Position: AI Should Not Be An Imitation Game: Centaur Evaluations

Andreas Haupt [1] **Erik Brynjolfsson** [1]

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 is the history of the field, from the Turing Test to Imagenet, which we discuss in Section 3. Another one lies in the cost and challenges in making centaur evaluations repdroducible, which we discuss in Sections 5 and 6.

We argue that increasing the amount of centaur evaluation in machine learning will benefit society with three arguments. First, centaur evaluations expand which capabilities of machines we can evaluate, in particular those involving human perception and dexterity (Section 4.1): "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 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 4.2). 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 4.3). There are clear incentives for imitation. Imitation-based evaluations are straightforward to formalize as supervised learning problems, humans provide ample training data in the behavior being imitated, and results are easy to communicate to the public, as most people have engaged in the behavior that systems are trained and evaluated to imitate.

Evaluation based on imitation, in turn, leads to incentives

[1]Digital Economy Lab, Stanford University, CA, USA. Correspondence to: Andreas Haupt <h4upt@stanford.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

for human replacement instead of human augmentation, which has led economists to call for human augmentation Acemoglu & Johnson (2023b); Brynjolfsson (2022); Brynjolfsson & McAfee (2011). Brynjolfsson (2022) introduces the *Turing Trap*, which is a possible future where humanity has created technologies that replace humans and leave them without economic and political power. The Turing Trap highlights the dangers of focusing too narrowly on AI systems that imitate human intelligence rather than augmenting it.

The argument in this position paper is structured as follows. We set the stage by defining centaur evaluations (Section 2). We then trace historical reasons for why, culturally, the machine learning community may engage in fewer centaur evaluations than other fields of computer science (Section 3). We expand on the main benefits of centaur evaluations, which we outlined in this introduction, in Section 4. We then discuss possible objections. Section 5 considers the cost of centaur evaluations, and argues that existing infrastructures from crowd work, A/B tests, and data science competitions, can help reducing the fixed costs of running centaur evaluations. We discuss replicability and sample efficiency of centaur evaluations in Section 6 and conclude in Section 7. Appendix A contains additional examples of centaur evaluations inspired by existing (non-centaur) evaluations and research papers in the social sciences of technology. We keep mathematical notation to a minimum for easier accessibility and only use it in Section 4.3 to highlight how centaur evaluations allow for a formalization of human augmentation.

2. Centaur Evaluations

We first define what centaur evaluations are (see the papers Lee et al. (2024); Shao et al. (2025) formalizations which focus on the interaction model).[1]

2.1. The Gold Standard of Centaur Evaluations

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.

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 component, which is helpful but not integral to centaur evaluations, is a way to communicate *transcripts*. For many cooperative tasks, *how* centaurs achieved a high score in a benchmark is helpful for human learning. Transcripts of successful centaurs allow humans and model developers to improve human-AI collaboration.

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* (compare (Chang et al., 2025) for an example of a centaurized benchmarks). 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). We provide additional centaurized evaluations in Appendix A.1.

Other evaluations are specifically designed with the additional affordances of centaur evaluations in mind. (The following is inspired by the economics paper Brynjolfsson et al. (2025)). A call center agent (human) interacts with a chatbot to help a client with a request via phone. 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 can, subject to the approval of the caller and the agent, be shared (transcript). We propose more (non-centaurized) centaur evaluations in Appendix A.2.

2.2. 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

[1] 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.

coding a functional HTTP server of a centaur compared to a machine learning model and a human alone. The paper Mozannar et al. (2024b) 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. (2024a) 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 centaur. 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. We argue that systematic centaur evaluations are beneficial. First, we argue why centaur benchmarks might not have the standing we argue they deserve.

2.3. Approximating Centaur Evaluations

Centaur benchmarks as a gold standard may be too expensive to run. As a gold standard, they can be approximated, and their calibration tested. *Synthetic centaur evaluations* approximate centaur evaluation using interactive evaluations, compare Park et al. (2023); Aher et al. (2023). For the evaluation of human-centered desiderata this can be conceptually particularly helpful. Consider evaluations of explainability. Envisioning an idealized centaur benchmark, that is then approximated in a synthetic centaur benchmark, which is actually run, can increase transparency on what the actual content of explainabilty in deployment would be.

3. Why Are There Few Centaur Evaluations?

Some of why there are relatively few high-profile centaur evaluations lies in the history of the field.

3.1. Turing

Alan Turing's eponymous test of intelligence of a machine (Turing, 1950) is arguably a main foundation of artificial intelligence, and a deeply human-imitating idea. If the Turing test is the main standard of intelligence, that is whether a human discriminator can distinguish what an algorithm says from what a human says, then the goal of machine learning is indeed human imitation. As generative adversarial networks taught us, developing technology with the goal of passing the Turing test eventually leads to imitation (Goodfellow et al., 2020). The imitation-based approach that Turing started is still engrained in the field. It views human involvement in systems as a distraction from the goal of intelligence, defined by its ability to be indistinguishable from (or surpass in performance) a human.

