low Toxicity in Conversion-Therapy Discourse

Adhiraj Chhoda

When Harmful Intent Lies Between the Lines: Pragmatics-Based Automated Red Teaming for LLM Safety Evaluation

Shuying Ye, Carolina Ternero and Fei Xia

Evaluating Automatically Inferred Linguistic Perspective as an Auxiliary Signal for Toxicity Detection

Meghana Gottapu

Annotating Extremist Severity in Brazilian Portuguese: Human-LLM Divergence Across Labels, Evidence, and Interpretation

Ricardo Penteado and Barbara McGillivray

MMJail: Compounded Multilingual and Multi-Turn Vulnerabilities in LLM Bias Resistance

Anisha Rao, Takyoung Kim and Dilek Hakkani-Tur

Toxicity to Harm: Dataset Archaeology of a Decade of Online Abuse Research

Srimonti Dutta and Akshata Moharir

14:15 - 15:30 *In-Person Poster Session*

Thursday, October 29, 2026 (continued)

RADAR: Retrieval-Augmented Detector with Adversarial Refinement for Adaptive LLM-Generated Fake News Detection

Song-Duo Ma, Yi-Hung Liu, Hsin-Yu Lin, Pin-Yu Chen, Hong-Yan Huang, Shau-Yung Hsu and Yun-Nung Chen

Unseen Harm: Measuring Cross-Script Safety Inconsistency with Missed-in-UrduScores in LLM Hate Speech Detection.

Fawzia Kara-Isitt, Sonal Khosla and Stephen Swift

RSD framework for the detection of covert online harms

Henna Paakki

Beyond Translation: Cross-Linguistic Expression Strategies of Online Incivility on YouTube: A Comparison of English and Russian

Mursal Dawodi, Yingping Sun, Albina Fatykhova, Dominic Lammert and Juer-gen Pfeffer

In-Browser Human-AI Co-Writing for Counterspeech: An Exploratory User Study of Agency, Perceived Response Quality, and Time

Tammy Schmidt and Luise Koch

Leveraging LLMs for Context-Aware Implicit Textual and Multimodal Hate Speech Detection

Joshua Brook and Ilia Markov

Detecting Toxic Communication in Hungarian: Theory, Corpus, and Thread-Conditioned Classification

Péter Hatvani and Gábor Madarász

Temporal and Geographic Patterns in Child Sexual Abuse Prosecutions: A Decade of Magistrates' Court Records from England and Wales

Tabassom Sedighi, Timothy McSweeney and Samantha Lundrigan

Beyond a Global Norm: Personalizing Toxicity Sensitivity in Language Models Without Retraining

Rares Diaconescu, Iulia Slanina, Alina Florea, Andrei Trache, Miruna Coroi, Anne Arzberger, Jie Yang and Enrico Liscio

Cross-Ideological Hallmarks of Extremist Rhetoric in Social Media

Hasan Omarzae and Björn Gambäck

How User-AI Mistreatment Occurs and Matters in Conversational Systems?

Fanqi Zeng, Sadid Hasan and Chaocheng He

Thursday, October 29, 2026 (continued)

Mind the Alignment Gap: Why General-Purpose Moderation Fails Children, and How a Child-Centric Taxonomy and Synthetic Data Close It

Jan Fikeis and Pavel Kordík

Who's Behind It? Annotating and Extracting Conspiratorial Actors from German Telegram Posts

Helena Mihaljevic, Jolanda Beer, Mareike Lisker and Katharina Soemer

Beyond Slurs: A Multilingual Analysis of Misogyny Detection Across Dutch, English, and Spanish

Jirky Dumon, Aaron Maladry and Els Lefever

Counterfactual Robustness and the Calibration Trap in Antisemitism Detection

Giacomo Case, Yasmin Chaqor and Leonardo Cartesegna

How Reliable Are LLM Annotators for Online Manipulation? Evidence from Phishing and Abuse Corpora

Senaid Popovic, Damien Riquet, Maxime Meyer, Fabien Lauer and Yannick Parmentier

Rule-Based Perturbation, Real Evaluation: How Far Do Fine-Tuned Encoders Go on Chinese Phonetically Cloaked Toxic Language?

