A Case for Centaur Evaluations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Benchmarks and evaluations are central to machine learning methodology and direct research in the field. Current evaluations commonly test systems in the absence of humans. This position paper argues that the machine learning community should increasingly use centaur evaluations, in which humans and AI jointly solve tasks. Centaur Evaluations refocus machine learning development toward human augmentation instead of human replacement, they allow for direct evaluation of human-centered desiderata, such as interpretability and helpfulness, and they can be more challenging and realistic than existing evaluations. By shifting the focus from automation toward collaboration between humans and AI, centaur evaluations can drive progress toward more effective and human-augmenting machine learning systems.

1 Introduction

Benchmarks and evaluations are central to machine learning methodology and direct machine learning research [Sculley et al., 2018]. As machine learning systems expand into many parts of society, broader impacts of evaluations become important. This position paper is concerned with how (or *how not*) AI system evaluation incorporates humans. **We argue that there should be more and more systematic centaur evaluations, in which humans and AI solve a task cooperatively.**

The progress of language models and their evaluation has been particularly rapid, leading to many new evaluation datasets in question-answer format [Hendrycks et al., 2021a, Wang et al., 2019, 2018, Chollet et al., 2024, Srivastava et al., 2023, Suzgun et al., 2023, Rein et al., 2024, Hendrycks et al., 2021b, Chen et al., 2021, Dua et al., 2019, Glazer et al., 2024, Chan et al., 2024] and interactive environments [Xie et al., 2024, Majumder et al., 2024, Deng et al., 2023, Zhou et al., 2024, Drouin et al., 2024]. Very few exceptions are *centaur evaluations* [Lee et al., 2024, Wijk et al., 2024, Shao et al., 2025] which include humans in the evaluation process.

There are several explanations for why centaur evaluations are relatively rare. One lies in the history and culture of the field of machine learning, from the Turing Test to Imagenet, which are based on the idea of imitating a human activity with a machine learning model. Even beyond cultural reasons, there are clear incentives to evaluate for human imitation. Not only are such evaluations straightforward to formalize as supervised learning problems, but they are also comparably cheap: humans provide ample training data in the behavior being imitated. Finally, results are easy to communicate to the public, as most people have engaged in the behavior that systems are trained and evaluated to imitate, or at least know what it means to take a mathematics test in school, or a law school exam.

We argue for the benefits of centaur evaluations in three arguments. First, centaur evaluations expand which capabilities of machines we can evaluate, in particular those involving human perception and dexterity (Section 3.1): “It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility.” (Moravec [1990], p.15) Centaur evaluations

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

might lead us away from evaluating AI with exams [Metz, 2025] and toward evaluations that more closely resemble human use of machine learning systems.

Our second argument for centaur evaluations is that they allow to *directly* evaluate human-centered features of machine learning models, such as interpretability [Casper et al., 2023], complementarity [Donahue et al., 2022], helpfulness [Bai et al., 2022], and the ability to ask follow-up questions [Li et al., 2023, Shaikh et al., 2024] (Section 3.2). This is in contrast to current evaluation methodologies, which require imperfect proxies for these desiderata.

Finally, and for us most importantly, centaur evaluations can re-center machine learning practice toward human augmentation and away from a destructive path of human replacement, leaving some without economic power and wealth and others with high amounts of both (Section 3.3). Several economists call for technical change that focuses on human augmentation rather than replacement Acemoglu and Johnson [2023a], Brynjolfsson [2022], Brynjolfsson and McAfee [2011], but there is limited translation of these aspirations into engineering practice. We aim to provide a definition and arguments for centaur benchmarks as such an intervention into engineering practice.

2 Defining Centaur Evaluations

We first define what centaur evaluations are; compare Lee et al. [2024], Shao et al. [2025] for other formalizations. We use the term *Centaur Evaluations* in the memory of centaur chess (also known as *advanced chess* or *freestyle chess*), in which humans use chess computers in their play [Sollinger, 2018]. This means direct involvement of humans in the testing process, not indirect process through labeling of evaluation datasets.