The thinking of Turing fits directly into one of the most important evaluation datasets, Imagenet (Deng et al., 2009). The dataset assembles *human* labels of images, testing how well a machine can learn what humans perceive in images. The model tries to imitate humans in a task.

Turing's imitative perspective is not the only basis for steering technological progress—other sub-fields of computer science started out differently.

3.2. Bush, Licklider, and Engelbardt

The difference played out in the early days of human-computer and human-robot interaction. Foundational thinkers envisioned technologies that amplify human capabilities rather than replacing them. One foundational example is Vannevar Bush's concept of the *memex* (Bush, 1945). The memex was conceived as a cognitive augmentation tool, enabling individuals to organize and retrieve information seamlessly through associative links, much like internet hyperlinks. Bush's vision prefigured many aspects of modern computing, including the web, and emphasized the potential of technology to augment human thought processes.

Two important thinkers were influenced by Bush's proposal. J.C.R. Licklider further advanced the concept of human-machine cooperation in his influential work on *man-computer symbiosis* (Licklider, 1960). Licklider envisioned a future where humans would handle planning and judgment tasks while machines would process data and perform calculations at unprecedented speeds. This collaboration aimed to improve decision-making efficiency and accuracy, illustrating the profound potential of human-machine partnerships.

Building on Bush's vision, Douglas Engelbart introduced the idea of *bootstrapping*, wherein tools are designed not only to assist humans directly but also to facilitate the creation of better tools (Engelbart, 1962), leading to Engelbart proposing many of modern computer's affordances in the "Mother of all demos" (Engelbart, 1968).

In the footsteps of these thinkers, Human-Computer Interaction works on making humans more productive through technology (Horvitz, 1999; Wang et al., 2020).

3.3. Robotics, Human-Robot Interaction, and the DARPA Grand Challenges

In physical domains and robotics, the Defense Advanced Research Projects Agency (DARPA)'s Grand Challenges demonstrate the principles of augmentation. The DARPA

Robotics Challenge allowed human operators to issue high-level commands, such as *drive forward*, while the robots autonomously handled the fine-grained motion control (Krotkov et al., 2018). This division of labor capitalized on human judgment and machine precision, enabling significant advancements in autonomous systems.

The DARPA Subterranean Challenge extended this idea further by integrating teams of robots with a human operator who had limited observability of the robots' actions (Rouček et al., 2020). This setup required effective communication and coordination, emphasizing the importance of human oversight in complex, dynamic environments. The interaction between humans and robots constitutes the field of Human-Robot Interaction (see, e.g., Lasota et al. (2017); Ajoudani et al. (2017)).

3.4. Current Evaluations in Artificial Intelligence

Why did centaur evaluations not take off in machine learning? It might be a combination of the existence of many imitative evaluations, Turing's lasting impact, or the cost of benchmarks, which makes them less widely accessible.

4. 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 4.1), they allow direct testing of human-centered desiderata like interpretability, human-augmentation, helpfulness, and grounding (Section 4.2), and, for us most importantly, re-center technological development toward human augmentation, while helping policymakers (Section 4.3).

4.1. 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.

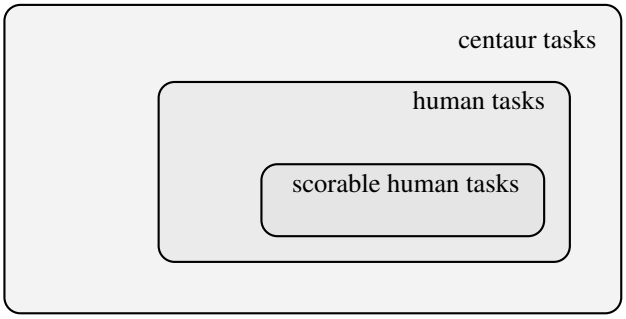

Figure 1: Variation of Brynjolfsson (2022), Figure 1. Imitative evaluations create a low ceiling for evaluations.

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, see Figure 1. 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 4.3). 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.

4.2. 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 theory model, and compare Yang et al. (2018); Okati et al. (2021); Mozannar & 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.

4.3. Centaur Evaluations Positively Impact Society

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

4.3.1. Directing Technological Change

Technology and automation play an important role in the inequality of power and wealth (Karabarbounis & 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 will view the performance of a model i on a centaur benchmark (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[2] Fitting a function, we obtain a *centaur production function* $Y = f_i(K, L)$ of the evaluations. For a moment assume that the evaluation performance is a good proxy for the monetary value of an economically relevant task. Then, the marginal value of human time, $\frac{\partial f_i}{\partial L}$ as a value of human augmentation. The reason for this is that, in a competitive market, the wage w of a worker in a productive task given by production function f_i satisfies

$$\frac{\partial f_i}{\partial L}(K, L) = w. \tag{1}$$

(To see why (1) holds, assume for example—and contradiction—$\frac{\partial f_i}{\partial L}(K, L) > w$. In this case, raising L by ε costs εw, but brings benefit $\varepsilon \frac{\partial f_i}{\partial L}(K, L) > \varepsilon w$, contradiction individual optimality in a market.) To motivate that $\frac{\partial f_i}{\partial L}$, which can only be estimated with a centaur benchmark, can be used to compare models, consider Figure 2 which sketches production functions for two different AI models (or interaction modules) f_1 and f_2, for a fixed level of computation. As a result of optimization, wages are the slope of the production function. As slopes for f_2 are higher than for f_1, for any value of human time, wages will be higher under f_2.

