

GEM 2023

**The 2023 Conference on Empirical Methods in Natural  
Language Processing**

**Proceedings of the Third Workshop on Natural Language  
Generation, Evaluation, and Metrics (GEM)**

December 6, 2023

©2023 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
209 N. Eighth Street  
Stroudsburg, PA 18360  
USA  
Tel: +1-570-476-8006  
Fax: +1-570-476-0860  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN 979-8-89176-049-3

## Introduction

We are excited to welcome you to GEM 2023, the 3rd Workshop on Generation Evaluation and Metrics. This year the workshop is being held in Singapore, on December 6, 2023, just before EMNLP 2023 that will take place December 8-10.

The GEM workshop aims to advance the field of evaluation, a field that has increased in importance as language generating models become more ubiquitous in everyday life. As models increase in variety of skills they perform, it is crucial to advance evaluation techniques at the same time. The workshop features a selection of papers on improving generation fluency, coherence, and faithfulness. It covers topics in automatic evaluation using learned and designed metrics. And it includes many works on assessing generated model outputs with humans.

We received 67 submissions this year, split between our main track, extended abstracts, and the inaugural industry track. This year, for the first time, we had six area chairs who were responsible to coordinate the review process for main track papers. We accepted 29 main papers, 9 extended abstracts, and 3 industry papers, for an overall acceptance rate of 62%. We additionally invited 34 papers accepted to Findings of EMNLP 2023 to present at the workshop, for a total of 75 presented works.

The Organizing Committee

# Organizing Committee

## Organizing Committee

Khyathi Raghavi Chandu, Allen Institute of AI  
Elizabeth Clark, Google Deepmind  
Kaustubh Dhole, Emory University  
Sebastian Gehrmann, Bloomberg LP  
João Sedoc, New York University  
Alex Wang, Cohere

## Industry Track Organizers

Enrico Santus, Bloomberg LP  
Hooman Sedghamiz, Bayer AG

# Program Committee

## Chairs

Khyathi Raghavi Chandu, Allen Institute of AI  
Elizabeth Clark, Google Research  
Kaustubh Dhole, Emory University  
Sebastian Gehrmann, Bloomberg LP  
Enrico Santus, Bloomberg  
Hooman Sedghamiz, Bayer AG  
João Sedoc, New York University  
Alex Wang, Cohere

## Program Committee

Anand A. Rajasekar, Flipkart Internet Private Limited  
Samuel Ackerman, IBM Research  
Tosin Adewumi, Luleå University of Technology  
Ameeta Agrawal, Portland State University  
Nestor Alvaro, Independent  
Pawan Sasanka Ammanamanchi, IIIT Hyderabad  
Anuoluwapo Aremu, Masakhane  
Shima Asaadi, Fraunhofer IIS  
Simone Balloccu, University of Aberdeen  
Samuel Cahyawijaya, HKUST  
Eduardo Calò, Utrecht University  
Ronald Cardenas, University of Edinburgh  
Boaz Carmeli, IBM Research - Haifa  
Silvia Casola, Fondazione Bruno Kessler, University of Padua  
Miruna Clinciu, Edinburgh Centre for Robotics  
Jordan Clive, Imperial College London  
Kordula De Kuthy, Universität Tübingen  
Ondrej Dusek, Charles University  
Chris Chinenye Emezue, Technical University of Munich  
Eitan Farchi, IBM research  
Lucie Flek, CAISA Lab, University of Bonn  
Subhasish Ghosh, TCS Research  
John Glover, 3M  
Ankita Gupta, University of Massachusetts Amherst  
Dilek Hakkani-Tur, Amazon Alexa AI  
Behnam Hedayatnia, Amazon  
Kaili Huang, Microsoft  
Rudali Huidrom, ADAPT Research Centre, Dublin City University  
Nikolai Ilinykh, University of Gothenburg  
Vasudevan Jagannathan, 3M  
Yangfeng Ji, University of Virginia  
Di Jin, Amazon  
Mayank Jobanputra, Saarland University  
Shailza Jolly, Amazon Alexa AI  
Mihir Kale, Google

