GenBench 2023

# **GenBench:** The first workshop on generalisation (benchmarking) in NLP

**Proceedings of the Workshop** 

December 6, 2023

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

Supported by



©2023 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

ISBN 979-8-89176-042-4

### 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). However, how generalisation should be defined and evaluated, or when it is particularly important, is a far from trivial question. The GenBench workshop on generalisation (benchmarking) in NLP aims to provide a platform to discuss challenging questions related to generalisation in NLP and establish a shared platform for state-of-the-art generalisation testing. We invited submitters to contribute work discussing generalisation in NLP and also held a collaborative benchmarking task, for which we called for submissions of challenging generalisation tests.

The first edition of the workshop was held at EMNLP 2023 in Singapore. For this edition, we accepted 11 archival papers in our main track, 7 archival papers for our collaborative benchmarking track, and 6 extended abstracts. The workshop also provided a platform for the authors of 29 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 5000 dollars, which we used to fund one of our invited speakers, to grant travel awards to allow participants that could otherwise not have attended to participate in the workshop, and to grant two awards, to the best submitted paper and best submitted benchmark. Lastly, we are grateful to our invited speakers, Adina Williams, Anna Rogers, and Tatsunori Hashimoto, for contributing to our programme.

### **Organizing Committee**

### **Workshop Organizers**

Dieuwke Hupkes, Meta Verna Dankers, University of Edinburgh Khuyagbaatar Batsuren, National University of Mongolia Koustuv Sinha, Meta Amirhossein Kazemnejad, McGill University and Mila Christos Christodoulopoulos, Amazon Research Ryan Cotterell, ETH Zürich Elia Bruni, University of Osnabrück

### **Program Committee**

#### Reviewers

Lisa Beinborn, Vrije Universiteit Amsterdam Jonathan Brophy, University of Oregon Lisa Bylinina, University of Groningen Benoit Crabbé, Université de Paris Ghazi Felhi, University Paris 13 Robert Frank, Yale University Mario Giulianelli, University of Amsterdam Yangfeng Ji, University of Virginia Robin Jia, University of Southern California Richard Johansson, Chalmers University Jenny Kunz, Linköping University Carolin Lawrence, NEC Laboratories Europe Alessandro Lenci, University of Pisa Sheng Liang, Ludwig-Maximilians-Universität München Tomasz Limisiewicz, Charles University Prague Matthias Lindemann, University of Edinburgh R. Thomas McCoy, Princeton University William Merrill, New York University Anmol Nayak, Bosch Joakim Nivre, Uppsala University Maria Ryskina, Massachusetts Institute of Technology Niladri S. Chatterji, Stanford University Hendrik Schuff, Technische Universität Darmstadt Rico Sennrich, University of Zurich Mattia Setzu, University of Pisa Sanchit Sinha, University of Virginia, Charlottesville Shane Steinert-Threlkeld, University of Washington, Seattle Aarne Talman, Basement AI Oskar van der Wal, University of Amsterdam Alex Warstadt, New York University Hao Yang, Beijing University of Post and Telecommunication Naoki Yoshinaga, Institute of Industrial Science, the University of Tokyo Dylan Z Slack, University of California, Irvine Sheng Zhang, Amazon Yichu Zhou, Yahoo

# Keynote Talk: Invited Talk 1

Anna Rogers IT University of Copenhagen



2023-12-06 - Time: 555 -

**Abstract:** One of the frequent points in the mainstream narrative about large language models is that they have "emergent properties" (sometimes even dangerous enough to be considered existential risk to mankind). However, there is much disagreement about even the very definition of such properties. If they are understood as a kind of generalization beyond training data - as something that a model does without being explicitly trained for it - I argue that we have not in fact established the existence of any such properties, and at the moment we do not even have the methodology for doing so.

**Bio:** Dr. Anna Rogers is an assistant professor at IT University of Copenhagen working on analysis, interpretability, and evaluation of NLP models, their societal impact, and NLP research methodology.

# Keynote Talk: Invited Talk 2

Adina Williams Meta AI



2023-12-06 - Time: 675 -

**Bio:** Adina is a Research Scientist at Meta on the Fundamental AI Research (FAIR) team in NYC. Her research spans several topics in NLP and computational linguistics, with a focus on dataset creation and model evaluation for humanlikeness, fairness, generalization and robustness.

