

GEM 2022

**2nd Workshop on Natural Language Generation, Evaluation,  
and Metrics**

**Proceedings of the Workshop**

December 7, 2022

©2022 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 978-1-959429-12-8

## **Introduction**

Welcome to the second workshop on Natural Language Generation, Evaluation, and Metrics (GEM), to be held on December 7, 2022 as part of EMNLP in Abu Dhabi. The workshop aims to bring together researchers interested in model audits, new evaluation approaches and meta evaluations. The workshop is privileged to present several invited talks this year and the results of the shared task on generation with limited resources.

We are grateful to the program committee for their careful and thoughtful reviews of the papers submitted this year. Likewise, we are thankful to the shared task organizers for their hard work in preparing the shared tasks. We are looking forward to a workshop covering a wide range of topics, and we hope for lively discussions.

GEM Workshop Organizers.

# Organizing Committee

## Organizers

Antoine Bosselut, EPFL  
Khyathi Chandu, Meta AI  
Kaustubh Dhole, Emory University  
Varun Gangal, Carnegie Mellon University  
Sebastian Gehrmann, Google Research  
Yacine Jernite, Hugging Face  
Jekaterina Novikova, Winterlight Labs  
Laura Perez-Beltrachini, University of Edinburgh

# Program Committee

## Reviewers

Tosin Adewumi, Ameeta Agrawal, Fatih Amasyali, David Aparicio, Samuel Arcadinho, Shima Asaadi

Simone Balloccu, Indrajit Bhattacharya, Şafak Bilici, Bernd Bohnet

Samuel Cahyawijaya, Pengshan Cai, Eduardo Calò, Ronald Cardenas, Boaz Carmeli, Silvia Casola, Khyathi Raghavi Chandu, Elizabeth Clark, Jordan Clive

Kordula De Kuthy, Mingkai Deng, Yuntian Deng, Daniel Deutsch, Mark Dingemanse, Esin Durmus, Ondřej Dušek

Moussa Kamal Eddine

Federico Fancellu, Raquel Fernandez

Albert Gatt, John Glover, Olga Golovneva

Tahmid Hasan, Behnam Hedayatnia, David M. Howcroft, Shulin Huang, Rudali Huidrom

Nikolai Ilinykh

Yacine Jernite, Mayank Jobanputra, Shailza Jolly

Emil Kalbaliyev, Mihir Kale, Marzena Karpinska, Noriaki Kawamae, Dimitar Kazakov, George Kour, Sergey Kovalchuk, Kalpesh Krishna, Rishu Kumar, Vanya Bannihatti Kumar

Harsh Lara, Alberto Lavelli, Hwanhee Lee, Jing Yang Lee, Yinghui Li, Paul Pu Liang, Andreas Liesenfeld, Sijia Liu, Yinhong Liu, Yixin Liu, Zhengzhong Liu, Ehsan Lotfi

Shirong Ma, Aman Madaan, Khyati Mahajan, Abinaya Mahendiran, Andreas Marfurt, Simon Mille, Sebastien Montella

Tapas Nayak, Vitaly Nikolaev, Tadashi Nomoto

Salomey Osei

Alexandros Papangelis, Cheoneum Park, Eunil Park, Tatiana Passali, Diogo Pernes, Sandro Pezzelle, Maja Popović, Jiashu Pu

Vipul Raheja, Anand A. Rajasekar, Vikas Raunak, Ehud Reiter, Leonardo F. R. Ribeiro, Daniele Riboni, Giuseppe Riccardi, Michael Ridenour, Terry Ruas

Rifat Shahriyar, Tianhao Shen, Anna Shvets, Arabella Sinclair, Marco Antonio Sobrevilla Cabezudo, Somayajulu Sripada, Yixuan Su, Barkavi Sundararajan

