GenBench 2024

GenBench: The second workshop on generalisation (benchmarking) in NLP

Proceedings of the Workshop

November 16, 2024

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

Supported by

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-182-7

Message from the Organisers

The ability to generalise well is often mentioned as one of the primary desiderata for models of natural language processing (NLP). Yet, there are still many open questions related to what it means for an NLP model to generalise well, and how generalisation should be evaluated. LLMs, trained on gigantic training corpora that are, at best, hard to analyse or might not be publicly available at all, bring a new set of challenges to the topic. The second GenBench workshop on generalisation (benchmarking) in NLP aims to serve as a cornerstone to catalyse research on generalisation in the NLP community. The workshop has two concrete goals: to bring together different expert communities to discuss challenging questions relating to generalisation in NLP and to establish a shared platform for state-of-the-art generalisation testing in NLP through our Collaborative Benchmarking Task (CBT). We started the CBT last year; this year's CBT is solely LLM-focused.

The second edition of the workshop was held at EMNLP 2024 in Miami, Florida. For this edition, we accepted 11 archival papers in our main track, 2 archival papers for our CBT, and 9 extended abstracts. The workshop also provided a platform for the authors of EMNLP Findings papers related to the workshop's topic to present their work as a poster at the workshop.

The workshop would not have been possible without the dedication of the programme committee, whom we would like to thank for their contributions. We would also like to thank Amazon for their sponsorship of 10,000 dollars, which we used to grant travel awards to allow participants who could otherwise not have attended to participate in the workshop, and to grant two best paper awards. Lastly, we are grateful to our invited speakers, Pascale Fung, Najoung Kim, and Sameer Singh, for contributing to our programme.

Organizing Committee

Workshop Organizers

Dieuwke Hupkes, Meta
Verna Dankers, University of Edinburgh
Khuyagbaatar Batsuren, Openstream AI
Amirhossein Kazemnejad, McGill University and Mila
Christos Christodoulopoulos, Amazon Research
Mario Giulianelli, ETH Zürich
Ryan Cotterell, ETH Zürich

Program Committee

Reviewers

Jonathan Brophy, University of Oregon
Lisa Bylinina, University of Groningen
Robert Frank, Yale University
Yangfeng Ji, University of Virginia
Jenny Kunz, Linköping University
Matthias Lindemann, University of Edinburgh
R. Thomas McCoy, Princeton University
Anmol Nayak, Bosch
Sanchit Sinha, University of Virginia, Charlottesville
Shane Steinert-Threlkeld, University of Washington, Seattle
Swetasudha Panda, Oracle
Koji Mineshima, Keio University
Tatiana Shavrina, Artificial Intelligence Research Institute
Jithendra Vepa, Idiap Research Institute
Rudolf Rosa, Charles University, Prague
Erenay Dayanik, Amazon
Antske Fokkens, VU University Amsterdam
Richard Futrell, University of California, Irvine
Djamé Seddah, Inria Paris
Lis Pereira, National Institute of Information and Communications Technology (NICT), National Institute of Advanced Industrial Science and Technology
Cassandra Jacobs, State University of New York, Buffalo
Marco Basaldella, Amazon
Houman Mehrafarin, Heriot-Watt University
Jean-Philippe Fauconnier, Apple
Deepanshu Gupta, Apple
Fabio Massimo Zanzotto, University of Rome Tor Vergata
Mira Moukheiber, Massachusetts Institute of Technology
Bryan Eikema, University of Amsterdam
Kate McCurdy, Universität des Saarlandes
Michal Štefánik, Masaryk University
Coleman Haley, University of Edinburgh
Michael Eric Goodale, Ecole Normale Supérieure de Paris
Rimvydas Rubavicius, Edinburgh University
Gautier Dagan, University of Edinburgh
Parsa Bagherzadeh, McGill University
Subham De, Meta AI
Churan Zhi, University of California, San Diego
Bogdan Kulynych, CHUV - University Hospital Lausanne
Jirui Qi, University of Groningen
Aditya Kaushik Surikuchi, University of Amsterdam

Keynote Talk
Invited Talk 1

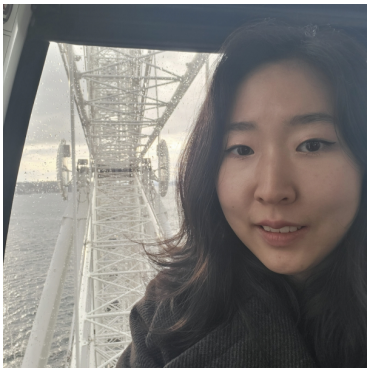
Pascale Fung
Hong Kong University of Science and Technology