# Keynote Talk: Invited Talk 3

Tatsunori Hashimoto Stanford University



2023-12-06 - Time: 840 -

**Abstract:** Instruction following language models have shown a remarkable ability to perform a wide range of tasks with little to no additional training data. Do these abilities come from a revolution in pretraining and instruction-following, or are there other more mundane explanations for how these models work? In this talk, I will discuss our efforts to answer these questions by replicating instruction-following models that generalize across tasks, studying the consistency of these models across different task formats, and building tests for benchmark contamination in pretraining.

**Bio:** Tatsunori Hashimoto is an Assistant Professor in the Computer Science Department at Stanford University. He is a member of the statistical machine learning and natural language processing groups at Stanford, and his research uses tools from statistics to make machine learning systems more robust and trustworthy — especially in complex systems such as large language models. He is a Kavli fellow, a Sony and Amazon research award winner, and his work has been recognized with best paper awards at ICML and CHI. Before becoming an Assistant Professor, he was a postdoctoral researcher at Stanford with Percy Liang and John Duchi and received his Ph.D. from MIT under the supervision of Tommi Jaakkola and David Gifford.

# **Table of Contents**

90% F1 Score in Relation Triple Extraction: Is it Real?Pratik Saini, Samiran Pal, Tapas Nayak and Indrajit Bhattacharya1
<i>GenCodeSearchNet: A Benchmark Test Suite for Evaluating Generalization in Programming Language Understanding</i> Andor Diera, Abdelhalim Dahou, Lukas Galke, Fabian Karl, Florian Sihler and Ansgar Scherp12
Adapt and Decompose: Efficient Generalization of Text-to-SQL via Domain Adapted Least-To-Most Prompting Aseem Arora, Shabbirhussain Bhaisaheb, Harshit Nigam, Manasi Patwardhan, Lovekesh Vig and
Gautam Shroff
<i>Robust Code Summarization</i> Debanjan Mondal, Abhilasha Lodha, Ankita Sahoo and Beena Kumari
Temporal Generalizability in Multimodal Misinformation DetectionNataliya Stepanova and Björn Ross76
Robust Generalization Strategies for Morpheme Glossing in an Endangered Language Documentation Context Michael Ginn and Alexis Palmer
Walking a Tightrope – Evaluating Large Language Models in High-Risk Domains   Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner and Carolin Lawrence99
Latent Feature-based Data Splits to Improve Generalisation Evaluation: A Hate Speech Detection Case Study
Maike Züfle, Verna Dankers and Ivan Titov112
Syntax-Guided Transformers: Elevating Compositional Generalization and Grounding in Multimodal Environments Danial Kamali and Parisa Kordjamshidi
<i>mSCAN: A Dataset for Multilingual Compositional Generalisation Evaluation</i> Amélie Reymond and Shane Steinert-Threlkeld
Inductive Bias Is in the Eye of the Beholder Michael Wilson and Robert Frank
Blackbird Language Matrices Tasks for GeneralizationPaola Merlo, Chunyang Jiang, Giuseppe Samo and Vivi Nastase163
In-Context Learning for Text Classification with Many Labels Aristides Milios, Siva Reddy and Dzmitry Bahdanau
GQG: Generalized Quantifier Generalization - A Dataset for Evaluating Quantifier Semantics Under- standing in Language Models Leroy Zhifei Wang and Shane Steinert-Threlkeld

Cross-Lingual Data Augmentation For Thai Question-Answering

Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet and Peerat Limkonchotiwat 193

On using distribution-based compositionality assessment to evaluate compositional ge	eneralisation in
machine translation	
Anssi Moisio, Mathias Creutz and Mikko Kurimo	
Shifted PAUQ: Distribution shift in text-to-SQL	
Oleg Somov and Elena Tutubalina	

### Program

#### Wednesday, December 6, 2023

- 09:00 09:15 Opening Remarks
- 09:15 10:00 Keynote 1 by Anna Rogers: A sanity check on emergent properties
- 10:00 11:15 *Poster Session 1*

Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models Jirui Qi, Raquel Fernández and Arianna Bisazza