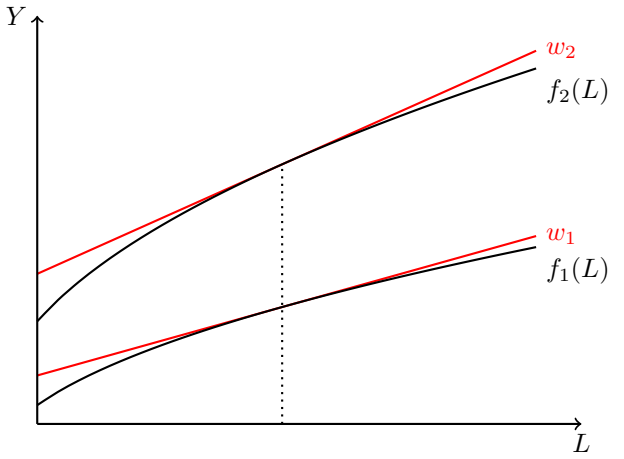

Figure 2: While model 1 is strictly more performant, it is only slightly more human-augmenting.

Informed by (1), we can give a (slightly informal) definition of technologies that are human-augmenting and which are human-replacing. Those machine learning systems that keep the marginal value of human time, and hence, according to (1) high, are called human-augmenting. If human time is irrelevant, the technology is human-replacing.

Definition 4.1. We call a machine learning system with production function f *human-augmenting* if $\frac{\partial f}{\partial L} \gg 0$ for relevant values K and L. If $\frac{\partial f}{\partial L} \approx 0$ 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. The point made here is supported by several economists; see, for example, (Acemoglu & Johnson, 2023a;b; 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).

Current evaluations are blind to human augmentation, as they evaluate $f_i(K, 0)$ or $\max_K f_i(K, 0)$. If the goal is to succeed in current evaluations, there are no incentives for human augmentation.

[2]This notation is inspired by macroeconomics. K, or *capital* is here played by computation, L or *labor* is the human input, Y or *output* is the performance on a task. We refer the interested reader to (Romer, David, 2018) for more macroeconomic modeling.

4.3.2. PRODUCING POLICY-RELEVANT ARTIFACTS

centaur evaluations allow us to produce evaluations with direct meaning for human augmentation and impacts for the value of human time.

- $\frac{\partial f_i}{\partial L}(K, L)$: human augmentation. The expected wage in a thought experiment is informed by (1).
- $f_i(K, L)$: task achievement, fixed resources in terms of both human and compute (compare Coleman et al. (2017) for resource-controlled computing).
- $\max_{K,L} f_i(K, L)$: maximal task achievement. The optimal performance of any centaur.

Using $\frac{\partial f_i}{\partial L}(K, L)$ as a benchmark allows to assess the marginal value of human time for a task. This can inform retraining of humans: If a new very performant ($f_i(K, L) \gg 0$), human-replacing ($\frac{\partial f_i}{\partial L} = 0$) technology arises, retraining toward other tasks is helpful. Conversely, if a new performant, human-augmenting ($\frac{\partial f}{\partial L} \gg 0$) technology is introduced, this makes training of (some) humans for these tasks desirable.

Even beyond tasks for which we *cannot* assume that success is a good proxy of monetary value (as for most tasks), marginal value $\frac{\partial f_i}{\partial L}(K, L)$ is a helpful measure. Consider centaurized MMLU (Hendrycks et al., 2021a). Evaluate the difference in performance between 15 minutes and 30 minutes of human time together with a chatbot to solve parts of a benchmark. We can view this as a finite-difference approximation of $\frac{\partial f_i}{\partial L}(K, L)$. If a system does not benefit at all from human input, we should see that this measure will be close to zero. Is it large, then humans will bring significant value to the system. High human augmentation is informative about whether human thought as opposed to mere knowledge recovery plays a role.

5. How to Run centaur evaluations?

A first concern about centaur evaluation is cost and reproducibility. In this section, we discuss three designs from crowd work, randomized controlled trials, and data science competitions, that can be used for centaur evaluations and how they reduce fixed costs and allow for the reporting of relevant quantities.

5.1. Centaur Evaluations with Crowd Workers

A first component of centaur benchmarks is the selection of humans. Crowd work platforms such as Amazon Mechanical Turk, allow for the recruitment of humans for centaur benchmarks, and has been employed in Lee et al. (2024). Crowd workers are reimbursed for participation in a centaur benchmark, and are chosen from a pool of crowd workers potentially with additional qualifications.

