AI Should Not Be an Imitation Game: Centaur Evaluations

Andreas Haupt
h4upt@stanford.edu
Stanford University
Palo Alto, CA, USA

Erik Brynjolfsson
erikb@mstanford.edu
Stanford University
Palo Alto, CA, USA

ABSTRACT

Benchmarks and evaluations are central to machine learning methodology and direct research in the field. Current evaluations test systems in the absence of humans. This position paper argues that the machine learning community should 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 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 AI systems.

CCS CONCEPTS

• **Applied computing** → *Economics*; • **Human-centered computing** → *Heuristic evaluations*; • **Social and professional topics** → Governmental regulations; • **Computing methodologies** → **Learning settings**.

KEYWORDS

evaluation, benchmarks, human augmentation, human replacement, Turing trap, centaurs

1 INTRODUCTION

Benchmarks and evaluations are central to machine learning methodology and direct machine learning research [67]. Machine learning systems also expand into many parts of society, which requires considering the broader impacts of evaluations. This position paper is concerned with how AI system evaluation incorporates humans. **We argue that there should be more (or any systematic) centaur evaluations in which humans and AI solve a task cooperatively.**[1]

Much progress is happening not only in development but also in their evaluation. However, among frequently used evaluations [17, 18, 20, 28, 33, 38, 39, 64, 71, 73, 80, 81], there is no explicit involvement of humans in LLM evaluations.[2] with very few exceptions [69, 83]. Human involvement in evaluation or tuning might be viewed as overfitting or, even worse, cheating. The gold standard of *solving* a task is full automation. A result of this is that models that are good at augmenting or complementing humans are not rewarded, and related capabilities are invisible in the most common evaluations.

We claim that increasing the amount of centaur evaluation in machine learning will benefit society and make three arguments to support our claim. First, centaur evaluations raise the bar for evaluating machine learning capabilities to those that involve human perception and dexterity (Section 4.1), in the spirit of Moravec's paradox: (Moravec [56], p.15): "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." Centaur evaluations might lead us away from evaluating AI with exams Metz [55] and toward evaluations that more closely resemble machine learning use.

Our second argument for centaur benchmarks is that they allow to directly evaluate of human-centered desiderata of machine learning models, such as interpretability [16], complementarity [26], and helpfulness [8] (Section 4.2). This is in contrast to current evaluation methodologies, which require (often unsatisfactory) proxies for these desiderata.

Finally, and for us most importantly, centaur evaluations can recenter 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 for human replacement instead of human augmentation, which has led economists to call for human augmentation Acemoglu and Johnson [2], Brynjolfsson [11], Brynjolfsson and McAfee [13]. Brynjolfsson [11] introduces the *Turing Trap* is the risk of creating technologies that replace humans and leave them without economic and political power. It highlights the dangers of focusing too narrowly on AI systems that imitate human intelligence rather than augment it.

The argument in this position paper is structured as follows. We set the stage by defining centaur evaluations in Section 2. We then trace historical reasons for why, despite several examples in computer science literature, the machine learning community pays little attention to evaluations with humans in Section 3. We expand on the main benefits of centaur evaluations, which we outlined in this introduction, in Section 4. We then go into possible objections. Section 5 considers models for running centaur benchmarks, using infrastructures from crowd work, randomized controlled trials, and competitions. We discuss alternative viewpoints in Section 6. Section 7 concludes. 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

[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. Centaur chess was proposed by former chess world champion Garri Kasparov [70].

[2] This even holds for more complex, multi-step interactive evaluations, compare [84] for travel planning, Majumder et al. [53] for scientific discovery, Deng et al. [25], Zhou et al. [88] for web navigation on tasks, and Drouin et al. [27] for broader knowledge worker tasks.

only use it in Section 4.3 to highlight how centaur benchmarks allow for a formalization of human augmentation.

2 CENTAUR EVALUATIONS

We informally define the benchmark we advocate for.

> A centaur benchmark for a machine learning system consists of three components:
>
> (1) A *task*, *i.e.* a distribution over environments that determine human and machine actions, as well as a selection criterion for humans who partake in the centaur benchmark.
> (2) An *interaction model*, *i.e.* available messages for human and machine learning system, and modalities of exchange of messages,
> (3) A *score* function, which scores actions taken in the environment. Beyond performance, it may also involve human time used and computations undertaken.
>
> A fourth component, which is helpful but not integral to centaur benchmarks, is a way to communicate *transcripts*. For many cooperative tasks, a high score of a system is much less informative than *how* the score was achieved. Transcripts of successful centaurs allow humans and model developers to improve human-AI collaboration.