A centaur benchmark for a machine learning system consists of three components:

Human A selection criterion for the human(s) involved in the evaluation, potentially allowing the model to be tested to train humans together with their model (“bring-your-own-human”) or from a distribution of humans, e.g., crowd workers.

Interface A set of actions that the machine learning system and the human can take to interact through an interface, the representation of this interface to the human and the format of submission of answers.

Scoring Scoring of submissions, which can be done through objective means or by a human preference [Chiang et al., 2024], only based on outcomes or also including process. It can also capture the resources, e.g., in terms of computation and human time, expended during the evaluation.

A fourth (optional) component is a way to communicate **transcripts**. For many cooperative tasks, *how* centaurs achieved a high score in a benchmark is helpful to improve machine learning systems, and train human collaborators.

In principle, there are two types of centaur evaluations. The first is raising the restriction of current evaluation practice that it must not involve humans. We call these *centaurized evaluations*. Consider, for example, the Massive Multitask Language Understanding benchmark (MMLU) Hendrycks et al. [2021a] without the requirement that no human should be involved in the solution of the task. MMLU prompts are provided to a human with given requirements (human). The human and AI can interact sequentially in a chat interface, and the human submits the outcome (interface). Correct responses are recorded, subject to costs or limitations on the amount of tokens and/or human time used (scoring). The transcripts of interactions can be recorded, e.g., as a screen capture (transcript).

Other evaluations are specifically designed with the additional affordances of centaur evaluations in mind. An example is an evaluation inspired by the paper Brynjolfsson et al. [2025] studying call center workers’ use of chatbots. A call center agent (human) interacts with a chatbot to help a client with a request. The agent and the LLM agent interact by chat (interaction). Satisfaction, time, and the number of tokens generated constitute the score (scoring). Finally, a transcript is shared to train call center agents and improve the chat bot (transcript).

2.1 Existing Centaur Evaluations

There are a few examples of centaur evaluations in the literature. Peng et al. [2023] find a high increase in speed in coding a functional HTTP server of a centaur compared to a machine learning model and a human alone. The paper Mozannar et al. [2024a] studies a random assignment of coders using machine learning-powered coding recommendations in Visual Studio Code, also finding high speed-ups, as do Peng et al. [2023]. Cui et al. [2024] studies in a randomized controlled trial the impact of equipping humans with a machine learning system for support and find large productivity increases. Barke et al. [2023], Mozannar et al. [2024b] analyze the micro-structure of the interaction of humans and machine learning systems. Shao et al. [2025] proposes an interface for interactions in centaur evaluations, using *collaborative agents* instead of our notion of centaurs. They implement an asynchronous computation and communication handler with an interface similar to OpenAI's Gym [Brockman et al., 2016]. Lee et al. [2024] conduct several centaur evaluations with crowdworkers in tasks of collaborative writing, summarization, and puzzles. While these are benchmarks, none of them is regularly reported for frontier models.

2.2 Centaur Evaluations as a Gold Standard

We argue that systematic centaur evaluations are beneficial. However, in many settings, this gold standard might be prohibitively expensive. In these cases, evaluation designers should be explicit about which centaur a benchmark aims to approximate, and test calibration. *Synthetic centaur evaluations* approximate centaur evaluation using interactive evaluations [Park et al., 2023, Aher et al., 2023] or even train tools in simulation [Wu et al., 2025].

3 Why There Should Be More Centaur Evaluations

We now make our case for centaur evaluations. First, centaur evaluations allow to evaluate AI more thoroughly (Section 3.1), they allow direct testing of human-centered desiderata like interpretability, human-augmentation, helpfulness, and grounding (Section 3.2), and, for us most importantly, re-center technological development toward human augmentation, while helping policymakers (Section 3.3).

