CONDA 2024

The First Data Contamination Workshop

Proceedings of the Workshop

August 16, 2024

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

Gold

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 979-8-89176-135-3

Introduction

Welcome to the Proceedings of the first iteration of the Workshop on Data Contamination (CONDA). The workshop is hosted at ACL 2024, in Thailand, on August 16, 2024.

Data contamination in NLP where evaluation data is inadvertently included in pre-training corpora, has become a concern in recent times. The growing scale of both models and data, coupled with unsupervised web crawling, has led to the inclusion of segments from evaluation benchmarks in the pre-training datasets of large language models (LLMs). The noisy nature of internet data makes it difficult to prevent this contamination from happening, or even detect when it has happened. Crucially, when evaluation data becomes part of pre-training data, it introduces biases and can artificially inflate the performance of LLMs on specific tasks or benchmarks. This poses a challenge for fair and unbiased evaluation of NLP models, as their performance may not accurately reflect their generalization capabilities.

We received 16 submissions, of which we accepted 13 for presentation at the workshop.

We extend heartfelt thanks to our program committee, our participants, and all authors who submitted papers for consideration—your engagement has been critical to the success of the workshop. We also thank Amazon, Google, and Hugging Face for generous sponsorship. Finally, we thank the ACL 2024 organizers for their hard work and support.

The CONDA Workshop Organizers,

Oscar Sainz, Iker García Ferrero, Eneko Agirre, Jon Ander Campos, Alon Jacovi, Yanai Elazar, Yoav Goldberg

Organizing Committee

Program Chairs

Oscar Sainz, HiTZ Center - Ixa, University of the Basque Country
Iker García Ferrero, HiTZ Center - Ixa, University of the Basque Country
Eneko Agirre, HiTZ Center - Ixa, University of the Basque Country
Jon Ander Campos, Cohere
Alon Jacovi, Bar-Ilan University
Yanai Elazar, Allen Institute for Artificial Intelligence, University of Washington
Yoav Goldberg, Bar-Ilan University, Allen Institute for Artificial Intelligence

Program Committee

Program Chairs

Eneko Agirre, University of the Basque Country (UPV/EHU)
Jon Ander Campos, Cohere
Yanai Elazar, Allen Institute for Artificial Intelligence and University of Washington
Iker García-Ferrero, University of the Basque Country (UPV/EHU)
Yoav Goldberg, Bar-Ilan University, Allen Institute for Artificial Intelligence and Bar Ilan University
Alon Jacovi, Google
Oscar Sainz, University of the Basque Country (UPV/EHU)

Reviewers

Rodrigo Agerri, Iñigo Alonso

Jeremy Barnes, Ander Barrena

Iker García-Ferrero, Shahriar Golchin, Itziar Gonzalez-Dios

EunJeong Hwang

Alon Jacovi

Yucheng LI, Oier Lopez De Lacalle

Ian Magnusson

Naiara Perez

Royi Rassin, Sahithya Ravi

Oscar Sainz, Hailey Schoelkopf, Preethi Seshadri

Yi Chern Tan, Kunvar Thaman, Kunvar Thaman

Keynote Talk
On the value of carefully measuring data

Margaret Mitchell
HuggingFace
2024-08-16 09:00:00 – Room: **TBA**

Abstract: Just as we evaluate models, we should measure data. Measuring data involves quantifying different aspects of its composition, such as counts of the top-represented domains, or correlations between sensitive identity terms and other concepts. In this talk, I will define the problem of measuring data and unpack how it can be applied to automatically curating distinct training and evaluation datasets for ML models.

Bio: Margaret Mitchell is a researcher focused on the ins and outs of machine learning and ethics-informed AI development in tech. She has published around 100 papers on natural language generation, assistive technology, computer vision, and AI ethics, and holds multiple patents in the areas of conversation generation and sentiment classification. She has recently received recognition as one of Time's Most Influential People of 2023. She currently works at Hugging Face as Chief Ethics Scientist, driving forward work in the ML development ecosystem, ML data governance, AI evaluation, and AI ethics. She previously worked at Google AI as a Staff Research Scientist, where she founded and co-led Google's Ethical AI group, focused on foundational AI ethics research and operationalizing AI ethics Google-internally. Before joining Google, she was a researcher at Microsoft Research, focused on computer vision-to-language generation; and was a postdoc at Johns Hopkins, focused on Bayesian modeling and information extraction. She holds a PhD in Computer Science from the University of Aberdeen and a Master's in computational linguistics from the University of Washington. While earning her degrees, she also worked from 2005-2012 on machine learning, neurological disorders, and assistive technology at Oregon Health and Science University. She has spearheaded a number of workshops and initiatives at the intersections of diversity, inclusion, computer science, and ethics. Her work has received awards from Secretary of Defense Ash Carter and the American Foundation for the Blind, and has been implemented by multiple technology companies. She likes gardening, dogs, and cats.

Keynote Talk

Evaluation data contamination: how much is there, and how much does it actually matter?

