Navigating the Modern Evaluation Landscape: Considerations in Benchmarks and Frameworks for Large Language Models (LLMs)

Leshem Choshen[*†], **Ariel Gera**[†], **Yotam Perlitz**[†], **Michal Shmueli-Scheuer**[†], **Gabriel Stanovsky**[◇]

[*]MIT, [†]IBM Research, [◇]The Hebrew University of Jerusalem

Abstract

General-Purpose language models have changed the world of natural language processing, if not the world itself. The evaluation of such versatile models, while supposedly similar to evaluation of generation models before them, in fact presents a host of new evaluation challenges and opportunities. This tutorial welcomes people from diverse backgrounds and assumes little familiarity with metrics, datasets, prompts and benchmarks. It will lay the foundations and explain the basics and their importance, while touching on the major points and breakthroughs of the recent era of evaluation. We will contrast new to old approaches, from evaluating on multi-task benchmarks rather than on dedicated datasets to efficiency constraints, and from testing stability and prompts on in-context learning to using the models themselves as evaluation metrics. Finally, we will present a host of open research questions in the field of robsut, efficient, and reliable evaluation.

Keywords: Language models, Benchmarks, efficient evaluation, language model as metrics,

1. Tutorial Description - Introduction

1.1. Background and Goals

Evaluation benchmarks have been a cornerstone of machine learning progress for years now. However, the introduction of pretrained models has profoundly altered the way benchmarks are used. Instead of focused questions, benchmarks now require assessing a vast and general set of abilities, for which diverse samples are collected (Liang et al., 2022; Gao et al., 2021). This is a first of many changes that are transforming the field of model evaluation, and that entail increasingly complex evaluation endeavours, compared to traditional single-task evaluation efforts.

On the other hand, the new era offers advantages in evaluation, requiring less data for training and better, flexible metrics. Evaluation is no longer done through fine-tuning, i.e. training on a train set for every task to be evaluated, but relies entirely on zero-shot or in-context learning. In that manner, instead of supplying training, the benchmark is a test set only. Another advantage of current models is that they can serve to evaluate other models, following the assumption that error detection is easier than generation. This approach offers a way to test answers in areas where it was hardly possible before.

With all of those changes, also comes great compute. Evaluating on a broad range of datasets, with more models, and with long and complex tasks, all brought growing compute needs, sometimes more costly than the model pretraining (Biderman et al., 2023).

This tutorial aims to introduce the still relevant concepts of evaluation (e.g., evaluation goals or N-gram based reference metrics) and contrast those with the new and changing needs of the general models we employ today. Such needs include leveraging another language model as an evaluator, a language model based metric, taking inference costs into account, evaluating each model on a diverse set of tasks, evaluating on diverse prompts, and more.

A complementary goal of the tutorial is to provide a structured and organized view of LLMs' benchmarking. Such a view is largely missing in the academic literature, where each paper typically addresses a specific problem in isolation, normally in an ad-hoc manner. This view is also missing from the practical solutions presented by the industry, where different decisions are taken without a proper explanation which might cause some vague or incomplete understanding by the community. We present a complete pipeline of LLMs benchmarking, and discuss decisions that need to be considered throughout the pipeline. We will also share our experience and lessons learned from evaluating LLMs. Finally, the tutorial will discuss future challenges of LLMs benchmarking.

1.2. Tutorial type

This is a *cutting-edge* tutorial that aims at bridging the gaps in this emerging field. The need for timely discussions of LLM benchmarking is ever more pressing in light of the rapid advancement in the field that has caused great shifts in benchmarking such as new evaluation paradigms (e.g., ICL), and ever growing benchmarks aiming to validate unprecedented amounts of new abilities. Specifically,

LREC-COLING 2024 Tutorials, pages 19–25

this tutorial differs from recent performance benchmarking tutorials (Coleman et al., 2019) that mainly deal with evaluations of training and inference performance for hardware, software, and services as opposed to our focus on quality. Others like (Boyd-Graber et al., 2022) focus on human evaluation and explainability of LLMs or NLG metrics (Khapra and Sai, 2021) which covers a small section of overall benchmarking considerations.

