Workshop on Meaningful, Efficient, and Robust Evaluation of LLMs (MERE)

1 Workshop Topic and Content

Recent work by the organizers and others have shown two major problems with current large language models evaluations. First, evaluations are brittle, leading to inconsistent and non-reproducible results (Sclar et al., 2023; Mizrahi et al., 2024; Hagmann et al., 2023). Second, they are inefficient, utilizing an unnecessary amount of compute (Perlitz et al., 2023; Polo et al., 2024). These issues stems from several novel evaluation challenges introduced by LLMs. Most notably, since LLMs are often not trained or finetuned for a specific downstream task, there are many seemingly arbitrary design choices that are required in order to construct a zero or few-shot prompt. For example, the phrasing of the task description in natural language, the choice of few-shot examples, their ordering, and more.

In this workshop we seek innovative research relating to the evaluation of large language models (LLMs). This includes, but is not limited to, robust, reproducible and efficient evaluation metrics, as well as new approaches for collecting evaluation data which can help in better differentiating between different LLMs and understanding their current bottlenecks.

LLM evaluation is broken. Recent research has found that the results obtained on popular benchmarks (e.g., BIG-bench (bench authors, 2023), MMLU (Hendrycks et al., 2020), or HELM (Liang et al., 2022)) are brittle, and that seemingly arbitrary prompt variation often lead to drastically different conclusions. Sclar et al. (2023) showed that LLMs are sensitive to minor prompt variations, such as the addition or omission of punctuation marks. They found that such changes may sometimes lead models to perform anywhere between 3% and 80% accuracy on the same test set. Voronov et al. (2024) showed that LLMs are sensitive to how in-context learning examples are formatted (e.g., if examples are separated with a space or a new line), and may lead to opposite ranking of LLMs based on this seemingly arbitrary decision. Finally, Mizrahi et al. (2024) has shown that prompt paraphrases can lead to vastly different results across a wide range of tasks and benchmarks, in both absolute as well as relative performance.

Taken together, since most benchmarks rely on only a single prompt template, these works and other raise serious doubts regarding any automatic evaluation carried on popular benchmarks, as well as on any claim made based on such evaluation.

Unclear why LLMs are so brittle. While the previously mentioned work identifies and quantifies the problems with current evaluation practices, it does not currently explain *why* this happens, nor *how* to solve it. Answering such questions is broad and multifaceted and stands at the core of scientific progress in the field, helping in identifying what are the current limitations in LLMs, and how different LLMs objectively compare to one another. Hence, we think it is important to conduct a community-wide effort soliciting ideas and approaches to achieve a meaningful and robust evaluation of LLMs.

To enable research into better evaluation, we will make available a large dataset with 1B model predictions over popular benchmarks like MMLU and BIG-bench. To facilitate and spur research in this field we will publish a large dataset of 1B model predictions together with prompts and gold standard references. This dataset will go beyond reporting just the accuracy of a model on a given sample, and will also include various axes which identify how the prompt was created and which were found to affect performance (instruction template, few-shot examples, their order, delimiters, etc.), as well as any known information about the model (pretraining corpora, type of instruction-tuning, different checkpoints, and more), and the annotated gold label. We have started collecting predictions to populate this dataset, and have thus far have over 200M predictions across our different sites. We plan to reach 1B instances by the call for papers deadline. Furthermore, this large dataset of model predictions allows us to apply a statistical reliability analysis (Hagmann et al., 2023; Riezler and Hagmann, 2024) that dissects the contribution of different noise sources to overall variance, and assesses LLM performance conditional on different properties of input data. This analysis shall serve as example for submis-

sions on novel evaluation methods for LLMs.

Examples of potential research questions to be addressed in MERE. Below we outline several possible research questions which we will welcome into our workshop, to get a sense of the type of work we would like to solicit. Researchers addressing these directions may benefit from the large 1B predictions dataset we are going to publish, but are not required to use it.

Meaningful evaluation. Do models from the same family exhibit similar prompt sensitivity? How does the instructions fine-tuning or in-context learning influence the evaluation of LLMs? Does prompt sensitivity change during training? Is there a universal prompting strategy that is best across different tasks?

Efficient evaluation. Does quantization or distillation increase prompt format sensitivity? What are efficient ways to collect annotated evaluation data?

Robust evaluation. Are larger models more robust? Are there robust statistical measures that analyze LLMs' sensitivity to prompts? Are there specific types of prompts that consistently cause failure (or success)?