Dieuwke Hupkes
Meta

2024-08-16 09:45:00 – Room: **TBA**

Abstract: With many of the current "SOTA" LLMs being closed sourced and their training data inaccessible, more and more questions arise that relate to potential contamination of the evaluation datasets used to claim their results. Various claims can be found online that range from suspicions of outright training on evaluation data to inflate results to suggestions that the definitions of contamination used may be inadequate and underestimate its impact. However, even with access to the training corpus, contamination and its impact is far from trivial to assess. In this talk, I discuss common ways of measuring contamination and provide empirical data into how much they impact results for a range of LLMs.

Bio: Dieuwke Hupkes is a research scientist at Meta. Among other things, she works on better understanding how (large) language models generalise, what they (don't) understand and what that even means, and more generally on how they can reasonably be evaluated. She is excited about the new opportunities such models bring us and the new scientific challenges that go hand in hand with that.

Keynote Talk
Contamination in Web-Scale Datasets and its Impact on Large Model Evaluations

Jesse Dodge
Allen Institute for AI
2024-08-16 11:00:00 – Room: **TBA**

Abstract: We are at a pivotal moment in the history of AI. The AI research community has driven progress for decades, but over the past couple years industry has started to make significant advances in model capabilities while purposely being closed about how. In this talk I'll start by discussing different types of contamination and how they appear in the wild. I'll then discuss some of our work on building massive datasets by scraping the web, including Dolma and C4. I'll discuss What's In My Big Data, a toolkit for documenting the contents of web-scale datasets, and some of our results on measuring contamination in different ways across a variety of popular pretraining corpora. I'll conclude by discussing evaluation of large models, and how current evaluations have low construct validity and how we don't have strong evaluations for the actual use cases that users care about.

Bio: Jesse Dodge is a Senior Research Scientist at the Allen Institute for AI, on the AllenNLP team, working on natural language processing and machine learning. He is interested in the science of AI and AI for science, and he works on reproducibility and efficiency in AI research. He is involved in many parts of OLMo, a project to create fully open large language models, including creation of Dolma (a web-scale training dataset), Palmoa (an evaluation benchmark for language models), and incorporating ethical principles at every stage of the machine learning pipeline. His research has highlighted the growing computational cost of AI systems, including the environmental impact of AI and inequality in the research community. He has worked extensively on improving transparency in AI research, including open sourcing and documenting datasets, data governance, and measuring bias in data. He has also worked on developing efficient methods, including model compression and improving efficiency of training large language models. His PhD is from the Language Technologies Institute in the School of Computer Science at Carnegie Mellon University. He created the NLP Reproducibility Checklist, which has been used by five main NLP conferences, including EMNLP, NAACL, and ACL, totaling more than 10,000 submissions, he helped create the Responsible NLP Checklist which is used for submissions to ARR (replacing the Reproducibility Checklist), and was an organizer for the ML Reproducibility Challenge 2020-2022. His research has won awards including a Best Student Paper at NAACL 2015 and a ten-year Test of Time award at ACL 2022, and is regularly covered by the press, including by outlets like The New York Times, Nature, MIT Tech Review, Wired, and others.

Keynote Talk
A Sanity Check on Emergent Properties

Anna Rogers
IT University of Copenhagen
2024-08-16 17:00:00 – Room: **TBA**

Abstract: One of the frequent points in the mainstream narrative about large language models is that they have "emergent properties", but there is a lot of disagreement about what that even means. 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: Anna Rogers is tenured associate professor at the Computer Science department at IT University of Copenhagen. She holds a PhD in computational linguistics from the University of Tokyo, followed by postdocs in machine learning for NLP (University of Massachusetts) and social data science (University of Copenhagen). Her research focuses on interpretability, robustness, and sociotechnical aspects of large language models.

Table of Contents

Program

Friday, August 16, 2024

08:55 - 09:00	*Opening Remarks*
09:00 - 09:45	*On the value of carefully measuring data.*
09:45 - 10:30	*Evaluation data contamination:how much is there, and how much does it actually matter?*
10:30 - 11:00	*Break*
11:00 - 11:45	*Contamination in Web-Scale Datasets and its Impact on Large Model Evaluations*
11:45 - 12:00	*Best paper presentation*
12:00 - 13:30	*Lunch Break*
13:30 - 15:30	*Poster Session*
	Evaluating Chinese Large Language Models on Discipline Knowledge Acquisition via Memorization and Robustness Assessment Chuang Liu, Renren Jin, Mark Steedman and Deyi Xiong
	Confounders in Instance Variation for the Analysis of Data Contamination Behzad Mehrbakhsh, Dario Garigliotti, Fernando Martínez-Plumed and Jose Hernandez-Orallo
	A Taxonomy for Data Contamination in Large Language Models Medha Palavalli, Amanda Bertsch and Matthew R. Gormley
15:30 - 16:00	*Break*
16:00 - 16:45	*A Sanity Check on Emergent Properties*
17:00 - 17:15	*Closing Remarks*