LREC-COLING 2024

The Third Workhop on Safety for Conversational AI (Safety4ConvAI) @LREC-COLING 2024

Workshop Proceedings

Editors
Tanvi Dinkar, Giuseppe Attanasio, Amanda Cercas Curry,
Ioannis Konstas, Dirk Hovy, Verena Rieser

21 May, 2024
Torino, Italia

Proceedings of Safety4ConvAI: The Third Workhop on Safety for Conversational AI @LREC-COLING 2024

ISBN 978-2-493814-44-9
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Message from the Organisers

This volume documents the Proceedings of the Third Workshop on Safety for Conversational AI (Safety4ConvAI), held on May 21st as part of the LREC-COLING 2024 conference (the joint international conference on Computational Linguistics, Language Resources and Evaluation) in Turin, Italy.

Recently, there has been an explosion of dialogue systems that often use large-scale language and vision models deployed in the real world. These systems have shown dramatic improvements in the ability to mimic conversational behaviours: they can hold long, multi-turn conversations, report facts and events, and engage through text, speech and images.

Conversational models have been quickly adopted by the general public for a range of different and emerging use cases. However, increasing adoption typically means new collateral risks. Like their NLP counterparts, these models still exhibit many concerning problems, such as learning undesirable features present in the training data (e.g. biased, toxic, or otherwise harmful language). Additionally, a fluent dialog agent may give a user false impressions of its 'expertise' and generate harmful advice in response to medically related user queries, manifesting in serious real-world harm. Beyond the context of the answers of these systems, there are aspects of how they present that also pose safety concerns: these systems learn from human data and are built to interact in a natural, 'human-like' way. Designers of these systems may co-opt these unique human-like ways to communicate to drive up user engagement or make a system sound more natural and, by default, more capable – i.e. these systems are anthropomorphised or personified. This anthropomorphism further contributes to the general public's overzealous adoption of these systems, and indeed attributing undue expertise to these systems.

This presents a challenge, as what is deemed as "offensive" or even "sensitive" is both contextually and culturally dependent, and picking up on more subtle examples of unsafe language often requires a level of language understanding that is well beyond current capabilities. For example, when considering interaction, what may be considered safe at an utterance level (e.g. the utterance 'Yes I agree'), may be unsafe at a contextual level (e.g. the utterance is agreeing to hateful/toxic language).

After the success of the second workshop on Safety for End-to-End Conversational AI at the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL) 2021 in Singapore, the Third Workshop on Safety for Conversational AI at LREC-COLING 2024 continued these reflections to promote research into these challenging technical and ethical questions. In this third edition, the workshop received 6 submissions. Of these, 5 contributions have been accepted, and the proceedings consist of 5 accepted archival research papers.

We would like to thank the members of the committee for their commitment to the review process and the authors of these contributions for their valuable investigations and for making this community more vibrant.

Organizing Committee, Safety4ConvAI 2024

Organizing Committee

Organizers:

Tanvi Dinkar, *Heriot Watt University*

Giuseppe Attanasio, *Bocconi University*

Amanda Cercas Curry, *Bocconi University*

Ioannis Konstas, *Heriot Watt University*

Dirk Hovy, *Bocconi University*

Verena Rieser, *Google DeepMind*

Advisory committee:

Gavin Abercrombie, *Heriot Watt University*

Debora Nozza, *Bocconi University*

Dave Howcroft, *Edinburgh Napier University*

Luca Arnaboldi, *University of Birmingham*

Mert Inan, *Northeastern University*

Dilek Hakkani-Tür, *University of Illinois Urbana-Champaign*

Javier Chiya Garcia, *Heriot Watt University*

Flor Miriam Plaza-del-Arco, *Bocconi University*

Angus Adelsee, *Heriot Watt University*

Alessandra Cervone, *Alexa AI*

Mahed Mousavi, *University of Trento*

Fatma Elsafoury, *Weizenbaum institute*

Vittorio Mazzia, *Alexa AI-NLU*

Rosa Alarcon, *Amazon*

Giuseppe Attanasio, *Bocconi University*

Table of Contents

Grounding LLMs to In-prompt Instructions: Reducing Hallucinations Caused by Static Pre-training Knowledge
Angus Addlesee 1

Diversity-Aware Annotation for Conversational AI Safety
Alicia Parrish, Vinodkumar Prabhakaran, Lora Aroyo, Mark Díaz, Christopher M. Homan, Greg Serapio-García, Alex S. Taylor and Ding Wang 8

Using Information Retrieval Techniques to Automatically Repurpose Existing Dialogue Datasets for Safe Chatbot Development
Tunde Oluwaseyi Ajayi, Gaurav Negi, Mihael Arcan and Paul Buitelaar 16

FairPair: A Robust Evaluation of Biases in Language Models through Paired Perturbations
Jane Dwivedi-Yu 28

Learning To See But Forgetting To Follow: Visual Instruction Tuning Makes LLMs More Prone To Jailbreak Attacks
Georgios Pantazopoulos, Amit Parekh, Malvina Nikandrou and Alessandro Suglia 40