Tagline: *Current LLM evaluation is unreliable. We look for new evaluation approaches, enabled by a new large-scale corpus of 1B model predictions.*

2 Invited Speakers

We have reached out to experts in the field and have got excited responses. Below we provide invited speakers who have tentatively agreed to present at our workshop, if accepted. The presentations will happen in two or three slots for individual invited talks, and a panel discussion of invited speakers.

- Leshem Choshen, Postdoctoral Researcher focusing on evaluation, efficient evaluation and open NLP; MIT and IBM
- Ehud Reiter, Professor; University of Aberdeen
- Christian Hardmeier, Associate Professor; IT University of Copenhagen
- Barbara Plank, Professor; LMU Munich

3 Workshop Size / Prior Events

- The workshop will take place as a one-day event.
- Based on the popularity of the topic we expect between 200-250 attendees.
- We'll support both online and in-person presentations, via the conference-provided platforms (e.g., Underline).

Preferred Venue: We have a slight preference for venues held in Europe for shorter travel for the majority of the organizers, e.g., ACL in Vienna. The later ACL date for call for paper also allows us more time to collect data in our dataset of model predictions which in turn will enable more research.

4 Diversity and Inclusion

Contribution to academic diversity: Conducting research on LLM evaluation is computationally expensive, requiring running LLMs on a vast number of samples to examine statistical trends and behaviors. This cost could potentially bar participation from under-resourced research groups. To address this concern, we put significant effort into creating a large corpus of 1B model predictions. These enable research groups to focus on explaining the data rather than worrying about how to get it. In addition, the workshop promotes a variety of topics related to diversity and fairness including fair evaluation for low-resource groups, reducing energy consumption, and model transperancy.

Diversifying representation: The OC is intentionally diverse across gender, ethnicity, geography, academic and industry backgrounds, as well as institutional affiliations. Likewise, our invited speakers represent a range of disciplines, career stages, and areas of expertise, ensuring a broad spectrum of perspectives.

Diversifying participation: We will promote the workshop through diverse channels, including global mailing lists, social media, ensuring outreach to underrepresented regions and groups in NLP and computational linguistics. We will specifically encourage submissions and participation from early-career researchers, students, and individuals from underrepresented groups, such as women, minorities, and researchers from less-represented geographic regions.

5 Workshop Organizers

- **Ofir Arviv**, IBM Research, ofir.arviv@ibm.com. Ofir is a member

of the LLMs Evaluation group in IBM, focusing on research of robust and efficient evaluation methods.

- **Eliya Habba**, PhD student at the Hebrew University, eliya.habba@mail.huji.ac.il, focuses on the evaluation and robustness of LLMs and the regulatory implications thereof.

- **Rotem Dror**, rdror@is.haifa.ac.il, Assitant professor at Haifa University, Interested in developing meaninful statistical measures for evaluating NLP and ML models. Co-organizer of the NLP4Science Workshop at EMNLP2024. Co-organizer of Data-centric Machine Learning Research (DMLR) Workshop at ICML2024. Organizer and Chair of the Statistical Challenges in Model-based Data Science Session at CFE-CMStatistics 2023. Co-organizer of Evaluation & Comparison of NLP Systems (Eval4NLP) Workshop at AACL202, Co-organizer of Data-centric Machine Learning Research (DMLR) Workshop at ICML2023, Student chair in the Student Research Workshop (SRW) at ACL 2020. Co-organizer of the Israeli Seminar on Computational Linguistics (ISCOL) 2018, 2024.

- **Michael Hagmann**, Heidelberg University, hagmann@cl.uni-heidelberg.de. Postdoctoral researcher in the Department of Computational Linguistics at Heidelberg University, Germany, since 2019. He has worked as a medical statistician at the medical faculty of Heidelberg University in Mannheim, Germany and in the section for Medical Statistics at the Medical University of Vienna, Austria. His research focus is on statistical methods for data science and, recently, NLP.

- **Hannaneh Hajishirzi** Associate Professor Paul G. Allen School of Computer Science and Engineering Adjunct at: UW Electrical and Computer Engineering, Linguistics, University of Washington, Senior Director, AllenNLP. hannaneh@cs.washington.edu, Co-organizer, 3rd Workshop on Knowledge-Augmented NLP, 2024 Co-organizer, 5th Workshop on Representation Learning for NLP (RepL4NLP), 2020 Task Co-organizer, Semeval-2019 task 10: math question answering Workshop at IJCAI 2019, Fourth Workshop on Declarative Learning Based Programming.