In principle, there are two types of centaur benchmarks. The first is raising the restriction of current benchmarks that they must not involve humans. We call these *centaurized benchmarks*. Consider, for example, MMLU Hendrycks et al. [38] without the requirement that no human should be involved in the solution of the task. MMLU prompts are provided to a human, who is asked to, after deliberation, provide a response (task). Humans and LLMs can exchange messages via text (interaction). 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). These are relatively low-effort ways to "centaurize" existing benchmarks. We provide additional centaurized evaluations in Appendix A.1.

Other evaluations are specifically designed with the additional affordances of centaur benchmarks in mind. (The following is inspired by the social science paper Brynjolfsson et al. [12]). A call center agent interacts with a chatbot to help a client with a request via phone (task). 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 centaur evaluations in Appendix A.2.

We want to highlight that there are systems proposed that standardize components of centaur evaluations. The concurrent research Shao et al. [69] proposes an interface for interactions in centaur evaluations (the authors of [69] use "collaborative agents" instead of centaurs). They implement an asynchronous computation and communication handler with an interface similar to OpenAI's Gym [10].

3 WHY ARE THERE FEW CENTAUR BENCHMARKS?

The reason for why benchmarks are so often based on imitation may be partly traced back to the historical roots of the machine learning and artificial intelligence research communities.

3.1 Turing

Alan Turing's eponymous test of intelligence of a machine [76] is arguably a main foundation of artificial intelligence, yet also a deeply human-imitating idea: The main standard of intelligence is whether a human discriminator can distinguish what an algorithm says from what a human says. As generative adversarial networks taught us, developing technology with the goal of passing the Turing test eventually leads to imitation [34]. The imitation-based approach that Turing started, we claim is still a foundation of what makes "good AI". It views human involvement in systems as a distraction from the goal of intelligence, defined by its ability to be indistinguishable from a human.

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 replace them. One foundational example is Vannevar Bush's concept of the *memex* [14]. 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* [50]. 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 [30], leading to Engelbart proposing many of modern computer's affordances in the "Mother of all demos" [31].

In the footsteps of these thinkers, Human-Computer Interaction works on making humans more productive through technology [40, 82].

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 [47]. 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 [66]. 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.*, Ajoudani et al. [4], Lasota et al. [48]).

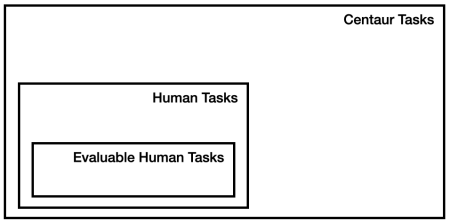


Figure 1: Variation of Brynjolfsson [11], Figure 1. Imitative benchmarks create a low ceiling for what productive use is possible with AI, as centaurs can act in strictly more environments.

3.4 Current Evaluations in Artificial Intelligence

Hence, we may see the reasons for a smaller number of centaur evaluations in the intellectual history (Turing *vs.* Bush/Lickleider/Engelbardt). We also saw in the Introduction that incentives (availability of imitation datasets, easy conceptualization of a gold standard) are pointing toward the evaluation of imitation. Centaur evaluation is, however, not without precedent, as we saw in the Grand Challenges. Centaurs are evaluated regularly in other fields of computer science, such as Human-Computer Interaction and Human-Robot Interaction.

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, and helpfulness (Section 4.2), and, for us most importantly, recenter technological development toward human augmentation, while helping policymakers (Section 4.3).

4.1 Centaur Evaluations Can Be Harder

Current evaluations "saturate", that is, AI models rapidly achieve very good results on benchmarks, leading to concerns that soon, humans might not be able to evaluate models [55, 62]. We contend that this worry might be a consequence of how restrictive current evaluation formats are rather than an imbalance in capability between humans and machine learning systems. So while benchmarks might be 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.*, Deng et al. [25], Drouin et al. [27], Majumder et al. [53], Shao et al. [69], Wijk et al. [83], Xie et al. [84], Zhou et al. [88]), there are strong reasons to consider centaur benchmarks for harder and more realistic benchmarks.