5.2. Centaur Evaluations via Trials

If the interest of a benchmark is not only the comparison of quantities humans and machine learning systems interact, but a comparison of $f_i(K, L)$ of $\frac{\partial f_i}{\partial L}(K, L)$ is desired, causal inference techniques may be used. Instead of only using separate human on for different machine learning systems i, one can randomly assign *treatments* that change the amount of computation used (e.g., by querying a machine learning model to use less chain-of-thought tokens) and providing incentives to humans to use less time (e.g., a Mechanical Turk bonus for each minute they finish early). This allows to learn a model of treatments to expected time spent and an unbiased estimate of $f_i(K, L)$ of $\frac{\partial f_i}{\partial L}(K, L)$, compare Heckman & Robb (1985); Joshua D. Angrist & Rubin (1996). (Compare also Ackerberg et al. (2015) for production function estimation in Economics.)

5.3. Centaur Evaluations via Competitions

So far, we considered examples of humans sampled from a pool. It is also possible to consider settings where the humans are chosen and trained by the entity producing the model to be evaluated. This naturally leads to leaderboards of performance in the spirit of `kaggle.com`. While the former two appraoches aim to choose a representative sample of humans to complete a task, leaderboards optimize both the AI system and the human. To provide statistical meaning to the evaluation with bring-your-own-human and not overfit to the selection of a particular task, this approach will require the humans to solve several tasks before their score is reported.

The usefulness of a leaderboard does not necessarily lie in the numeric evaluation results like in the first two approaches but rather in the transcripts that are produced. Humans can learn from the best humans using AI very productively for a task and improve their actions—a success of social learning.

An important feature of centaur evaluations, we predict, is some amount of adaptability to the discovery of unintended ways to solve a task (glitches, jailbreaks, shortcuts, etc.). We are optimistic that such norms can be found in an online community, as a parallel case of the speed run community shows. In a speed run in a videogame, a human tries to "complete" a video game, that is, reach a particular game state as fast as possible, to rank in a global leaderboard. Leaderboards such as `speedrun.com` have human moderators who determine which glitches, shortcuts, and hardware setups are allowed and which are forbidden (compare the study Scully-Blaker (2016) of the speedrunning community).

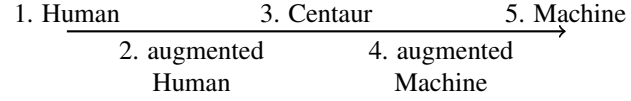

Figure 3: Five stages of automation.

6. Alternative Views

We discuss three additional arguments in opposition to our argument. The first argument Section 6.1 roughly states that *human augmentation is an illusion* and that in relevant tasks centaurs usually perform worse than humans or machine learning systems alone. The second focuses on statistical issues and contends that centaur *evaluations* do not work. The third says that centaur evaluations are broadly not *worth their opportunity cost*.

6.1. Human Augmentation Does Not Exist

Argument. There are many tasks for which centaurs are demonstrably worse than algorithms or humans alone. For example, (Ludwig & Mullainathan, 2021; Kleinberg et al., 2018) show that biases of humans lead to worse performance of centaurs compared to machine learning in several settings of social relevance. Judges perform worse than counterfactual decisions made by algorithms alone (Angelova et al., 2024; Yang et al., 2018), radiology screening algorithms outperform radiologists (Yu et al., 2024), and human-AI systems might be less fair than algorithms alone (McLaughlin et al., 2022). More broadly, the metastudy Vaccaro et al. (2024) finds that there is mostly no human augmentation in studies published in 2020 to 2023.

Rebuttal. This argument only highlights that centaurs' performance is task-dependent (an observation that (Dell'Acqua et al., 2023) formulates). While the argument lists examples where centaurs do not perform well, there are many tasks for which centaurs outperform humans and/or AI. Examples, where such human uplift was demonstrated, are in child protective services reviews (Grimon & Mills, 2022), call centers (Brynjolfsson et al., 2025) and entrepreneurship (Otis et al., 2023).

We also believe that a presumption of a failure of centaurs steers technology in the wrong direction. We rather think of technological automation in *five stages of automation*, see Figure 3 through which all tasks proceed at different speeds. First, humans are doing the task, and technology is too immature to be at all helpful. With more and more capable technology and well-trained humans, centaur performance increases. Finally, machines are capable enough to not benefit from human involvement anymore. One example of such automation is chess. In the last 80 years, we have gone through all five stages of automation for the game of chess and the use of chess agents. During the war, chess did not benefit from computation, and humans were playing it by themselves. More and more, computers helped humans, and in the 2000s, centaur chess tournaments tested different centaurs against each other, see an interview about this time (Sollinger, 2018). Roughly ten years later, there is no benefit to centaurs compared to computers alone, according to (Emerson, 2013).

While sufficient engineering effort can move all tasks through the stages of automation, how this transition works depends on the machine learning community. If the only goal is to reach the final stage of automation, there will be no productive centaur in the time of transition as there is no technology supporting this stage of automation. With the current culture of machine learning evaluation, we sacrifice performance in stages 2 to 4 while waiting for stage 5, at which point large inequality in wealth and power arises. Given how undesirable this outcome is, we believe investing in the development in successful centaur systems is societally beneficial.