Moussa Kamal Eddine, École polytechnique  
Marzena Karpinska, University of Massachusetts Amherst  
Noriaki Kawamae, NTT Comware  
Sergey Kovalchuk, Huawei  
Kalpesh Krishna, Google  
Saurabh Kulshreshtha, University of Massachusetts Lowell  
Dhruv Kumar, Grammarly  
Harsh Lara, Google Research  
Alberto Lavelli, FBK  
Jing Yang Lee, Nanyang Technological University  
Hwanhee Lee, Chung-Ang University  
Yinghui Li, Tsinghua University  
Terry Lima Ruas, University of Gottingen  
Yixin Liu, Yale University  
Yinhong Liu, University of Cambridge  
Michela Lorandi, Dublin City University  
Ehsan Lotfi, University of Antwerp, CLiPS  
Mounica Maddela, Georgia Institute of Technology  
Khyati Mahajan, University of North Carolina at Charlotte  
Saad Mahamood, trivago N.V  
Abinaya Mahendiran, Mphasis NEXT Labs  
Pedro Henrique Martins, Instituto de Telecomunicações, Instituto Superior Técnico  
Joshua Maynez, Google  
Sebastien Montella, Huawei Ltd.  
Seyed Mahed Mousavi, University of Trento  
Tadashi Nomoto, National Institute of Japanese Literature  
Alexandros Papangelis, Amazon Alexa AI  
Soham Parikh, ServiceNow Inc  
Cheoneum Park, Hyundai Motor Group  
Eunil Park, Sungkyunkwan University  
Tatiana Passali, Aristotle University of Thessaloniki  
Dina Pisarevskaya, PhD student at QMUL  
Maja Popović, ADAPT, Dublin City University  
Mahima Pushkarna, Google  
Vipul Raheja, Grammarly  
Vikas Raunak, Microsoft  
Marek Rei, Imperial College London  
Ehud Reiter, University of Aberdeen  
Leonardo F. R. Ribeiro, Amazon Alexa AI  
Giuseppe Riccardi, University of Trento  
Gabriel Roccabruna, University of Trento  
Juan Diego Rodriguez, The University of Texas at Austin  
Sashank Santhanam, University of North Carolina at Charlotte/ Apple  
Thomas Schaaf, 3M | M\*Modal  
Rifat Shahriyar, Bangladesh University of Engineering and Technology  
Tatiana Shavrina, AIRI  
Tianhao Shen, Tianjin University  
Anastasia Shimorina, Orange  
Anna Shvets, FabLab by Inetum  
Arabella Sinclair, University of Aberdeen  
Somayajulu Sripada, Arria NLG Plc and University of Aberdeen

Hendrik Strobelt, MIT IBM Watson AI Lab  
Barkavi Sundararajan, University of Aberdeen  
Bowen Tan, Carnegie Mellon University  
Craig Thomson, University of Aberdeen  
Ashish Upadhyay, Robert Gordon University  
David Vilar, Google  
Y e n - H s i a n g Wang, National Chung Hsing University  
John Wieting, University of Illinois; TTI-Chicago; CMU; Google  
Xinnuo Xu, University of Edinburgh  
Bing Yan, New York University  
Guanqun Yang, Stevens Institute of Technology  
Akhila Yerukola, Carnegie Mellon University  
Naoki Yoshinaga, Institute of Industrial Science, The University of Tokyo  
Alessandra Zarcone, Technische Hochschule Augsburg  
Yian Zhang, Amazon  
Justin Zhao, Google  
Yongxin Zhou, Université Grenoble Alpes, LIG  
Jiawei Zhou, Harvard University  
Qi Zhu, Tsinghua University