Bowen Tan, Katherine Thai, Craig Thomson, Grigorios Tsoumakas

Ashish Upadhyay

Jan Philip Wahle, Shira Wein, Michael White, Genta Winata, Zixiu Wu

Yadong Xi, Deyi Xiong, Xinnuo Xu, Yumo Xu

Li Yangning

Alessandra Zarcone, Jiawei Zhou, Yongxin Zhou, Qi Zhu

# Keynote Talk: Challenges in evaluating safety for LLMs

Emily Dinan  
FAIR (Meta AI)

**Abstract:** While research on large language models (LLMs) continues to accelerate, much recent work has called attention to anticipated risks and harms from their use in society. We will discuss challenges in evaluating the relative safety of these models as well as current approaches for doing so. Finally, we will highlight avenues for future research into evaluating and mitigating these harms.

**Bio:** Emily Dinan is a Research Engineer at FAIR (Meta AI) in New York. Her research interests include conversational AI, natural language processing, and safety and responsibility in these fields. Recently she has focused on methods for preventing conversational agents from reproducing biased, toxic, or otherwise harmful language. Prior to joining FAIR, she received her master's degree in Mathematics from the University of Washington.

# Keynote Talk: Instructable and Collaborative Language Models

**Timo Schick**  
FAIR (Meta AI)

**Abstract:** Textual content is often the output of a collaborative writing process — which includes writing text, making comments and changes, finding references, and asking others for help —, but today’s NLP models are only trained to generate the final output of this process. In this talk, we will discuss an alternative approach where models are trained to imitate the entire writing process. We will look at examples of how this enables models to plan and explain their actions, to correct their own mistakes, and to better collaborate with humans. We will also discuss how to make such models better at following human-written instructions.

**Bio:** Timo Schick is a research scientist at FAIR working on few-shot learning in NLP. Previously, he did his PhD at the Center for Information and Language Processing (CIS) in Munich and worked in industry as a data scientist for several years. Timo’s current research focuses on instruction-based learning and teaching language models to collaborate with other entities.



# Keynote Talk: Reflections on Trusting Untrustworthy Language Generators

Sean Welleck

University of Washington

**Abstract:** In his 1984 Turing Award Lecture “Reflections on Trusting Trust”, Ken Thompson famously said “You can’t trust code that you did not totally create yourself”. These words are especially relevant today, as powerful and flexible language models generate natural language and code that is increasingly human-like. However, these same systems challenge our trust, exhibiting odd degeneracies, amplifying biases, and producing flawed reasoning. In this talk, I will introduce two directions for harnessing the potential of these language models while mitigating the risks. First, I will discuss unlearning: removing undesirable behaviors by integrating feedback and learning. Second, I will discuss how integrating language models with trustworthy symbolic systems can open the door to tackling challenging mathematical reasoning tasks. Join me as we explore the path towards trusting untrustworthy language generators.

**Bio:** Sean Welleck is a Postdoctoral Scholar at the University of Washington and the Allen Institute for Artificial Intelligence, working with Yejin Choi. His research focuses on algorithms for natural language generation and machine reasoning, with the aim of minimizing the effort needed to trust the output of AI systems. He has developed unlearning, decoding, and evaluation algorithms for controllable neural language generation, and methods for integrating language models with symbolic systems, with a particular focus on mathematical reasoning. He received his Ph.D. from New York University, where he was advised by Kyunghyun Cho. Outside of his research activities, he hosts the Thesis Review Podcast and enjoys running long distances.

