Eval4NLP 2023

The 4th Workshop on Evaluation and Comparison of NLP Systems

Proceedings of the Workshop

November 1, 2023

The Eval4NLP organizers gratefully acknowledge the support from the following sponsors.

Sponsors

The Federal Ministry of Education and Research (BMBF) via the grant “Metrics4NLG”

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

ISBN 979-8-89176-021-9

Introduction

Welcome to the Fourth Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP 2023).

The current year has brought astonishing achievements in NLP. Generative large language models (LLMs) like ChatGPT and GPT4 demonstrate wide capabilities in understanding and performing tasks from in-context descriptions without fine-tuning, bringing world-wide attention to the risks and opportunities that arise from current and ongoing research. Further, the release of open-source models like LLaMA and Falcon LLM, better quantization techniques for inference and training, as well as the adaptation of efficient fine-tuning techniques such as LORA accelerate the research progress by allowing hardware and runtime efficiency. Given the ever growing speed of research, fair evaluations and comparisons are of fundamental importance to the NLP community in order to properly track progress. This concerns the creation of benchmark datasets that cover typical use cases and blind spots of existing systems, the designing of metrics for evaluating the performance of NLP systems on different dimensions, and the reporting of evaluation results in an unbiased manner.

We believe that new insights and methodology, particularly in the last 2-3 years, have led to much renewed interest in the workshop topic. The first workshop in the series, Eval4NLP'20, was the first workshop to take a broad and unifying perspective on the subject matter. The second (Eval4NLP'21) and third (Eval4NLP'22) workshop extended this perspective. We believe the fourth workshop continues the tradition of being a reputed platform for presenting and discussing latest advances in NLP evaluation methods and resources.

This year we especially encouraged the submission of works that consider the evaluation of LLMs and their generated content as well as works that leverage LLMs in their evaluation strategies. In fact, to encourage research in this direction, we ran a successful shared task this year on prompting LLMs as explainable metrics. Participants were given a set of open-source LLMs and were tasked with designing prompts and score retreival strategies for automatically scoring machine translation and automatic text summarization outputs without using a reference text.

Our workshop and shared task attracted a lot of attention from the research community. Among the 15 submissions, 9 were accepted for presentation after thorough consideration by the program committee. In addition, there were 9 teams that participated in the shared task. This year's program covers a wide range of topics, including creating a benchmark dataset for identifying and quantifying sexism in language models, evaluation metrics for named entity recognition, probing techniques for large language models, and much more.

We would like to thank all of the authors for their contributions, the program committee for their thoughtful reviews, the keynote speaker for sharing their perspective, and all the attendees for their participation. We believe that all of these will contribute to a lively and successful workshop. Looking forward to meeting you all at Eval4NLP 2023!

Eval4NLP 2023 Organizing Committee,
Daniel Deutsch, Rotem Dror, Steffen Eger, Yang Gao, Christoph Leiter, Juri Opitz, Andreas Rücklé

Organizing Committee

Organizing Committee

Daniel Deutsch, Google
Rotem Dror, University of Haifa
Steffen Eger, Bielefeld University
Yang Gao, Google Research
Christoph Leiter, Bielefeld University
Juri Opitz, Heidelberg University
Andreas Rücklé, Amazon

Program Committee

Reviewers

Omri Abend
Jonas Belouadi
Yanran Chen
Daniel Deutsch
Li Dong
Zi-Yi Dou
Rotem Dror
Steffen Eger
George Foster
Anette Frank
Markus Freitag
Yang Gao
Claire Gardent
Juraj Juraska
Ji-Ung Lee
Christoph Leiter
Lucy Lin
Juri Opitz
Ines Rehbein
Ehud Reiter
Leonardo Ribeiro
Ori Shapira
Julius Steen
Benyou Wang
Ran Zhang
Shiyue Zhang
Wei Zhao

Table of Contents