6.2. Centaur Evaluations are Noisy and Irreproducible

Argument. Evaluation is the core of machine learning methodology (see Rahimi & Recht (2017); Kolter (2024)), so we should be careful with changes to how we evaluate. Centaur evaluations, at their core, are lab studies, and suffer from issues inherent in them: brittleness and dependence on experimental details (Tversky & Kahneman, 1974; 1981), noisy data, and high sample complexity even for moderately tight comparisons of models.

Rebuttal. We first point out that centaur evaluations are more than RCTs when designed as leaderboards Section 5.3. We engage with the rest of the argument. First, on their dependence on experimental conditions. In our view, the dependence on experimental features is a feature rather than a bug of centaur evaluations. For the design of human-augmenting technologies, it is important to capture human decision-making, which is dependent on details of the setup, and may be viewed as "brittle". Assuming that human selection criteria are transparent enough and interfaces are flexible enough for the designers of systems to design good interfaces, we view this as an opportunity for good design.

Second, on sample complexity. We agree that it might be necessary to run a task on a few hundred (Lee et al., 2024) to a few thousand (Brynjolfsson et al., 2025) humans. This might be expensive, and does not allow for routine evaluation of all models and model iterations. Given the concentration with a handful of performant models, selection of models may be possible.

6.3. Opportunity Costs

Argument. Centaur evaluations are expensive and eat into resources that would otherwise go into the rapid iteration of models. Losing the ability to rapidly iterate is a huge loss for the machine learning community, and it is unclear whether the benefits of centaur evaluations offset that loss.

Rebuttal. The gold standard of a pure centaur benchmark might not be appropriate for all tasks, and approximations might be necessary. For the leading models, we still believe that the benefits for steering technologies outweigh the costs of running the benchmarks. In addition, we have already seen successful centaur benchmarks, which show that they are feasible (Peng et al., 2023; Mozannar et al., 2024b; Peng et al., 2023; Cui et al., 2024; Barke et al., 2023; Mozannar et al., 2024a; Shao et al., 2025; Lee et al., 2024).

7. Summary

Evaluations are crucial for machine learning methodology. Most prominent evaluations of machine learning systems consider the systems in isolation from humans, leading to easily saturated benchmarks, hard-to-formalize human-centered desiderata, and a bias of technological development toward human replacement instead of human augmentation. Human replacement exacerbates an existing imbalance of power and wealth.

We argue that all of these concerns about current evaluations are addressed by *centaur evaluations* in which humans and machine learning systems complete tasks together in a shared environment. Centaur evaluations require a specification of the selection of humans, the human-machine interface, and a scoring mechanism. They may also provide transcripts of interactions, and can be run based on existing infrastructure for crowdsourcing, RCTs, and data science competitions.

Centaur evaluations allow us to identify tasks where human augmentation is most beneficial, as well as those in which machine learning systems outperform humans. The current practice of machine learning system evaluation leads to under-performing centaurs until full automation, upon which many humans lose economic bargaining power and income.

Acknowledgements

This research was supported by funding from the Schmidt Futures and the Project Liberty Institute. We thank David Autor and Stewart Slocum for helpful conversations.

References

Acemoglu, D. and Johnson, S. *Power and progress*. PublicAffairs, May 2023a.

Acemoglu, D. and Johnson, S. Rebalancing ai. *International Monetary Fund*, 2023b.

Ackerberg, D. A., Caves, K., and Frazer, G. Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451, 2015. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/43866416.

Aher, G., Arriaga, R. I., and Kalai, A. T. 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.

Ajoudani, A., Zanchettin, A. M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., and Khatib, O. Progress and prospects of the human–robot collaboration. *Autonomous Robots*, 42(5):957–975, October 2017. ISSN 1573-7527. doi: 10.1007/s10514-017-9677-2. URL http://dx.doi.org/10.1007/s10514-017-9677-2.

Alvaredo, F., Chancel, L., Piketty, T., Saez, E., and Zucman, G. *World Inequality Report 2022*. World Inequality Lab, 2022. URL https://wir2022.wid.world/.

Angelova, V., Dobbie, W., and Yang, C. S. Algorithmic recommendations when the stakes are high: Evidence from judicial elections. In *AEA Papers and Proceedings*, volume 114, pp. 633–637. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2024.

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.

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

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11405–11414, 2021.

Barke, S., James, M. B., and Polikarpova, N. 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.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym, 2016. URL https://arxiv.org/abs/1606.01540.

Brynjolfsson, E. 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.

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

Brynjolfsson, E., Li, D., and Raymond, L. 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.

Bush, V. As we may think. *Atlantic Monthly, July*, 1945.

Cai, H., Cai, X., Chang, J., Li, S., Yao, L., Wang, C., Gao, Z., Wang, H., Li, Y., Lin, M., Yang, S., Wang, J., Xu, M., Huang, J., Fang, X., Zhuang, J., Yin, Y., Li, Y., Chen, C., Cheng, Z., Zhao, Z., Zhang, L., and Ke, G. SciAssess: Benchmarking LLM proficiency in scientific literature analysis, 2024. URL https://arxiv.org/abs/2403.01976.