## Table of Contents

<i>Improving abstractive summarization with energy-based re-ranking</i> Diogo Pernes, Afonso Mendes and André F. T. Martins .....	1
<i>Task-driven augmented data evaluation</i> Olga Golovneva, Pan Wei, Khadige Abboud, Charith Peris, Lizhen Tan and Haiyang Yu .....	18
<i>Generating Coherent Narratives with Subtopic Planning to Answer How-to Questions</i> Pengshan Cai, Mo Yu, Fei Liu and Hong Yu .....	26
<i>Weakly Supervised Context-based Interview Question Generation</i> Samiran Pal, Kaamraan Khan, Avinash Kumar Singh, Subhasish Ghosh, Tapas Nayak, Girish Palshikar and Indrajit Bhattacharya .....	43
<i>Analyzing Multi-Task Learning for Abstractive Text Summarization</i> Frederic Thomas Kirstein, Jan Philip Wahle, Terry Ruas and Bela Gipp .....	54
<i>CLSE: Corpus of Linguistically Significant Entities</i> Aleksandr Chuklin, Justin Zhao and Mihir Kale .....	78
<i>Revisiting text decomposition methods for NLI-based factuality scoring of summaries</i> John Glover, Federico Fancellu, Vasudevan Jagannathan, Matthew R. Gormley and Thomas Schaaf .....	97
<i>Semantic Similarity as a Window into Vector- and Graph-Based Metrics</i> Wai Ching Leung, Shira Wein and Nathan Schneider .....	106
<i>Towards In-Context Non-Expert Evaluation of Reflection Generation for Counselling Conversations</i> Zixiu Wu, Simone Balloccu, Rim Helaoui, Diego Reforgiato Recupero and Daniele Riboni ..	116
<i>WikiOmnia: filtration and evaluation of the generated QA corpus on the whole Russian Wikipedia</i> Dina Pisarevskaya and Tatiana Shavrina .....	125
<i>Evaluation of Response Generation Models: Shouldn't It Be Shareable and Replicable?</i> Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella and Giuseppe Riccardi .....	136
<i>Enhancing and Evaluating the Grammatical Framework Approach to Logic-to-Text Generation</i> Eduardo Calò, Elze van der Werf, Albert Gatt and Kees van Deemter .....	148
<i>Controllable Text Generation for All Ages: Evaluating a Plug-and-Play Approach to Age-Adapted Dialogue</i> Lennert Jansen, Štěpán Lars Laichter, Arabella Sinclair, Margot van der Goot, Raquel Fernandez and Sandro Pezzelle .....	172
<i>Template-based Contact Email Generation for Job Recommendation</i> Qiuchi Li and Christina Lioma .....	189
<i>Are Abstractive Summarization Models truly 'Abstractive'? An Empirical Study to Compare the two Forms of Summarization</i> Vinaysheshkar Bannihatti Kumar and Rashmi Gangadharaiah .....	198
<i>Transfer learning for multilingual vacancy text generation</i> Anna Lorincz, David Graus, Dor Lavi and Joao Lebre Magalhaes Pereira .....	207