3.1 Centaur Evaluations Can Be Harder

Current evaluations "saturate" fast. That is, AI models rapidly achieve very good results on evaluations, leading to concerns that soon, humans might not be able to evaluate models [Arc Prize, 2025, Metz, 2025]. We contend that this worry might be a consequence of how restrictive current evaluation formats are rather than a general limitations of humans in evaluating machine learning systems. Additionally, while most imitative evaluations might soon saturated, benchmark results may not transfer to real-world tasks because much of the hardness of operation in the real world stems from complex feedback loops and heterogeneity that only comes out in interaction with humans. Hence, while we laud more complex, realistic, and interactive evaluations (e.g., Xie et al. [2024], Majumder et al. [2024], Deng et al. [2023], Zhou et al. [2024], Drouin et al. [2024], Lee et al. [2024], Shao et al. [2025], Wijk et al. [2024]), there are strong reasons to consider centaur evaluations for harder and more realistic evaluations.

One way in which centaur evaluations can be harder is mechanistic: Humans have more actions and more sensors available than even the most powerful multimodal models. Consider a call center benchmark. Human raters are still often able to distinguish whether they are talking to an AI or a human and will rate AI differently. In this case, a human replacement evaluation will have limited success unless the auditive Turing test is passed, and we can replace most call center workers altogether (more on this in Section 3.3). Similarly, many security-critical actions are exclusive to humans, which likely will persist into the future. Evaluating interactions with safety-critical systems requires evaluating a centaur. In contrast to a call center or a security-relevant setting, current evaluations look synthetic: school-level [Hendrycks et al., 2021b] and researcher-level mathematics [Glazer et al., 2024], general knowledge questions [Hendrycks et al., 2021a], and reading comprehension [Dua et al., 2019], among others. What they do have in common is that they have text as input, text as output, and a correct answer. The format of evaluations is restrictive and makes it hard for humans to create truly hard evaluations.

3.2 Centaur Evaluations Simplify the Evaluation of Human-Centered Desiderata

Centaur evaluations also simplify the evaluation of human-centered desiderata such as explainability, interpretability, helpfulness, or grounding. One such desideratum, *explainability*, has received attention in policy for example in the European Union's AI Act (European Union [2024], Art. 13, compare also Art. 52): "High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent *to enable deployers* to interpret a system's output and use it appropriately." (emphasis added). Explainability is measured with explicit reference to humans, in this case, deployers. On the other hand, much of explainability evaluation uses proxies of explainability or mechanistic techniques, compare Casper et al. [2023]. With centaur evaluations, explainability can be directly evaluated as the ability of a human to act correctly based on system outputs.

Additionally, current evaluations cloak achievements in human-centered development technology. One concrete example is the learning-to-defer literature, which studies when a machine learning system should defer to a human for a decision (see Bansal et al. [2021] for a theoretical model, and compare Yang et al. [2018], Okati et al. [2021], Mozannar and Sontag [2021], Madras et al. [2018], Keswani et al. [2022], Vodrahalli et al. [2022], Bansal et al. [2021], De et al. [2021]). In current evaluations that do not consider human-AI interplay, learning-to-defer is irrelevant. Successful deferral helps in real-world use, but current evaluations are blind to it.

3.3 Reporting Relevant Artifacts

Finally, centaur evaluations re-center the direction of progress in machine learning and can help decision-makers decide where to steer technological development.

Technology and automation play an important role in the inequality of power and wealth [Karabarbounis and Neiman, 2014, Autor, 2019]. One of the main channels through which inequality arises is that capital (so any non-human input to production) becomes more important and is owned by a smaller group than a few decades ago [Alvaredo et al., 2022]. We believe that keeping humans productive (as we formalize in this subsection) is important for machine learning development.