## Table of Contents

<i>Contextualizing the Limits of Model Evaluation Dataset Curation on Semantic Similarity Classification Tasks</i>	
Daniel Theron .....	1
<i>Dialogue Quality and Emotion Annotations for Customer Support Conversations</i>	
John Mendonca, Patrícia Pereira, Miguel Menezes, Vera Cabarrão, Ana C Farinha, Helena Moniz, Alon Lavie and Isabel Trancoso .....	9
<i>Formalizing content creation and evaluation methods for AI-generated social media content</i>	
Christian Jensen and Axel Højmark .....	22
<i>Automatic Evaluation of Generative Models with Instruction Tuning</i>	
Shuhaib Mehri and Vered Shwartz .....	42
<i>Effective Proxy for Human Labeling: Ensemble Disagreement Scores in Large Language Models for Industrial NLP</i>	
Wei Du, Laksh Advani, Yashmeet Gambhir, Daniel Perry, Prashant Shiralkar, Zhengzheng Xing and Aaron Colak .....	53
<i>Automatic Reflection Generation for Peer-to-Peer Counseling</i>	
Emma O’neil, João Sedoc, Diyi Yang, Haiyi Zhu and Lyle Ungar .....	62
<i>One-Shot and Few-Shot Exemplification Modeling</i>	
John Harvill, Hee Suk Yoon, Eunseop Yoon, Mark H a s e g a w a - J o h n s o n and Chang Yoo	76
<i>Leveraging Large Language Models for Enhanced Product Descriptions in eCommerce</i>	
Jianghong Zhou, Bo Liu, Jhalak Acharya, Yao Hong, K u a n g - C h i h Lee and Musen Wen .	88
<i>QAMPARI: A Benchmark for Open-domain Questions with Many Answers</i>	
Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig and Jonathan Berant	97
<i>Unveiling Safety Vulnerabilities of Large Language Models</i>	
George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Fandina, Ateret Anaby Tavor, Orna Raz and Eitan Farchi .....	111
<i>Adapting Pre-trained Generative Models for Extractive Question Answering</i>	
Prabir Mallick, Tapas Nayak and Indrajit Bhattacharya .....	128
<i>Predicting Question-Answering Performance of Large Language Models through Semantic Consistency</i>	
Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi and Ateret Anaby Tavor .....	138
<i>Towards Effective Long-Form QA with Evidence Augmentation</i>	
Mengxia Yu, Sara Rosenthal, Mihaela Bornea and Avi Sil .....	155
<i>Harnessing the Plug-and-Play Controller by Prompting</i>	
Hao Wang and Lei Sha .....	165
<i>Context and Literacy Aware Learnable Metric for Text Simplification</i>	
Jeongwon Kwak, Hyeryun Park, Kyungmo Kim and Jinwook Choi .....	175
<i>Synthetic Dialogue Dataset Generation using LLM Agents</i>	
Yelaman Abdullin, Diego Molla, Bahadorreza Ofoghi, John Yearwood and Qingyang Li . . . .	181

<i>An Empirical Bayes Framework for Open-Domain Dialogue Generation</i>	
Jing Yang Lee, Kong Aik Lee and Woon Seng Gan .....	192
<i>Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models</i>	
Joseph Marvin Imperial and Harish Tayyar Madabushi .....	205
<i>ChatGPT as a Java Decompiler</i>	
Bradley Mcdanel and Zhanhao Liu .....	224
<i>Multi-domain Summarization from Leaderboards to Practice: Re-examining Automatic and Human Evaluation</i>	
David Demeter, Oshin Agarwal, Simon Ben Igeri, Marko Sterbentz, Neil Molino, John Conroy and Ani Nenkova .....	233
<i>Targeted Image Data Augmentation Increases Basic Skills Captioning Robustness</i>	
Valentin Barriere, Felipe Del Rio, Andres Carvallo, Carlos Aspillaga, Eugenio Herrera-Berg and Cristian Buc .....	243
<i>Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses</i>	
Xenia Ohmer, Elia Bruni and Dieuwke Hupkes .....	258
<i>Text Encoders Lack Knowledge: Leveraging Generative LLMs for Domain-Specific Semantic Textual Similarity</i>	
Joseph Gatto, Omar Sharif, Parker Seegmiller, Philip Bohlman and Sarah Preum .....	277
<i>To Burst or Not to Burst: Generating and Quantifying Improbable Text</i>	
Kuleen Sasse, Efsun Sarioglu Kayi, Samuel Barham and Edward Staley .....	289
<i>Are Large Language Models Reliable Judges? A Study on the Factuality Evaluation Capabilities of LLMs</i>	
Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen and Shashi Bhushan Tn .....	310
<i>RankAug: Augmented data ranking for text classification</i>	
Tiasa Roy and Priyam Basu .....	317
<i>Separating the Wheat from the Chaff with BREAD: An open-source benchmark and metrics to detect redundancy in text</i>	
Isaac Caswell, Lisa Wang and Isabel Papadimitriou .....	324
<i>Elo Uncovered: Robustness and Best Practices in Language Model Evaluation</i>	
Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker and Marzieh Fadaee .....	339
<i>PersonalityChat: Conversation Distillation for Personalized Dialog Modeling with Facts and Traits</i>	
Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann and Walter Daelemans .....	353
<i>How well ChatGPT understand Malaysian English? An Evaluation on Named Entity Recognition and Relation Extraction</i>	
Mohanraj Chanthran, Lay-Ki Soon, Ong Huey Fang and Bhawani Selvaretnam .....	372
<i>Post Turing: Mapping the landscape of LLM Evaluation</i>	
Alexey Tikhonov and Ivan Yamshchikov .....	398
<i>A Simple yet Efficient Ensemble Approach for AI-generated Text Detection</i>	
Harika Abburi, Kalyani Roy, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen and Sanmitra Bhattacharya .....	413