2. Target Audience

While the tutorial will present the current state of the art and cutting-edge research, it should accommodate entry-level audience. The tutorial assumes little to no knowledge about evaluation, merely expecting some understanding of what Language Models are currently capable of and why they are useful. Thus, the tutorial is the best fit for people who have worked on a specific aspect of evaluation, but are less familiar with the big picture, researchers who are new to evaluation, and researchers who are less familiar with new challenges specific to large language models, such as benchmarking across many datasets, evaluating in open-domain tasks and prompting.

3. Outline

Part 1: Introduction (35 min)
Part 1.1: Introduction to Benchmarking

- What are the goals of model evaluation?
- Benchmarking building blocks- task, dataset, and metric

Part 1.2: Introduction to LLM Benchmarking

- Models: what do we evaluate?
- What are the main challenges? or, why it is not trivial?
- Common and important tasks
- Measurements - automatic metrics and human evaluation
- Benchmarking paradigms - fine-tuning, zero shot learner, few shot leaner
- Other important hyperparameters, instructions, prompts matter
- Reviewing general benchmarks
- Reviewing specific downstream tasks
- How do objectives and considerations (what, when, and whom) affect benchmarking decisions?

Part 2: Framework for Benchmarking (10 min)

- What are the requirements from the framework?
- Open source frameworks (e.g., HELM, OpenAI Evals, LM-evaluation-harness)
- Business frameworks

Part 3: Metrics (45 min)

- Classic N-gram based metrics
- Language Model based metrics
- Reference-less Metrics
- Language models as evaluators
- Fine-grained and specialized metrics
- Challenge sets, perturbation and data-based metrics

Part 4: Prompts (45 min)

- The importance of prompts
  - Who writes the prompts? What goals do they serve?
- Overview of evaluation protocol for prompts
  - Typically, a single prompt is used to evaluate across models
- Prompt banks
- Different desiderata for different use-cases
  - LLM developers
  - Developers for targeted downstream applications
  - Developers of open-ended user-facing aplciations

Part 5: Efficient Benchmark Design (45 min)

- Benchmarks Objectives
- Benchmarks Compute (survey)
- Benchmark decisions, or, common ways to reduce compute (survey)
- What makes a good benchmark (validity, reliability)
- Best practices for compute reduction in LLM benchmarks

Part 6: Manual Evaluation Efforts (30 min)

- Is human evaluation being abandoned?
- The alignment paradigm
- LLM-Human feedback loops

4. Diversity Considerations

The tutorial promotes a variety of topics related to diversity and fairness including efficient benchmarking to enable fair evaluation for low-resource groups, and reducing energy consumption. In addition, some of the topics are directly related to increasing transparency around model evaluation.

The presenters are diverse in terms of gender, age, background, location and affiliation.

5. Reading List

1. Surveys on evaluation of LLMs (Chang et al., 2023; Ziyu et al., 2023; Gehrmann et al., 2023)
2. Pre-training paradigms (Min et al., 2023)
3. Current benchmarks: HELM (Liang et al., 2022), big-bench (Srivastava et al., 2022), LM-evaluation-harness (Gao et al., 2021)
4. Prompts: creating paraphrases (Lester et al., 2021; Gonen et al., 2022; Honovich et al., 2022), robustness to paraphrases (Gu et al., 2022; Sun et al., 2023; Mizrahi et al., 2024)
5. Metrics: survey (Sai et al., 2022), models as evaluators (Zheng et al., 2023)
6. Efficient-benchmarking: (Perlitz et al., 2023a; Vivek et al., 2023; Liang et al., 2022),
7. Manual Evaluation: survey (Bojic et al., 2023), reproducibility (Belz et al., 2023)

6. Presenters

Leshem Choshen

leshem.choshen@mail.huji.ac.il

Leshem Choshen is a postdoctoral researcher at MIT/IBM, aiming to collaboratively pretrain through model recycling (Don-Yehiya et al., 2022b; Yadav et al., 2023), efficient evaluation (Choshen et al., 2022b; Perlitz et al., 2023a), and manageable pretraining research (e.g., co-organizing the babyLM shared task (Warstadt et al., 2023)). Before leading a small research group at IBM, he received the postdoctoral Rothschild and Fulbright fellowships as well as IAAI and Blavatnik best Ph.D. awards. With broad NLP and ML interests, he also worked on Reinforcement Learning, and Understanding of how neural networks learn (Choshen et al., 2022a; Din et al., 2023), with a specific interest in evaluation (Choshen and Abend, 2019; Choshen et al., 2020), evaluation of evaluation (Choshen and Abend, 2018b,a), reference-less metrics (Choshen and Abend, 2018c; Honovich et al., 2021), quality estimation (Don-Yehiya et al., 2022a) and related topics. In parallel, he participated in Project Debater, creating a machine that could hold a formal debate, ending in a Nature cover and live debate (Slonim et al., 2021).