<i>Plug-and-Play Recipe Generation with Content Planning</i> Yinhong Liu, Yixuan Su, Ehsan Shareghi and Nigel Collier .....	223
<i>Towards Attribute-Entangled Controllable Text Generation: A Pilot Study of Blessing Generation</i> Shulin Huang, Shirong Ma, Yinghui Li, Li Yangning, Shiyang Lin, Haitao Zheng and Ying Shen	235
<i>Towards Attribute-Entangled Controllable Text Generation: A Pilot Study of Blessing Generation</i> Andreas Marfurt and James Henderson .....	248
<i>A Corpus and Evaluation for Predicting Semi-Structured Human Annotations</i> Andreas Marfurt, Ashley Thornton, David Sylvan, Lonneke van der Plas and James Henderson	262
<i>T5QL: Taming language models for SQL generation</i> Samuel David Arcadinho, David Aparicio, Hugo Veiga and Antonio Alegria .....	276
<i>Human perceiving behavior modeling in evaluation of code generation models</i> Sergey V. Kovalchuk, Vadim Lomshakov and Artem Aliev .....	287
<i>Nearest Neighbor Language Models for Stylistic Controllable Generation</i> Severino Trotta, Lucie Flek and Charles Welch .....	295
<i>On reporting scores and agreement for error annotation tasks</i> Maja Popović and Anya Belz .....	306
<i>Answerability: A custom metric for evaluating chatbot performance</i> Pranav Gupta, Anand A. Rajasekar, Amisha Patel, Mandar Kulkarni, Alexander Sunell, Kyung Kim, Krishnan Ganapathy and Anusua Trivedi .....	316
<i>Improved Evaluation of Automatic Source Code Summarisation</i> Jesse Phillips, David Bowes, Mahmoud El-Haj and Tracy Hall .....	326
<i>Most NLG is Low-Resource: here's what we can do about it</i> David M. Howcroft and Dimitra Gkatzia .....	336
<i>GiCCS: A German in-Context Conversational Similarity Benchmark</i> Shima Asaadi, Zahra Kolagar, Alina Liebel and Alessandra Zarcone .....	351
<i>Control Prefixes for Parameter-Efficient Text Generation</i> Jordan Clive, Kris Cao and Marek Rei .....	363
<i>A Survey of Recent Error Annotation Schemes for Automatically Generated Text</i> Rudali Huidrom and Anya Belz .....	383
<i>What's in a (dataset's) name? The case of BigPatent</i> Silvia Casola, Alberto Lavelli and Horacio Saggion .....	399
<i>Measuring the Measuring Tools: An Automatic Evaluation of Semantic Metrics for Text Corpora</i> George Kour, Samuel Ackerman, Eitan Daniel Farchi, Orna Raz, Boaz Carmeli and Ateret Anaby Tavor .....	405
<i>Multilingual Social Media Text Generation and Evaluation with Few-Shot Prompting</i> Mack Blackburn .....	417
<i>Assessing Inter-metric Correlation for Multi-document Summarization Evaluation</i> Michael Ridenour, Ameeta Agrawal and Olubusayo Olabisi .....	428

<i>Factual Error Correction for Abstractive Summaries Using Entity Retrieval</i> Hwanhee Lee, Cheoneum Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Juae Kim and Kyomin Jung .....	439
<i>Coherent Long Text Generation by Contrastive Soft Prompt</i> Guandan Chen, Jiashu Pu, Yadong Xi and Rongsheng Zhang .....	445
<i>Error Analysis of ToTTo Table-to-Text Neural NLG Models</i> Barkavi Sundararajan, Somayajulu Sripada and Ehud Reiter .....	456
<i>Improving Dialogue Act Recognition with Augmented Data</i> Khyati Mahajan, Soham Parikh, Quaizar Vohra, Mitul Tiwari and Samira Shaikh .....	471
<i>Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks? A Case Study of Image Paragraph Generation</i> Nikolai Ilinykh and Simon Dobnik .....	480
<i>20Q: Overlap-Free World Knowledge Benchmark for Language Models</i> Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann and Walter Daelemans .....	494
<i>What Was Your Name Again? Interrogating Generative Conversational Models For Factual Consistency Evaluation</i> Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann and Walter Daelemans .....	509
<i>Narrative Why-Question Answering: A Review of Challenges and Datasets</i> Emil Kalbaliyev and Kairit Sirts .....	520
<i>Exploring a POS-based Two-stage Approach for Improving Low-Resource AMR-to-Text Generation</i> Marco Antonio Sobrevilla Cabezudo and Thiago Pardo .....	531
<i>What Makes Data-to-Text Generation Hard for Pretrained Language Models?</i> Moniba Keymanesh, Adrian Benton and Mark Dredze .....	539
<i>Don't Say What You Don't Know: Improving the Consistency of Abstractive Summarization by Constraining Beam Search</i> Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy and Doug Downey	555

# Program

## Wednesday, December 7, 2022

09:00 - 10:30     *Opening Remarks and Keynote (Sean Welleck)*

10:30 - 11:00     *Coffee Break*

11:00 - 12:30     *Talk Session*

12:30 - 14:00     *Lunch Break*

14:00 - 15:30     *Poster Session*

15:30 - 16:00     *Coffee Break*

16:00 - 17:00     *Keynote (Timo Schick)*

17:00 - 18:30     *Talk Session*

20:00 - 21:00     *Virtual Keynote (Emily Dinan)*

21:00 - 22:30     *Virtual Poster Session*