2024-11-16 – Time: **09:15 – 10:00** –

Bio: Pascale Fung is a Chair Professor at the Department of Electronic & Computer Engineering at The Hong Kong University of Science & Technology (HKUST), and a visiting professor at the Central Academy of Fine Arts in Beijing. She is an elected Fellow of the Association for the Advancement of Artificial Intelligence (AAAI) for her significant contributions to the field of conversational AI and to the development of ethical AI principles and algorithms", an elected Fellow of the Association for Computational Linguistics (ACL) for her "significant contributions towards statistical NLP, comparable corpora, and building intelligent systems that can understand and empathize with humans". She is an Fellow of the Institute of Electrical and Electronic Engineers (IEEE) for her "contributions to human-machine interactions" and an elected Fellow of the International Speech Communication Association for "fundamental contributions to the interdisciplinary area of spoken language human-machine interactions". She is the Director of HKUST Centre for AI Research (CAiRE), an interdisciplinary research centre promoting human-centric AI. She co-founded the Human Language Technology Center (HLTC). She is an affiliated faculty with the Robotics Institute and the Big Data Institute at HKUST. She is the founding chair of the Women Faculty Association at HKUST. She is an expert on the Global Future Council, a think tank for the World Economic Forum. She represents HKUST on Partnership on AI to Benefit People and Society. She is on the Board of Governors of the IEEE Signal Processing Society. She is a member of the IEEE Working Group to develop an IEEE standard - Recommended Practice for Organizational Governance of Artificial Intelligence. She was a Distinguished Consultant on Responsible AI at Meta in 2022, and a Visiting Faculty Researcher at Google in 2023. Her research team has won several best and outstanding paper awards at ACL, ACL and NeurIPS workshops.

Keynote Talk
Invited Talk 2

Najoung Kim
Boston University



2024-11-16 – Time: **11:00 – 11:45** –

Bio: Najoung Kim is an Assistant Professor at the Department of Linguistics and an affiliate faculty in the Department of Computer Science at Boston University. She is also currently a visiting faculty researcher at Google DeepMind. Before joining BU, she was a Faculty Fellow at the Center for Data Science at New York University and received her PhD in Cognitive Science at Johns Hopkins University. She is interested in studying meaning in both human and machine learners, especially ways in which they generalize to novel inputs and ways in which they treat implicit meaning. Her research has been supported by NSF and Google, and has received awards at venues such as ACL and *SEM.

Keynote Talk

Invited Talk 3

Sameer Singh
University of California, Irvine

2024-11-16 – Time: **15:00 – 15:45** –

Bio: Dr. Sameer Singh is a Professor of Computer Science at UC Irvine. He is working primarily on the robustness and interpretability of machine learning algorithms and models that reason with text and structure for natural language processing. Sameer was a postdoctoral researcher at the University of Washington and received his Ph.D. from the University of Massachusetts, Amherst. He has been named the Kavli Fellow by the National Academy of Sciences, received the NSF CAREER award, UCI Distinguished Early Career Faculty award, the Hellman Faculty Fellowship, and was selected as a DARPA Riser. His group has received funding from Allen Institute for AI, Amazon, NSF, DARPA, Adobe Research, Hasso Plattner Institute, NEC, Base 11, and FICO. Sameer has published extensively at machine learning and natural language processing venues and received conference paper awards at KDD 2016, ACL 2018, EMNLP 2019, AKBC 2020, ACL 2020, and NAACL 2022.

Table of Contents

Evaluating the fairness of task-adaptive pretraining on unlabeled test data before few-shot text classification
Kush Dubey 1

From Language to Pixels: Task Recognition and Task Learning in LLMs
Janek Falkenstein, Carolin M. Schuster, Alexander H. Berger and Georg Groh 27

The SlayQA benchmark of social reasoning: testing gender-inclusive generalization with neopronouns
Bastian Bunzeck and Sina Zarrieß 42

Automated test generation to evaluate tool-augmented LLMs as conversational AI agents
Samuel Arcadinho, David Oliveira Aparicio and Mariana S. C. Almeida 54

MMLU-SR: A Benchmark for Stress-Testing Reasoning Capability of Large Language Models
Wentian Wang, Sarthak Jain, Paul Kantor, Jacob Feldman, Lazaros Gallos and Hao Wang 69

MLissard: Multilingual Long and Simple Sequential Reasoning Benchmarks
Mirelle Candida Bueno, Roberto Lotufo and Rodrigo Frassetto Nogueira 86

MultiPragEval: Multilingual Pragmatic Evaluation of Large Language Models
Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park and Sungeun Lee 96

Beyond the Numbers: Transparency in Relation Extraction Benchmark Creation and Leaderboards
Varvara Arzt and Allan Hanbury 120

Is artificial intelligence still intelligence? LLMs generalize to novel adjective-noun pairs, but don't mimic the full human distribution
Hayley Ross, Kathryn Davidson and Najoung Kim 131

CHIE: Generative MRC Evaluation for in-context QA with Correctness, Helpfulness, Irrelevancy, and Extraneousness Aspects
Wannaphong Phatthiyaphaibun, Surapon Nonesung, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Jitkapat Sawatphol, Ekapol Chuangsuwanich and Sarana Nutanong 154