Casper, S., Li, Y., Li, J., Bu, T., Zhang, K., and Hadfield-Menell, D. Benchmarking interpretability tools for deep neural networks. *arXiv preprint arXiv:2302.10894*, 4, 2023.

Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., Weng, L., and Mądry, A. Mle-bench: Evaluating machine learning agents on machine learning engineering, 2024. URL https://arxiv.org/abs/2410.07095.

Chang, S., Anderson, A., and Hofman, J. M. Chatbench: From static benchmarks to human-ai evaluation, 2025. URL https://arxiv.org/abs/2504.07114.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. Chatbot arena: An open platform for evaluating LLMs by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Chollet, F., Knoop, M., Kamradt, G., and Landers, B. Arc Prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024.

Coleman, C., Narayanan, D., Kang, D., Zhao, T., Zhang, J., Nardi, L., Bailis, P., Olukotun, K., Ré, C., and Zaharia, M. Dawnbench: An end-to-end deep learning benchmark and competition. *Training*, 100(101):102, 2017.

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

Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., and Salz, T. 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.

De, A., Okati, N., Zarezade, A., and Gomez Rodriguez, M. 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.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4573321. URL http://dx.doi.org/10.2139/ssrn.4573321.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., and Su, Y. 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.

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

Drouin, A., Gasse, M., Caccia, M., Laradji, I. H., Del Verme, M., Marty, T., Vazquez, D., Chapados, N., and Lacoste, A. WorkArena: How capable are web agents at solving common knowledge work tasks? In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11642–11662. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/drouin24a.html.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Burstein, J., Doran, C., and Solorio, T. (eds.), *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)*, pp. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246/.

Emerson, D. Computers are now beating humans at 'advanced chess'. *Business Insider*, 2013. URL https://www.businessinsider.com/computers-beating-humans-at-advanced-chess-2013-11. Accessed: 2025-01-30.

Engelbart, D. C. Augmenting human intellect: A conceptual framework. Technical report, Stanford Research Institute, Menlo Park, CA, 1962. URL https://www.dougengelbart.org/pubs/augment-3906.html. Last accessed: January 22, 2025.

Engelbart, D. C. Demo: The augmented knowledge workshop, 1968. URL https://dougengelbart.org/content/view/209/. Last accessed: January 22, 2025.

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.

Glazer, E., Erdil, E., Besiroglu, T., Chicharro, D., Chen, E., Gunning, A., Olsson, C. F., Denain, J.-S., Ho, A.,

de Oliveira Santos, E., Järviniemi, O., Barnett, M., Sandler, R., Vrzala, M., Sevilla, J., Ren, Q., Pratt, E., Levine, L., Barkley, G., Stewart, N., Grechuk, B., Grechuk, T., Enugandla, S. V., and Wildon, M. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024. URL https://arxiv.org/abs/2411.04872.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, October 2020. ISSN 0001-0782. doi: 10.1145/3422622. URL https://doi.org/10.1145/3422622.

Grimon, M.-P. and Mills, C. The impact of algorithmic tools on child protection: Evidence from a randomized controlled trial. *Job market paper*, 2022.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023. URL https://arxiv.org/abs/2308.11462.

Heckman, J. J. and Robb, R. Alternative methods for evaluating the impact of interventions. *Journal of Econometrics*, 30(1–2):239–267, October 1985. ISSN 0304-4076. doi: 10.1016/0304-4076(85)90139-3. URL http://dx.doi.org/10.1016/0304-4076(85)90139-3.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. 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.

Horvitz, E. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pp. 159–166, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 0201485591. doi: 10.1145/302979.303030. URL https://doi.org/10.1145/302979.303030.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Joshua D. Angrist, G. W. I. and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902.

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

Keswani, V., Lease, M., and Kenthapadi, K. Designing closed human-in-the-loop deferral pipelines, 2022. URL https://arxiv.org/abs/2202.04718.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293, 2018.

Kolter, J. Z. Is this really science? a lukewarm defense of alchemy. Talk at NeurIPS 2024 Workshop: Scientific Methods for Understanding Neural Networks, December 2024. Accessed online: https://neurips.cc/virtual/2024/107918.

Krotkov, E., Hackett, D., Jackel, L., Perschbacher, M., Pippine, J., Strauss, J., Pratt, G., and Orlowski, C. The DARPA robotics challenge finals: Results and perspectives. *The DARPA robotics challenge finals: Humanoid robots to the rescue*, pp. 1–26, 2018.

Lasota, P. A., Fong, T., and Shah, J. A. A survey of methods for safe human-robot interaction. *Foundations and Trends® in Robotics*, 5(4):261–349, 2017. ISSN 1935-8253. doi: 10.1561/2300000052. URL http://dx.doi.org/10.1561/2300000052.

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., Gerard-Ursin, I., Li, X. L., Ladhak, F., Rong, F., Wang, R. E., Kwon, M., Park, J. S., Cao, H., Lee, T., Bommasani, R., Bernstein, M., and Liang, P. Evaluating human-language model interaction, 2024. URL https://arxiv.org/abs/2212.09746.