- **Itay Itzhak**, PhD student at the Technion and the Hebrew University, interested in understanding the limitations and behavioral biases of language models, and designing robust evaluation methods. Published work exploring the evaluation of model biases and implicitly learned abilities.

- **Yotam Perlitz**, IBM Research AI, advocating for more transparent, roubust and efficient LLM benchmarks, factually correct Data-to-text generation and data-efficient LLM training Previously, investigated efficient methods for objects detection as well as exotic transmission phenomena through organic phases of condensed matter. Co-organized the Navigating the Modern Evaluation Landscape tutorial at LREC-COLING 2024

- **Michal Shmueli-Scheuer**, IBM Research, shmueli@il.ibm.com. Michal is a principal researcher, leading the work of LLMs Evaluation in IBM. She was an organizer of the 1st and 2nd Scientific Document Processing (SDP) workshops at 2020 (EMNLP) and 2021 (COLING), and co-organized shared tasks for Scientific document summarization in those workshops, and the Navigating the Modern Evaluation Landscape tutorial at LREC-COLING 2024.

- **Stefan Riezler**, Heidelberg University, riezler@cl.uni-heidelberg.de. Full professor in the Department of Computational Linguistics at Heidelberg University, Germany since 2010, and also co-opted in Informatics at the Department of Mathematics and Computer Science. His research focus is on interactive machine learning for natural language processing. He co-organized shared tasks at WMT and served as program chair for EACL and CoNLL.

- **Gabriel Stanovsky**, Hebrew University and AI2, gabriel.stanovsky@mail.huji.ac.il. Interested in developing NLP models which deal with real-world texts and help answer multi-disciplinary research questions, e.g., in archaeology, law, medicine, and more. Co-organized the Workshop on Gender Bias in Natural Lan-

guage Processing and the Workshop on Semantic Evaluation (SemEval) both collocated with NAACL 2022, a shared task on Math Question Answering collocated with NAACL 2019, and the Navigating the Modern Evaluation Landscape tutorial at LREC-COLING 2024.

- **Oyvind Tafjord**, AI2, oyvindt@allenai.org. Developing the evaluation system for the OLMo models, interested in how to make LLM evaluations more meaningful and useful when making decisions about their use.

- **Jiangjiang Yang**, AI2, jjyang@allenai.org. Engineered data repository and retrieval systems for scientific knowledge graph (SemanticScholar) and LLM evaluation.

6 Program Committee

We expect to receive 30-50 submissions, and hence aim for roughly that number of reviewers which will each review 3 submissions on average. We are also happy to accept excellent submissions from the ACL rolling review platform.

PCs who have accepted our invitation. Felipe Maia Polo, Roy Schwartz, Ella Rabinovich, Roni Friedman-Melamed, Qinyuan Ye, Robin Jia, Adam Nohejl, Jannis Bulian, Gili Lior, Uri Berger, Yonatan Bitton, Yonatan Belinkov, Ben Bogin, Arie Cattan, Valentina Pyatkin, Paul Roit, Yu Zhao, Fan Bai, Ameya Prabhu, Justin Ray, Ronan Le Bras, Hila Gonen, Vered Shwartz, Yusuke Sakai, Harvey Fu, Taro Watanabe, Noam Dahan, Max Ryabinin, Yanai Elazar, Aviv Slobodkin, Melanie Sclar, Akshita Bhagia, Ori Shapira, Tim Dettmers, Xiaotang Du, Yuval Reif.

Pending Invitations. Shir Ashoury-Tahan, Ariel Gera, Elad Venezian, Vishaal Udandarao, Xiang Ren, Yusuke Sakai, Jiangnan Hang, Hidetaka Kamigaito, Sotiris Anagnostidis, Jan Philip Wahle, Terry Ruas, Yang Xu, Bela Gipp, Michael Hassid, Tom Hope, Anton Voronov, Lena Wolf, Ori Yoran, Tomer Wolfson, Ohad Rubin, Shmuel Amouyal, Shmuel Amouyal, Shmuel Amouyal, Amit Ben Artzi.

References

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Michael Hagmann, Philipp Meier, and Stefan Riezler. 2023. Towards inferential reproducibility of machine learning research. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2023. Efficient benchmarking (of language models). *arXiv preprint arXiv:2308.11696*.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples.

Stefan Riezler and Michael Hagmann. 2024. *Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science*, second edition. Springer.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766*.