

GenBench 2024: The Second Workshop on Generalisation (Benchmarking) in NLP

Miami, Florida, USA
16 November 2024

ISBN: 979-8-3313-0847-6

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2024) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006

Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Evaluating the fairness of task-adaptive pretraining on unlabeled test data before few-shot text classification</i>	
Kush Dubey	1
<i>From Language to Pixels: Task Recognition and Task Learning in LLMs</i>	
Janek Falkenstein, Carolin M. Schuster, Alexander H. Berger and Georg Groh	27
<i>The SlayQA benchmark of social reasoning: testing gender-inclusive generalization with neopronouns</i>	
Bastian Bunzeck and Sina Zarriß	42
<i>Automated test generation to evaluate tool-augmented LLMs as conversational AI agents</i>	
Samuel Arcadinho, David Oliveira Aparicio and Mariana S. C. Almeida	54
<i>MMLU-SR: A Benchmark for Stress-Testing Reasoning Capability of Large Language Models</i>	
Wentian Wang, Sarthak Jain, Paul Kantor, Jacob Feldman, Lazaros Gallos and Hao Wang	69
<i>MLlissard: Multilingual Long and Simple Sequential Reasoning Benchmarks</i>	
Mirelle Candida Bueno, Roberto Lotufo and Rodrigo Frassetto Nogueira	86
<i>MultiPragEval: Multilingual Pragmatic Evaluation of Large Language Models</i>	
Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park and Sungeun Lee	96
<i>Beyond the Numbers: Transparency in Relation Extraction Benchmark Creation and Leaderboards</i>	
Varvara Arzt and Allan Hanbury	120
<i>Is artificial intelligence still intelligence? LLMs generalize to novel adjective-noun pairs, but don't mimic the full human distribution</i>	
Hayley Ross, Kathryn Davidson and Najoung Kim	131
<i>CHIE: Generative MRC Evaluation for in-context QA with Correctness, Helpfulness, Irrelevancy, and Extraneousness Aspects</i>	
Wannaphong Phatthiyaphaibun, Surapon Nonesung, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Jitkapat Sawatphol, Ekapol Chuangsuwanich and Sarana Nutanong	154
<i>Investigating the Generalizability of Pretrained Language Models across Multiple Dimensions: A Case Study of NLI and MRC</i>	
Ritam Dutt, Sagnik Ray Choudhury, Varun Venkat Rao, Carolyn Rose and V.G.Vinod Vydiswaran	165
<i>OmniDialog: A Multimodal Benchmark for Generalization Across Text, Visual, and Audio Modalities</i>	
Anton Razzhigaev, Maxim Kurkin, Elizaveta Goncharova, Irina Abdullaeva, Anastasia Lysenko, Alexander Panchenko, Andrey Kuznetsov and Denis Dimitrov	183
<i>Towards a new Benchmark for Emotion Detection in NLP: A Unifying Framework of Recent Corpora</i>	
Anna Koufakou, Elijah Nieves and John Peller	196