Investigating the Generalizability of Pretrained Language Models across Multiple Dimensions: A Case Study of NLI and MRC
Ritam Dutt, Sagnik Ray Choudhury, Varun Venkat Rao, Carolyn Rose and V.G.Vinod Vydiswaran 165

OmniDialog: A Multimodal Benchmark for Generalization Across Text, Visual, and Audio Modalities
Anton Razzhigaev, Maxim Kurkin, Elizaveta Goncharova, Irina Abdullaeva, Anastasia Lysenko, Alexander Panchenko, Andrey Kuznetsov and Denis Dimitrov 183

Towards a new Benchmark for Emotion Detection in NLP: A Unifying Framework of Recent Corpora
Anna Koufakou, Elijah Nieves and John Peller 196

Program

Saturday, November 16, 2024

09:00 - 09:15 *Opening Remarks*

09:15 - 10:00 *Keynote 1 by Pascale Fung*

10:00 - 10:30 *Oral presentations*

Is artificial intelligence still intelligence? LLMs generalize to novel adjective-noun pairs, but don't mimic the full human distribution
Hayley Ross, Kathryn Davidson and Najoung Kim

Investigating the Generalizability of Pretrained Language Models across Multiple Dimensions: A Case Study of NLI and MRC
Ritam Dutt, Sagnik Ray Choudhury, Varun Venkat Rao, Carolyn Rose and V.G.Vinod Vydiswaran

10:30 - 11:00 *Morning Coffee Break*

11:00 - 11:45 *Keynote 2 by Najoung Kim*

11:45 - 12:30 *Spotlight talks*

The SlayQA benchmark of social reasoning: testing gender-inclusive generalization with neopronouns
Bastian Bunzeck and Sina Zarrieß

MMLU-SR: A Benchmark for Stress-Testing Reasoning Capability of Large Language Models
Wentian Wang, Sarthak Jain, Paul Kantor, Jacob Feldman, Lazaros Gallos and Hao Wang

MLissard: Multilingual Long and Simple Sequential Reasoning Benchmarks
Mirelle Candida Bueno, Roberto Lotufo and Rodrigo Frassetto Nogueira

MultiPragEval: Multilingual Pragmatic Evaluation of Large Language Models
Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park and Sungeun Lee

OmniDialog: A Multimodal Benchmark for Generalization Across Text, Visual, and Audio Modalities
Anton Razzhigaev, Maxim Kurkin, Elizaveta Goncharova, Irina Abdullaeva, Anastasia Lysenko, Alexander Panchenko, Andrey Kuznetsov and Denis Dimitrov

Saturday, November 16, 2024 (continued)

12:30 - 13:45 *Lunch break*

13:45 - 15:00 *Poster session*

Evaluating the fairness of task-adaptive pretraining on unlabeled test data before few-shot text classification
Kush Dubey

From Language to Pixels: Task Recognition and Task Learning in LLMs
Janek Falkenstein, Carolin M. Schuster, Alexander H. Berger and Georg Groh

Automated test generation to evaluate tool-augmented LLMs as conversational AI agents
Samuel Arcadinho, David Oliveira Aparicio and Mariana S. C. Almeida

Beyond the Numbers: Transparency in Relation Extraction Benchmark Creation and Leaderboards
Varvara Arzt and Allan Hanbury

CHIE: Generative MRC Evaluation for in-context QA with Correctness, Helpfulness, Irrelevancy, and Extraneousness Aspects
Wannaphong Phatthiyaphaibun, Surapon Nonesung, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Jitkapat Sawatphol, Ekapol Chuangsuwanich and Sarana Nutanong

Towards a new Benchmark for Emotion Detection in NLP: A Unifying Framework of Recent Corpora
Anna Koufakou, Elijah Nieves and John Peller

MLissard: Multilingual Long and Simple Sequential Reasoning Benchmarks
Mirelle Candida Bueno, Roberto Lotufo and Rodrigo Frassetto Nogueira

MultiPragEval: Multilingual Pragmatic Evaluation of Large Language Models
Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park and Sungeun Lee

OmniDialog: A Multimodal Benchmark for Generalization Across Text, Visual, and Audio Modalities
Anton Razzhigaev, Maxim Kurkin, Elizaveta Goncharova, Irina Abdullaeva, Anastasia Lysenko, Alexander Panchenko, Andrey Kuznetsov and Denis Dimitrov

The SlayQA benchmark of social reasoning: testing gender-inclusive generalization with neopronouns
Bastian Bunzeck and Sina Zarrieß

Saturday, November 16, 2024 (continued)

MMLU-SR: A Benchmark for Stress-Testing Reasoning Capability of Large Language Models
Wentian Wang, Sarthak Jain, Paul Kantor, Jacob Feldman, Lazaros Gallos and Hao Wang

15:00 - 15:45 *Keynote 3 by Sameer Singh*

15:45 - 16:00 *Afternoon Coffee Break*

16:00 - 16:30 *Panel*

16:30 - 16:45 *Closing Remarks and Best Paper Award*