To define human augmentation and human replacement precisely, we use notation from macroeconomics (but the following should be self-contained. In this notation, K denotes *capital*, or the material means of production, L or *labor* is the human input, Y or *output* is the performance on a task, often measured in monetary terms. $f\colon (K, L) \mapsto Y$ is commonly called a production function. (We refer the interested reader to Romer, David [2018] for more macroeconomic modeling.) We will view model i's performance on a centaur evaluation (including human, interface, and scoring components) through the lens of triples (i, K, L, Y) where K denotes the amount of compute, L the amount of time a human time spent, and Y the performance on an economically relevant task. Fitting a function, we obtain the evaluation's *centaur production function*

$$Y = f_i(K, L).$$

Definition 3.1. We call a machine learning system i with centaur production function f_i *human-augmenting* if the marginal value of a human minute $\frac{\partial f_i}{\partial L} \gg 0$ for relevant values K and L. If the marginal value of a human minute is approximately zero, $\frac{\partial f_i}{\partial L} \approx 0$, for relevant values K and L, we call it *human-replacing*.

Human augmenting technologies are more likely to produce high wages and sustain economic bargaining power for those who do not own capital, as supported by economists [Acemoglu and Johnson, 2023b,a, Brynjolfsson, 2022]. Even institutions at the center of technological disruption call for ways to increase the number of jobs, see Y Combinator's open letter Combinator [2024].

Centaur evaluations allow us to produce evaluations with direct meaning for human augmentation and impacts for the value of human time. In addition to human augmentation, we could evaluate $f_i(K, L)$: task achievement, fixed resources in terms of both human and compute (compare Coleman et al. [2017] for resource-controlled computing). Or we could evaluate $\max_{K,L} f_i(K, L)$: maximal task achievement. Current evaluations, in contrast, are blind to human augmentation, as they evaluate $f_i(K, 0)$ (total task achievement absent humans under limited compute budget) or $\max_K f_i(K, 0)$ (total task achievement absent humans under limited compute budget). If the goal is to succeed in current evaluations, there are no incentives for human augmentation.

References

D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor, 2018. URL `https://openreview.net/forum?id=rJWF0Fywf`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://aclanthology.org/W18-5446/`.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc Prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=uyTL5Bvosj`. Featured Certification.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *ACL (Findings)*, pages 13003–13051, 2023. URL `https://doi.org/10.18653/v1/2023.findings-acl.824`.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=Ti67584b98`.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021b. URL `https://openreview.net/forum?id=7Bywt2mQsCe`.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL `https://aclanthology.org/N19-1246/`.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024. URL `https://arxiv.org/abs/2411.04872`.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. Mle-bench: Evaluating machine learning agents on machine learning engineering, 2024. URL `https://arxiv.org/abs/2410.07095`.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. TravelPlanner: a benchmark for real-world planning with language agents. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models, 2024. URL `https://arxiv.org/abs/2407.01725`.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=kiYqbO3wqw`.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=oKn9c6ytLx`.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. WorkArena: How capable are web agents at solving common knowledge work tasks? In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11642–11662. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/drouin24a.html`.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. Evaluating human-language model interaction, 2024. URL `https://arxiv.org/abs/2212.09746`.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, et al. Re-bench: Evaluating frontier AI r&d capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration, 2025. URL `https://arxiv.org/abs/2412.15701`.

Hans P Moravec. *Mind children*. Harvard University Press, London, England, July 1990.

Cade Metz. A.i. poses humanity's "last exam." are we ready? *The New York Times*, January 2025. URL https://www.nytimes.com/2025/01/23/technology/ai-test-humanitys-last-exam.html.

Stephen Casper, Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang, and Dylan Hadfield-Menell. Benchmarking interpretability tools for deep neural networks. *arXiv preprint arXiv:2302.10894*, 4, 2023.

Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness, 2022. URL https://arxiv.org/abs/2202.08821.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models, 2023. URL https://arxiv.org/abs/2310.11589.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. Grounding gaps in language model generations. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.348. URL https://aclanthology.org/2024.naacl-long.348/.

Daron Acemoglu and Simon Johnson. Rebalancing ai. *International Monetary Fund*, 2023a.