Li, B. Z., Tamkin, A., Goodman, N., and Andreas, J. Eliciting human preferences with language models, 2023. URL https://arxiv.org/abs/2310.11589.

Li, H., Cao, Y., Yu, Y., Javaji, S. R., Deng, Z., He, Y., Jiang, Y., Zhu, Z., Subbalakshmi, K., Xiong, G., Huang, J., Qian, L., Peng, X., Xie, Q., and Suchow, J. W. Investorbench: A benchmark for financial decision-making tasks with llm-based agent, 2024. URL https://arxiv.org/abs/2412.18174.

Licklider, J. C. Man-computer symbiosis. *IRE transactions on human factors in electronics*, HFE(1):4–11, 1960.

Ludwig, J. and Mullainathan, S. Fragile algorithms and fallible decision-makers: lessons from the justice system. *Journal of Economic Perspectives*, 35(4):71–96, 2021.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: Improving fairness and accuracy by learning to defer, 2018. URL https://arxiv.org/abs/1711.06664.

Majumder, B. P., Surana, H., Agarwal, D., Mishra, B. D., Meena, A., Prakhar, A., Vora, T., Khot, T., Sabharwal, A., and Clark, P. DiscoveryBench: Towards data-driven discovery with large language models, 2024. URL https://arxiv.org/abs/2407.01725.

McLaughlin, B., Spiess, J., and Gillis, T. On the fairness of machine-assisted human decisions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 890, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533152. URL https://doi.org/10.1145/3531146.3533152.

Metz, C. 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.

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

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

Mozannar, H., Bansal, G., Fourney, A., and Horvitz, E. 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, 2024a. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3641936. URL https://doi.org/10.1145/3613904.3641936.

Mozannar, H., Chen, V., Alsobay, M., Das, S., Zhao, S., Wei, D., Nagireddy, M., Sattigeri, P., Talwalkar, A., and Sontag, D. The RealHumanEval: Evaluating Large Language Models' abilities to support programmers, 2024b. URL https://arxiv.org/abs/2404.02806.

Okati, N., De, A., and Gomez-Rodriguez, M. Differentiable learning under triage, 2021. URL https://arxiv.org/abs/2103.08902.

Otis, N. G., Clarke, R. P., Delecourt, S., Holtz, D., and Koning, R. The uneven impact of generative AI on entrepreneurial performance, Dec 2023. URL osf.io/hdjpk.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. 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.

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. The impact of AI on developer productivity: Evidence from github copilot, 2023. URL https://arxiv.org/abs/2302.06590.

Rahimi, A. and Recht, B. Reflections on random kitchen sinks, December 2017. URL https://archives.argmin.net/2017/12/05/kitchen-sinks/. Accessed: 2025-01-26.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level Google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

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

Rouček, T., Pecka, M., Čížek, P., Petříček, T., Bayer, J., Šalanskỳ, V., Heřt, D., Petrlík, M., Báča, T., Spurnỳ, V., et al. Darpa subterranean challenge: Multi-robotic exploration of underground environments. In *Modelling and Simulation for Autonomous Systems: 6th International Conference, MESAS 2019, Palermo, Italy, October 29–31, 2019, Revised Selected Papers 6*, pp. 274–290. Springer, 2020.

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

Scully-Blaker, R. *Re-curating the accident: Speedrunning as community and practice*. PhD thesis, Concordia University, 2016.

Shaikh, O., Gligoric, K., Khetan, A., Gerstgrasser, M., Yang, D., and Jurafsky, D. Grounding gaps in language model generations. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 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/.

Shao, Y., Samuel, V., Jiang, Y., Yang, J., and Yang, D. Collaborative gym: A framework for enabling and evaluating human-agent collaboration, 2025. URL https://arxiv.org/abs/2412.15701.

Sollinger, M. 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.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A. K., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ferri, C., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, C. D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-Lopez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H. F. A., Schuetze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocon, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Oliveros-Colón, L., Metz, L., Senel, L. K., Bosma, M., Sap, M., Hoeve, M. T., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Ramirez-Quintana, M. J., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P. W., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R. A., Lee, S. R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S.,

Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S. S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S., Shieber, S., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., vinay uday prabhu, Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, S., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. 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.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *ACL (Findings)*, pp. 13003–13051, 2023. URL https://doi.org/10.18653/v1/2023.findings-acl.824.

Turing, A. M. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. URL https://courses.cs.umbc.edu/471/papers/turing.pdf. Last accessed: January 22, 2025.

Tversky, A. and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157): 1124–1131, 1974. doi: 10.1126/science.185.4157.1124. URL https://www.science.org/doi/abs/10.1126/science.185.4157.1124.

Tversky, A. and Kahneman, D. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. doi: 10.1126/science.7455683. URL https://www.science.org/doi/abs/10.1126/science.7455683.

Vaccaro, M., Almaatouq, A., and Malone, T. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8 (12):2293–2303, October 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-02024-1. URL http://dx.doi.org/10.1038/s41562-024-02024-1.