Ariel Gera

ariel.gera1@ibm.com

Ariel is a research scientist at IBM Research AI, with diverse interests in both NLG and text classification. Ariel is currently pursuing research on utilizing the outputs of different model layers (Gera et al., 2023) and on efficient and reliable evaluation for NLG tasks. Following his research on argumentation (Bilu et al., 2019) as part of Project Debater (Slonim et al., 2021), he has worked on numerous threads related to training models with limited supervision. These include studies of active learning (Ein-Dor et al., 2020; Perlitz et al., 2023c), few-shot (Shnarch et al., 2022a) and zero-shot (Gera et al., 2022), as well as development of the Label Sleuth platform for building text classifiers with a human in the loop (Shnarch et al., 2022b). Ariel has an MSc in Cognitive Science from the Hebrew University, for psychological studies of emotion perception.

Yotam Perlitz

yotam.perlitz@ibm.com

Yotam Perlitz is an AI Research scientist at IBM Research AI, advocating for more transparent and efficient LLM benchmarks (Perlitz et al., 2023a; Bandel et al., 2024), factually correct Data-to-text generation (Perlitz et al., 2023b, 2022) and data-efficient LLM training (Gera et al., 2022; Perlitz et al., 2023c). Previously, Yotam had investigated coarse to fine methods for objects detection (Dana et al., 2021) as well as exotic transmission phenomena through various phases of matter (Perlitz and Michaeli, 2018) as part of his M.Sc at the Weizmann institute of Science.

Michal Shmueli-Scheuer

shmueli@il.ibm.com

Michal is a principal researcher in the Language and Retrieval research group in IBM Research AI. Her area of expertise is in the fields of NLG and NLP including data to text, conversational bots, summarization of scientific documents, and affective computing. Michal is leading the work of LLMs Evaluation in IBM. She has published in leading NLP and AI conferences and journals, including ACL, EMNLP, NAACL, AAAI, and IUI. She regularly reviews for top NLP and AI conferences. 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. Michal received her PhD from the University of California, Irvine in 2009.

Gabriel Stanovsky
gabriel.stanovsky@mail.huji.ac.il
Gabriel Stanovsky is a senior lecturer (assistant professor) in the school of computer science and engineering at the Hebrew University of Jerusalem, and a research scientist at the Allen Institute for AI (AI2). He did his postdoctoral research at the University of Washington and AI2 in Seattle, working with Prof. Luke Zettlemoyer and Prof. Noah Smith, and his PhD with Prof. Ido Dagan at Bar-Ilan University. He is interested in developing natural language processing models which deal with real-world texts and help answer multi-disciplinary research questions, in archaeology, law, medicine, and more. His work has received awards at top-tier venues, including ACL, NAACL, and CoNLL, and recognition in popular journals such as Science and New Scientist, and The New York Times.

7. Ethics Statement

During the tutorial, we will emphasize the importance of being aware of and addressing biases in benchmarks and frameworks. We will advocate for transparency in benchmark creation and evaluation methodologies. In addition, we will acknowledge the environmental impact of large-scale models by discussing efficient benchmarking approaches. Finally, we will highlight the importance of community engagement and collaboration for the benefit of diverse perspectives and the benefit of science.

References

Elron Bandel, Yotam Perlitz, Elad Venezian, Roni Friedman-Melamed, Ofir Arviv, Matan Orbach, Shachar Don-Yehyia, Dafna Sheinwald, Ariel Gera, Leshem Choshen, et al. 2024. Unitxt: Flexible, shareable and reusable data preparation and evaluation for generative ai. *arXiv preprint arXiv:2401.14019*.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raf. 2023. Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*.

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 2019. Argument invention from first principles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1013–1026, Florence, Italy. Association for Computational Linguistics.