*Temporal Generalizability in Multimodal Misinformation Detection* Nataliya Stepanova and Björn Ross

Robust Generalization Strategies for Morpheme Glossing in an Endangered Language Documentation Context Michael Ginn and Alexis Palmer

*Walking a Tightrope – Evaluating Large Language Models in High-Risk Domains* Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner and Carolin Lawrence

*The ICL Consistency Test* Lucas Weber, Elia Bruni and Dieuwke Hupkes

Generalizability and Robustness of Large Language Models Detecting Alzheimer's Disease from Speech Jekaterina Novikova

Syntax-Guided Transformers: Elevating Compositional Generalization and Grounding in Multimodal Environments Danial Kamali and Parisa Kordjamshidi

*Inductive Bias Is in the Eye of the Beholder* Michael Wilson and Robert Frank

On using distribution-based compositionality assessment to evaluate compositional generalisation in machine translation Anssi Moisio, Mathias Creutz and Mikko Kurimo

#### Wednesday, December 6, 2023 (continued)

Morning Coffee Break

- 11:15 12:00 *Keynote 2 by Adina Williams: Evaluation after the LLM boom: frustrations, fallacies, and the future*
- 12:00 12:30 *CBT Spotlights*

10:30 - 11:00

GenCodeSearchNet: A Benchmark Test Suite for Evaluating Generalization in Programming Language Understanding Andor Diera, Abdelhalim Dahou, Lukas Galke, Fabian Karl, Florian Sihler and

Ansgar Scherp

Latent Feature-based Data Splits to Improve Generalisation Evaluation: A Hate Speech Detection Case Study Maike Züfle, Verna Dankers and Ivan Titov

On using distribution-based compositionality assessment to evaluate compositional generalisation in machine translation Anssi Moisio, Mathias Creutz and Mikko Kurimo

Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models

Jirui Qi, Raquel Fernández and Arianna Bisazza

- 12:30 14:00 Lunch break
- 14:00 14:45 *Keynote 3 by Tatsunori Hashimoto: Understanding generalization for instruction following and black-box language models*
- 14:45 15:30 Oral presentations

#### Evaluating Neural Language Models as Cognitive Models of Language Acquisition

Hector Javier Vazquez Martinez, Annika Lea Heuser, Charles Yang and Jordan Kodner

*Robust Code Summarization* Debanjan Mondal, Abhilasha Lodha, Ankita Sahoo and Beena Kumari

### Cross-Lingual Data Augmentation For Thai Question-Answering

Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet and Peerat Limkonchotiwat

#### Wednesday, December 6, 2023 (continued)

- 15:30 16:00 Afternoon Coffee Break
- 16:00 17:00 Poster session 2 (hybrid)

90% F1 Score in Relation Triple Extraction: Is it Real? Pratik Saini, Samiran Pal, Tapas Nayak and Indrajit Bhattacharya

mSCAN: A Dataset for Multilingual Compositional Generalisation Evaluation Amélie Reymond and Shane Steinert-Threlkeld

GQG: Generalized Quantifier Generalization - A Dataset for Evaluating Quantifier Semantics Understanding in Language Models Leroy Zhifei Wang and Shane Steinert-Threlkeld

Fighting Bias with Bias: Promoting Model Robustness by Amplifying Dataset **Biases** Yuval Reif and Roy Schwartz

GenCodeSearchNet: A Benchmark Test Suite for Evaluating Generalization in Programming Language Understanding

Andor Diera, Abdelhalim Dahou, Lukas Galke, Fabian Karl, Florian Sihler and Ansgar Scherp

Latent Feature-based Data Splits to Improve Generalisation Evaluation: A Hate Speech Detection Case Study Maike Züfle, Verna Dankers and Ivan Titov

Blackbird Language Matrices Tasks for Generalization Paola Merlo, Chunyang Jiang, Giuseppe Samo and Vivi Nastase

In-Context Learning for Text Classification with Many Labels Aristides Milios, Siva Reddy and Dzmitry Bahdanau

Shifted PAUQ: Distribution shift in text-to-SQL Oleg Somov and Elena Tutubalina

17:00 - 17:30 Pannel

## Wednesday, December 6, 2023 (continued)

17:30 - 17:45 Closing Remarks and Best Paper Award