Erik Brynjolfsson. The Turing Trap: The promise & peril of human-like artificial intelligence. *Daedalus*, 151(2):272–287, 05 2022. ISSN 0011-5266. doi: 10.1162/daed_a_01915. URL https://doi.org/10.1162/daed_a_01915.

Erik Brynjolfsson and Andrew McAfee. *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Brynjolfsson and McAfee, 2011.

Marc Sollinger. Garry kasparov and the game of artificial intelligence, 2018. URL https://theworld.org/stories/2018/01/05/garry-kasparov-and-game-artificial-intelligence. Last accessed: January 22, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 02 2025. ISSN 0033-5533. doi: 10.1093/qje/qjae044. URL https://doi.org/10.1093/qje/qjae044.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The impact of AI on developer productivity: Evidence from github copilot, 2023. URL https://arxiv.org/abs/2302.06590.

Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, Manish Nagireddy, Prasanna Sattigeri, Ameet Talwalkar, and David Sontag. The RealHumanEval: Evaluating Large Language Models' abilities to support programmers, 2024a. URL https://arxiv.org/abs/2404.02806.

Kevin Zheyuan Cui, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. The Productivity Effects of Generative AI: Evidence from a Field Experiment with GitHub Copilot. *An MIT Exploration of Generative AI*, mar 27 2024. https://mit-genai.pubpub.org/pub/v5iixksv.

Shraddha Barke, Michael B. James, and Nadia Polikarpova. Grounded copilot: How programmers interact with code-generating models. *Proc. ACM Program. Lang.*, 7(OOPSLA1), April 2023. doi: 10.1145/3586030. URL `https://doi.org/10.1145/3586030`.

Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. Reading between the lines: Modeling user behavior and costs in ai-assisted programming. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3641936. URL `https://doi.org/10.1145/3613904.3641936`.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym, 2016. URL `https://arxiv.org/abs/1606.01540`.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL `https://doi.org/10.1145/3586183.3606763`.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. CollabLLM: From passive responders to active collaborators. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=DmH4HHVb3y`.

Arc Prize. OpenAI's GPT-4o and the next frontier in AI research, 2025. URL `https://arcprize.org/blog/oai-o3-pub-breakthrough`. Accessed: 2025-01-29.

European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending various regulations and directives (Artificial Intelligence Act). Official Journal of the European Union, 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`. Accessed: 2025-01-29.

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414, 2021.

Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 designing interactive systems conference*, pages 585–596, 2018.

Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage, 2021. URL `https://arxiv.org/abs/2103.08902`.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert, 2021. URL `https://arxiv.org/abs/2006.01862`.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer, 2018. URL `https://arxiv.org/abs/1711.06664`.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Designing closed human-in-the-loop deferral pipelines, 2022. URL `https://arxiv.org/abs/2202.04718`.

Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of Human-AI interactions, 2022. URL `https://arxiv.org/abs/2107.07015`.

Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez Rodriguez. Classification under human assistance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):5905–5913, May 2021. doi: 10.1609/aaai.v35i7.16738. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16738`.

Loukas Karabarbounis and Brent Neiman. The global decline of the labor share. *The Quarterly journal of economics*, 129(1):61–103, 2014.

David H Autor. Work of the past, work of the future. In *AEA Papers and Proceedings*, volume 109, pages 1–32. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2019.

Facundo Alvaredo, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. *World Inequality Report 2022*. World Inequality Lab, 2022. URL `https://wir2022.wid.world/`.

Romer, David. *Advanced Macroeconomics*. McGraw-Hill Education, Columbus, OH, 5 edition, November 2018.

Daron Acemoglu and Simon Johnson. *Power and progress*. PublicAffairs, May 2023b.

Y Combinator. One million jobs 2.0, 2024. URL `https://www.youtube.com/watch?v=BAeBkS2gBpo`. Accessed: 2025-01-22.

Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Dawnbench: An end-to-end deep learning benchmark and competition. *Training*, 100(101):102, 2017.