

PrivateNLP 2024

**The Fifth Workshop on Privacy in Natural Language  
Processing**

**Proceedings of the Workshop**

August 15, 2024

©2024 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](mailto:acl@aclweb.org)

ISBN 979-8-89176-139-1

## Introduction

Welcome to the Fifth Workshop on Privacy in Natural Language Processing. Co-located with ACL 2024 in Bangkok, Thailand, the workshop is scheduled for August 15, 2024. 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 agenda features a keynote speech, hybrid talk sessions both for long and short papers, and a poster session. This year we received 29 submissions. We accepted 23 submissions after a thorough peer-review. Five accepted submissions preferred the non-archival option and thus are not included in this proceedings. Moreover, our poster session features additional four ACL-Findings papers.

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

Abhilasha Ravichander, Allen Institute for AI, United States

Vijayanta Jain, University of Maine, United States

Patricia Thaine, Private AI, Canada

Timour Igamberdiev, Technical University of Darmstadt, Germany

Niloofer Miresghallah, University of Washington, United States

Oluwaseyi Feyisetan, Amazon, United States

# Program Committee

## Program Committee

Andrea Atzeni, Polytechnic Institute of Turin  
Asma Aloufi, Taif University  
Eleftheria Makri, Leiden University  
Erion Cano, Universität Paderborn  
Eugenio Martínez-Cámara, Universidad de Jaén  
Gergely Acs, Technical University of Budapest  
Isar Nejadgholi, National Research Council Canada  
Jaydeep Borkar, Northeastern University  
Kambiz Ghazinour, SUNY Canton  
Lizhen Qu, Monash University  
Mattia Salnitri, Polytechnic Institute of Milan  
Mousumi Akter, Technische Universität Dortmund  
Natasha Fernandes, Macquarie University  
Pengwei Li, Meta  
Peter Story, Clark University  
Pierre Lison, Norwegian Computing Center  
Rocky Slavin, University of Texas at San Antonio  
Ruyu Zhou, University of Notre Dame  
Sai Peddinti, Google  
Sebastian Ochs, Technische Universität Darmstadt  
Shomir Wilson, Pennsylvania State University  
Timour Igamberdiev, Technische Universität Darmstadt  
Travis Breaux, Carnegie Mellon University

## Table of Contents

<i>Noisy Neighbors: Efficient membership inference attacks against LLMs</i> Filippo Galli, Luca Melis and Tommaso Cucinotta .....	1
<i>Don't forget private retrieval: distributed private similarity search for large language models</i> Guy Zyskind, Tobin South and Alex 'Sandy' Pentland .....	7
<i>Characterizing Stereotypical Bias from Privacy-preserving Pre-Training</i> Stefan Arnold, Rene Gröbner and Annika Schreiner .....	20
<i>Protecting Privacy in Classifiers by Token Manipulation</i> Re'em Harel, Yair Elboher and Yuval Pinter .....	29
<i>A Collocation-based Method for Addressing Challenges in Word-level Metric Differential Privacy</i> Stephen Meisenbacher, Maulik Chevli and Florian Matthes .....	39
<i>Preset-Voice Matching for Privacy Regulated Speech-to-Speech Translation Systems</i> Daniel Platnick, Bishoy Abdelnour, Eamon Earl, Rahul Kumar, Zahra Rezaei, Thomas Tsangaris and Faraj Lagum .....	52
<i>PII-Compass: Guiding LLM training data extraction prompts towards the target PII via grounding</i> Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang and Xuebing Zhou .....	63
<i>Unlocking the Potential of Large Language Models for Clinical Text Anonymization: A Comparative Study</i> David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André V. Carreiro and Vitor Rolla .....	74
<i>Anonymization Through Substitution: Words vs Sentences</i> Vasco Alves, Vitor Rolla, João Alveira, David Pissarra, Duarte Pereira, Isabel Curioso, André V. Carreiro and Henrique Lopes Cardoso .....	85
<i>PocketLLM: Enabling On-Device Fine-Tuning for Personalized LLMs</i> Dan Peng, Zhihui Fu and Jun Wang .....	91
<i>Smart Lexical Search for Label Flipping Adversarial Attack</i> Alberto José Gutiérrez-Megías, Salud María Jiménez-Zafra, L. Alfonso Ureña and Eugenio Martínez- Cámara .....	97
<i>Can LLMs get help from other LLMs without revealing private information?</i> Florian Hartmann, Duc-Hieu Tran, Peter Kairouz, Victor Cărbune and Blaise Aguera Y Arcas	107
<i>Cloaked Classifiers: Pseudonymization Strategies on Sensitive Classification Tasks</i> Arij Riabi, Menel Mahamdi, Virginie Mouilleron and Djamé Seddah .....	123
<i>Improving Authorship Privacy: Adaptive Obfuscation with the Dynamic Selection of Techniques</i> Hemanth Kandula, Damianos Karakos, Haoling Qiu and Brian Ulicny .....	137
<i>Deconstructing Classifiers: Towards A Data Reconstruction Attack Against Text Classification Models</i> Adel Elmahdy and Ahmed Salem .....	143
<i>PrivaT5: A Generative Language Model for Privacy Policies</i> Mohammad Al Zoubi, Santosh T.y.s.s, Edgar Ricardo Chavez Rosas and Matthias Grabmair .	159

<i>Reinforcement Learning-Driven LLM Agent for Automated Attacks on LLMs</i>	
Xiangwen Wang, Jie Peng, Kaidi Xu, Huaxiu Yao and Tianlong Chen .....	170
<i>A Privacy-preserving Approach to Ingest Knowledge from Proprietary Web-based to Locally Run Models for Medical Progress Note Generation</i>	
Sarvesh Soni and Dina Demner-Fushman .....	178