

Workshop on Challenges & Perspectives in Creating Large Language Models

Dublin, Ireland
27 May 2022

ISBN: 978-1-7138-6727-2

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© (2022) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2023)

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>Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora</i> Xisen Jin, Dejjiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold and Xiang Ren	1
<i>Using ASR-Generated Text for Spoken Language Modeling</i> Nicolas Hervé, Valentin Pelloin, Benoit Favre, Franck Dary, Antoine Laurent, Sylvain Meignier and Laurent Besacier	17
<i>You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings</i> Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla and Oskar Van Der Wal	26
<i>Diverse Lottery Tickets Boost Ensemble from a Single Pretrained Model</i> Sosuke Kobayashi, Shun Kiyono, Jun Suzuki and Kentaro Inui	42
<i>UNIREX: A Unified Learning Framework for Language Model Rationale Extraction</i> Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren and Hamed Firooz	51
<i>Pipelines for Social Bias Testing of Large Language Models</i> Debora Nozza, Federico Bianchi and Dirk Hovy	68
<i>Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0</i> Francesco De Toni, Christopher Akiki, Javier De La Rosa, Clémentine Fourrier, Enrique Manja- vacas, Stefan Schweter and Daniel Van Strien	75
<i>A Holistic Assessment of the Carbon Footprint of Noor, a Very Large Arabic Language Model</i> Imad Lakim, Ebtessam Almazrouei, Ibrahim Abualhaol, Merouane Debbah and Julien Launay	84
<i>GPT-NeoX-20B: An Open-Source Autoregressive Language Model</i> Sidney Black, Stella Biderman, Eric Hallahan, Quentin Gregory Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Martin Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang and Samuel Weinbach	95
<i>Dataset Debt in Biomedical Language Modeling</i> Jason Alan Fries, Natasha Seelam, Gabriel Altay, Leon Weber, Myungsun Kang, Debajyoti Datta, Ruisi Su, Samuele Garda, Bo Wang, Simon Ott, Matthias Samwald and Wojciech Kusa	137
<i>Emergent Structures and Training Dynamics in Large Language Models</i> Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin and Aaron Gokaslan	146
<i>Foundation Models of Scientific Knowledge for Chemistry: Opportunities, Challenges and Lessons Learned</i> Sameera Horawalavithana, Ellyn Ayton, Shivam Sharma, Scott Howland, Megha Subramanian, Scott Vasquez, Robin Cosbey, Maria Glenski and Svitlana Volkova	160