

Fifth Workshop on Privacy in Natural Language Processing (PrivateNLP 2024)

Bangkok, Thailand
15 August 2024

ISBN: 979-8-3313-0173-6

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2024) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

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 Vítor Rolla	74
<i>Anonymization Through Substitution: Words vs Sentences</i>	
Vasco Alves, Vítor 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 Moulleron 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