Vodrahalli, K., Daneshjou, R., Gerstenberg, T., and Zou, J. Do humans trust advice more if it comes from ai? an analysis of Human-AI interactions, 2022. URL https://arxiv.org/abs/2107.07015.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A. (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., and Wang, Q. From human-human collaboration to human-ai collaboration: Designing AI systems that can work together with people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pp. 1–6, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368193. doi: 10.1145/3334480.3381069. URL https://doi.org/10.1145/3334480.3381069.

Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen, M., Clymer, J., Dhyani, J., et al. Re-bench: Evaluating frontier AI r&d capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024.

Xie, J., Zhang, K., Chen, J., Zhu, T., Lou, R., Tian, Y., Xiao, Y., and Su, Y. 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.

Yang, Q., Scuito, A., Zimmerman, J., Forlizzi, J., and Steinfeld, A. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 designing interactive systems conference*, pp. 585–596, 2018.

Yu, F., Moehring, A., Banerjee, O., Salz, T., Agarwal, N., and Rajpurkar, P. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine*, 30(3):837–849, 2024.

Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Menders, C., Lin, J. W., Jones, E., Hussein, G., Liu, S., Jasper, D., Peetathawatchai, P., Glenn, A., Sivashankar, V., Zamoshchin,

D., Glikbarg, L., Askaryar, D., Yang, M., Zhang, T., Alluri, R., Tran, N., Sangpisit, R., Yiorkadjis, P., Osele, K., Raghupathi, G., Boneh, D., Ho, D. E., and Liang, P. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models, 2024. URL https://arxiv.org/abs/2408.08926. Accessed: 2025-01-30.

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., and Neubig, G. 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.

A. Examples of Centaur Evaluations

We list additional examples of centaur evaluations, which are each inspired by either social science studies on human augmentation of technology or non-centaur evaluations.

A.1. Centaurized Evaluations

Example (Inspired by Guha et al. (2023)). A lawyer (human) works alongside an AI contract analysis system to identify potential risks, inconsistencies, or missing clauses in legal documents. The lawyer can ask questions, request clarifications, and accept or reject AI suggestions through a structured review interface (interface). The benchmark score is determined by the accuracy of risk identification, the time spent by the lawyer, and the computational costs associated with the AI model (scoring). A transcript of the lawyer-AI interaction can be stored to understand patterns in effective collaboration (transcript).

Example (Inspired by Cai et al. (2024)). A researcher (human) is given a set of papers and collaborates with an AI system to extract key insights, generate summaries, and identify relevant citations for a literature review. The researcher and AI interact via a text-based interface where the AI provides ranked lists of references, extracts key points, and the researcher can refine queries or adjust summarization parameters (interface). The performance is graded based on the relevance and accuracy of extracted information, the efficiency of the process, and the cost in terms of human effort and AI-generated tokens (scoring). The transcript of interactions, including refinements and queries, is exported (transcript).

Example (Inspired by Li et al. (2024)). A financial planner (human) works with an AI-powered financial model to provide investment recommendations tailored to a client's risk profile and goals (task). The financial planner receives AI-generated insights, including risk analyses and portfolio optimizations, and can modify, approve, or reject them through a structured advisory interface (interface). Performance is graded based on investment outcomes, client satisfaction, time spent on decisions, and computational costs (scoring). Transcripts of these interactions are shared (transcript).

Example (Inspired by Zhang et al. (2024)). A security analyst (human) collaborates with an AI threat detection system to solve capture-the-flag problems. The AI system flags suspicious activities and provides automated recommendations while the human analyst interprets, refines, and executes security measures (interface). The accuracy of threat detection, speed of response, and costs in terms of computational resources and human oversight are evaluated (scoring). The transcript records and shares decision-making patterns (transcript).

Example (Inspired by Jimenez et al. (2023)). A software engineer (human) collaborates with an AI debugging assistant to fix Github issues. The AI suggests possible bug locations, offers code fixes and explains error causes, while the human verifies, modifies, or rejects suggestions (interface). The benchmark evaluates debugging accuracy, time efficiency, and human-AI interaction costs (scoring). Transcripts show the messages that humans send to the system, and the history of edits (transcript).

A.2. Novel Centaur Evaluations

Example (Inspired by Brynjolfsson et al. (2025)). A support agent (human) uses an AI assistant to resolve customer queries more efficiently. The AI suggests responses, retrieves relevant documentation, and assists in troubleshooting, while the human agent makes final decisions and personalizes responses (interface). The benchmark score is based on resolution accuracy, customer satisfaction, and cost in terms of human effort and AI-generated tokens (scoring). Transcripts contained exchanged messages and text transcripts, conditional on consent, of the client conversation (transcript).

Example (Inspired by Yu et al. (2024)). A radiologist (human) collaborates with an AI-powered image analysis tool to diagnose medical conditions from X-rays or MRIs. The radiologist and the AI system communicate through an interface where the AI can highlight potential areas of concern, provide confidence scores, and suggest diagnoses while the human can query, approve, or override suggestions (interface). The evaluation consists of diagnostic accuracy, time taken per case, and any associated costs for human-AI interaction (scoring). Transcripts of these interactions, including decision-making paths and disagreements, are shared. (transcript).