One way in which centaur benchmarks 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. Humans are still often able to distinguish whether they are talking to an AI or a human and will treat 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 last 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 benchmarks look synthetic: school-level [39] and researcher-level mathematics [33], general knowledge questions [38], and reading comprehension [28], 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 provide 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, or helpfulness. One such desideratum, *explainability*, has received attention in policy for example in the European Union's AI Act (European Union [32], 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. [16]. With centaur evaluations, explainability can be directly evaluated as the ability of a human to act together with the system.

Additionally, current benchmarks 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. [9] for a theory model, and compare Bansal et al. [9], De et al. [23], Keswani et al. [44], Madras et al. [52], Mozannar and Sontag [57], Okati et al. [58], Vodrahalli et al. [79], Yang et al. [85]). In current evaluations that do not consider human-AI interplay, learning-to-defer does not have a benefit. Successful deferral helps in real-world use, but current evaluations are blind to it.

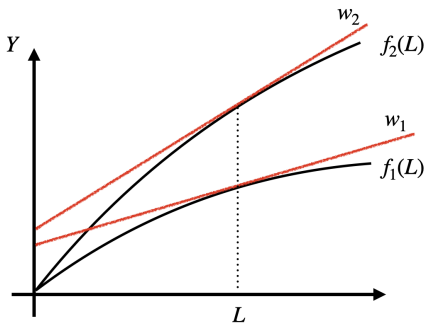


Figure 2: Higher marginal productivity of human time leads to higher wages.

4.3 Centaur Evaluations Positively Impact Society

Finally, centaur evaluations recenter the direction of progress in machine learning and can help policymakers.

4.3.1 Directing Technological Change. Technology and automation play an important role in the inequality of power and wealth [7, 43]. 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 [5]. 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 precise, we will view a centaur benchmark (including task, interaction model, and score function) of economically relevant tasks through the lens of triples (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[3] Regressing output on the $Y = f(K, L)$, we obtain a function, which we call the *production function* of the benchmark. As one of the strong assumptions for this setting, we assume that Y is a good proxy for the monetary benefits of the economically relevant task so that we can compare Y to wages that a human earns. We propose to use the marginal value of human time, $\frac{\partial f}{\partial L}$ 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 satisfies

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

(To see why (1) holds, assume for example—and contradiction—$\frac{\partial f}{\partial L}(K, L) > w$. In this case, raising L by ε costs εw, but brings benefit $\varepsilon \frac{\partial f}{\partial L}(K, L) > \varepsilon w$, contradiction individual optimality in a market.) To motivate that $\frac{\partial f}{\partial L}$, 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.

Informed by (1), we can give a (slightly informal) definition of technologies that are human-augmenting and which are human-replacing. Informally, those that keep the marginal value of human time, and hence, according to (1), wages, high, are called human-augmenting. If human time is (close to) 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, [1, 2, 11]. Even institutions at the center of technological disruption call for ways to increase the number of jobs [22].

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

4.3.2 Producing Policy-Relevant Artifacts. Centaur benchmarks allow us to produce more policy-relevant objects, which we define here (we leave the actual estimation of the objects for future research). Examples of such objects are:

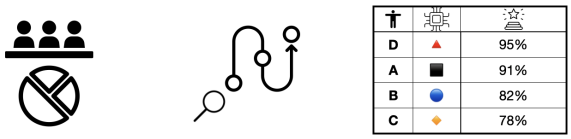
- $f(K, L)$: task achievement, fixed resources. This is the value controlling for resource use (compare Coleman et al. [21] for monitoring use of compute)
- $\max_{K,L} f(K, L)$: maximal task achievement. The optimal performance of any centaur.
- $\frac{\partial f}{\partial L}(K, L)$: human augmentation. The expected wage in a thought experiment is informed by (1).

Using $\frac{\partial f}{\partial L}(K, L)$ 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 ($f(K, L) \gg 0$), human-replacing ($\frac{\partial f}{\partial L} = 0$) 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 is a signal to train more humans in this task.

Even beyond tasks for which we *cannot* assume that success is a good proxy of monetary value (as for most tasks), marginal value is helpful. Consider a centaurized version of MMLU [38]. 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 $\frac{\partial f}{\partial L}(K, L)$. 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. If there is high human value, it might be a good sign for the system in work with knowledge workers.