Mingguì Duan and Shijie Liu

A Multilingual Dataset for Dehumanisation Detection in Russian and Ukrainian Wartime Communication

Elizaveta Chernenko and Balazs Vedres

Minion: An LLM-Powered Browser Extension to Support Users in Negotiating Harmful Value Conflicts with AI Companions

Qing Xiao, Xianzhe Fan, Xuhui Zhou, Yuran Su, Zhicong Lu, Maarten Sap and Hong Shen

Introducing the Privacy-HSD Trade-off: Hate Speech Detection, but not at the Cost of Privacy

Stephen Meisenbacher, Vlad Gabuz, Chirill Donos, Maxim Dnestreanschi, Gabriel Creanga, Andreea-Elena Bodea, Thomas Lampert and Jana Diesner

SWARM: A Multilingual Human-Annotated Dataset for Russian Propaganda Detection in Search Engine Results

Manuel Tonneau, Abhinav Dubey, Farhan Shaikh, Ilaria Vitulano, Martha Stolze, Hale Dedeoglu, Clara Riechert, Ella Kuka, Maryna Sydorova, Mykola Makhortkh and Elizaveta Kuznetsova

GENDIS: Gender-Related Discourse and Stance in Multilingual YouTube Comments from India, Kenya, and Nigeria

Luise Koch, Manuel Tonneau, Niyati Malhotra, Sunny Rai, Ana Maria Munoz Boudet, Victor Orozco-Olvera, Samuel Fraiberger and Sharath Chandra Guntuku

Thursday, October 29, 2026 (continued)

You Never Can Tell: A Survey of Preferences on LLM-Generated Counterspeech in an Austrian Context

Pia Pachinger and Arianna Muti

Disagreeing Rationales: Rethinking Classification and Explainability Evaluation in Hate Speech Detection

Benedetta Muscato, Beiduo Chen, Gizem Gezici, Siyao Peng, Barbara Plank and Fosca Giannotti

Bye Bye Perspective API: Lessons for Measurement Infrastructure in NLP, CSS and LLM Evaluation

David Hartmann, Manuel Tonneau, Angelie Kraft, LK Seiling, Dimitri Staufer, Pieter Delobelle, Jan Fillies, Anna Ricarda Luther, Jan Batzner and Mareike Lisker

Propaganda Forensics: Recovering the Generation Pipeline of an AI-Driven Influence Campaign

Benjamin Icard, Elouan Vuichard, Louis Lefebvre, Lila Sainero, Thomas Girault, Alice Breton, Tanguy Launay, Gauvain Bourgne, Morgane Casanova, Guillaume Gadek, Victor Klötzer, Michel Le Nouy, Guillaume Gravier, Jean-Gabriel Ganasia and Paul Égré

You Frame It: How Conceptual Representations Shape LLM Detection and Reasoning about Antisemitism

Katharina Soemer and Helena Mihaljević

HurtNet: a Multilingual Dictionary of Hurtful Words

Maria Alexandra Roussopoulou, Eliana Di Palma, Andrea Di Fabio, Adel Mahmoud Wizani, Vivian Stamou, Anaïs Ollagnier, Petya Osenova, Valerio Basile and Stella Markantonatou

From Specialization to Generalization: Instruction-tuned LLMs for Robust Harmful Content Mitigation

Lukas Edman, Daryna Dementieva and Alexander Fraser

15:30 - 16:00 *Coffee Break*

16:00 - 16:45 *Invited Talk 4 - Sarah T. Roberts*

16:45 - 17:25 *Panel Discussion*

17:25 - 17:30 *Closing Remarks*