Iva Bojic, Jessica Chen, Si Yuan Chang, Qi Chwen Ong, Shafiq Joty, and Josip Car. 2023. Hierarchical evaluation framework: Best practices for human evaluation. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems (HumEval)*.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Leshem Choshen and Omri Abend. 2018a. Automatic metric validation for grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1372–1382, Melbourne, Australia. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2018b. Inherent biases in reference-based evaluation for grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Melbourne, Australia. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2018c. Reference-less measure of faithfulness for grammatical error correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 124–129, New Orleans, Louisiana. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2019. Automatically extracting challenge sets for non-local phenomena in neural machine translation. In

Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 291–303, Hong Kong, China. Association for Computational Linguistics.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022a. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Leshem Choshen, Dmitry Nikolaev, Yevgeni Berzak, and Omri Abend. 2020. Classifying syntactic errors in learner language. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 97–107, Online. Association for Computational Linguistics.

Leshem Choshen, Elad Venezian, Shachar Don-Yehia, Noam Slonim, and Yoav Katz. 2022b. Where to start? analyzing the potential value of intermediate models. *arXiv preprint arXiv:2211.00107*.

Cody Coleman, Wen mei Hwu, Gennady Pekhimenko Vijay Janapa Reddi, Carole-Jean Wu, and Jinjun Xiong. 2019. mlperf-bench: Benchmarking deep learning systems. Tutorial,IEEE International Symposium on Performance Analysis of Systems and Software.

Alexandra Dana, Maor Shutman, Yotam Perlitz, Ran Vitek, Tomer Peleg, and Roy J Jevnisek. 2021. You better look twice: a new perspective for designing accurate detectors with reduced computations.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2023. Jump to conclusions: Short-cutting transformers with linear transformations. *ArXiv*, abs/2303.09435.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2022a. PreQuEL: Quality estimation of machine translation outputs in advance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11183, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2022b. Cold fusion: Collaborative descent for distributed multitask finetuning. *ArXiv*, abs/2212.01378.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for bert: An empirical study. In *EMNLP*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Ariel Gera, Roni Friedman, Ofir Arviv, Chulaka Gunasekara, Benjamin Sznajder, Noam Slonim, and Eyal Shnarch. 2023. The benefits of bad advice: Autocontrastive decoding across model layers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10406–10420, Toronto, Canada. Association for Computational Linguistics.

Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1119, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Jiasheng Gu, Hanzi Xu, Liangyu Nie, and Wenpeng Yin. 2022. Robustness of learning from task instructions. *arXiv preprint arXiv:2212.03813*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Mitesh M Khapra and Ananya B Sai. 2021. A tutorial on evaluation metrics used in natural language generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 15–19.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

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*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

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. 2023a. Efficient benchmarking (of language models). *ArXiv*, abs/2308.11696.

Yotam Perlitz, Liat Ein-Dor, Dafna Sheinwald, Noam Slonim, and Michal Shmueli-Scheuer. 2023b. Diversity enhanced table-to-text generation via type control.

Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein-Dor. 2023c. Active learning for natural language generation.

Yotam Perlitz and Karen Michaeli. 2018. Helical liquid in carbon nanotubes wrapped with DNA molecules. *Physical Review B*, 98(19).

Yotam Perlitz, Dafna Sheinwald, Noam Slonim, and Michal Shmueli-Scheuer. 2022. nbiig: A neural bi insights generation system for table reporting.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2022a. Cluster & tune: Boost cold start performance in text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7639–7653, Dublin, Ireland. Association for Computational Linguistics.

Eyal Shnarch, Alon Halfon, Ariel Gera, Marina Danilevsky, Yannis Katsis, Leshem Choshen, Martin Santillan Cooper, Dina Epelboim, Zheng Zhang, and Dakuo Wang. 2022b. Label sleuth: From unlabeled text to a classifier in a few hours. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 159–168, Abu Dhabi, UAE. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nature*, 591:379 – 384.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2023. Anchor points: Benchmarking models with much fewer examples. *ArXiv*, abs/2309.08638.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Ethan Wilcox, Williams Adina, Chengxu Zhuang, Linzen Tal, and Ryan Cotrerell. 2023. Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In

Proceedings of the BabyLM Challenge. Association for Computational Linguistics (ACL).

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Resolving interference when merging models. *ArXiv*, abs/2306.01708.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109.