PrivateNLP 2026

The Seventh Workshop on Privacy in Natural Language Processing

Proceedings of the Workshop

July 3, 2026

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 979-8-89176-397-5

Introduction

Welcome to the Seventh Workshop on Privacy in Natural Language Processing. Co-located with ACL 2026 in San Diego (CA), USA, the workshop is scheduled for Friday, July 3rd, 2026. To facilitate the participation of the global NLP community, we continue running the workshop in a hybrid format.

Privacy-preserving language data processing has become essential in the age of Large Language Models (LLMs) where access to vast amounts of data can provide gains over tuned algorithms. A large proportion of user-contributed data comes from natural language, e.g., text transcriptions from voice assistants. It is therefore important to curate NLP datasets while preserving the privacy of the users whose data is collected, and train ML models that only retain non-identifying user data. The workshop brings together practitioners and researchers from academia and industry to discuss the challenges and approaches to designing, building, verifying, and testing privacy preserving systems in the context of Natural Language Processing.

Our half-day agenda features a keynote speech and hybrid talk sessions both for long and short papers. This year we received 22 direct submissions and three 'fast-track' submissions. We accepted 14 submissions after a thorough peer-review, corresponding to 64% acceptance rate. Four accepted submissions decided for a non-archival track and thus are not included in these proceedings. We also desk-rejected one highly suspicious LLM-generated paper from a highly suspicious account that had already been deactivated by OpenReview for posting fake papers to numerous other venues in 2026; we hope our workshop will keep attracting only good-faith researchers interested in scientific discourse and true research exchange in its future editions.

We would like to deeply thank to all the authors, committee members, keynote speaker, and participants to help us make this research community grow both in quantity and quality.

Workshop Chairs

Organizing Committee

Program Chairs

Ivan Habernal, Ruhr-University Bochum, Germany
Sepideh Ghanavati, University of Maine, United States
Sara Haghighi, University of Maine, United States
Krithika Ramesh, Johns Hopkins University, United States
Timour Igamberdiev, University of Vienna, Austria
Shomir Wilson, Pennsylvania State University, United States

Program Committee

Reviewers

Stefan Arnold, Friedrich-Alexander-Universität Erlangen-Nürnberg
Andrea Atzeni, Politecnico di Torino
Travis Breaux, Carnegie Mellon University
Christos Dimitrakakis, Université de Neuchâtel
Mark Dras, Macquarie University
James Flemings, University of Southern California
Pierre Lison, Norwegian Computing Center
Christina Lohr, Universität Leipzig
Eugenio Martínez-Cámara, Universidad de Jaén
Stephen Meisenbacher, Technical University of Munich
Isar Nejadgholi, National Research Council Canada and University of Ottawa
Sebastian Ochs, Technische Universität Darmstadt
Ildikó Pilán, Norwegian Computing Center
Lizhen Qu, Monash University
Peter Story, Clark University
Juraj Vladika, Technische Universität München
Ruyu Zhou, University of Notre Dame

Table of Contents

From Conventional Web Privacy to Agentic Disclosure: How Tool Schemas May Invite LLM Oversharing
Shahriar Shayesteh and Shomir Wilson 1

The Challenge of Identifying the Origin of Black-Box Large Language Models
Ziqing Yang, Yixin Wu, Yun Shen, Wei Dai, Michael Backes and Yang Zhang 7

SecureLLM: Using Inference-time Compositionality to Build Secure Language Models
Abdulrahman Alabdulkareem, Christian Michael Arnold, Yerim Lee, Pieter M Feenstra, Conner Arnold, Boris Katz, Andrei Barbu and Brian Cheung 26

STAMP-R: Stylometric Text Anonymization with Memory-guided Policy Rewriting
Zhan Shi, Yefeng Yuan, Liang Cheng and Yuhong Liu 53

Loss Masking Under the Hood: Backdoor Concealment and Private Data Memorization in LLMs
Tagore Rao Kosireddy and Evan Lucas 69

Prompt Stylometry for On-Device Affect-Adaptive AI: A Feasibility Study in Linguistic Signal Detection and Response Steering
Debmalya Pal 80

Differentially-Private Text Rewriting reshapes Linguistic Style
Stefan Arnold 96

Linguistic Identity Leakage: When Language Reveals Identity in Anonymized Text
Wajdi Zaghouani 107

A Systematic Exploration of Text Decomposition and Budget Distribution in Differentially Private Text Obfuscation
Stephen Meisenbacher, Angelo Kleinert and Florian Matthes 118

Safer Reasoning Traces: Measuring and Mitigating Chain-of-Thought Leakage in LLMs
Patrick Ahrend, Tobias Eder, Xiyang Yang, Zhiyi Pan and Georg Groh 140