4.3.3 Providing Transcripts to Train Humans. Finally, centaur evaluations yield transcripts, which help human inductive learning for better cooperation with systems. We discuss the possibilities of

[3] 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 [65] for more macroeconomic modeling.



(a) Crowd Workers (b) RCTs (c) Competitions

Figure 3: Three models for centaur benchmarks.

	Representativity	Quality	Affordability
Crowd work		✓*	✓*
Trial	✓	✓*	✓*
Competition		✓	✓

Table 1: Three Models for centaur evaluations.

transcripts further in the competition model of centaur benchmarks below; see Section 5.3.

5 HOW TO RUN CENTAUR BENCHMARKS?

A first concern about centaur evaluation and costs. In this section, we relate centaur benchmarks to existing infrastructure from crowd work, randomized controlled trials, and competitions, which we call *models of centaur evaluation*. We summarize the comparative advantages of the models on representativity of tasks, quality of human participation, and cost in Table 1. We also discuss specific challenges of the approaches and how they can be overcome.

This section focuses on how to find humans that partake in centaur evaluations. We are confident that the research community will develop successful interfaces, as it has done in other domains (*e.g.*, Pei et al. [60], Perry [61], Tkachenko et al. [74] for text domains, see Shao et al. [69] for interaction models for centaur evaluations). We also do not focus on the design of tasks but give examples in Appendix A.

5.1 Centaur Evaluations via Crowd Work

A first model of centaur evaluations relies on the infrastructure of crowd work, *e.g.*, from Amazon Mechanical Turk. In this case, crowd workers choose to participate in centaur benchmarks for reimbursement and are incentivized to solve a task through a score-dependent bonus (Figure 3a). The distribution of the performance of different crowd workers is reported as the centaur evaluation.

This approach has benefits in standardization: Tasks, participant selection criteria, and time used can be specified. On the flip side, this means that tasks are less representative of real use.

Both the quality and cost of centaur evaluations with crowd workers will depend on the qualifications of crowd workers doing the task. While tests without particular qualifications may be affordable, professional qualifications (*e.g.*, in software engineering) might be more expensive.

5.2 Centaur Evaluations via Trials

Centaur benchmarks with panels might still be quite expensive because humans are not productive beyond the evaluation when completing it. Randomized controlled trials may be used, in combination with causal inference techniques, to run centaur evaluations (Figure 3b).

In a classical randomized controlled trial (RCT), humans are assigned to a treatment arm, and differences in (potentially conditionally) average outcomes are determined [37]. For example, the difference in the likelihood of a click may be compared for different user interfaces on a web application. It is not direct to put the estimation of an object like a production function $f(K, L)$ or even human augmentation in this framework.

Thankfully, causal inference techniques can help for such estimation. (See Ackerberg et al. [3] for related work on production function estimation in Economics.) To illustrate the possibility, assume we are running an RCT which treats software engineers to different bonuses for each minute they finish before a certain time. Engineers choose the amount of time and computation they use, partly based on features of the task that are unobservable to us, leading to bias in naïve estimation of $f(K, L)$. Learning a model of how engineers choose time and compute based on treatments allows to *debias* estimates, compare Joshua D. Angrist and Rubin [42].

5.3 Centaur Evaluations via Competitions

A quite different approach to centaur evaluations is the leaderboard, inspired by platforms like kaggle.com (Figure 3c). 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.

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 benchmarks, 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 (see Scully-Blaker [68] on the speedrunning community)

Beyond the three models outlined here, additional work, we are optimistic, will help make centaur benchmarks more affordable, reliable, and representative.

The rest of this position paper considers objections to our argument.

6 ALTERNATIVE VIEWPOINTS

We discuss two additional arguments in opposition to our argument. The first argument Section 6.1 roughly states that *centaurs do not improve performance*. The second focuses on statistical issues and contends that centaur *evaluations* do not work.

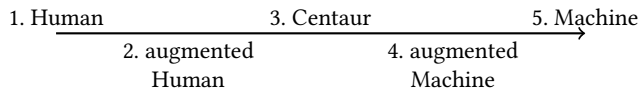


Figure 4: Five stages of automation.

6.1 Human Augmentation Does Not Exist

Argument. There are many tasks for which centaurs are demonstrably worse than algorithms alone. For example, [45, 51] show that biases of humans lead to worse performance than in combination with AI in several settings of social relevance. Judges perform worse than counterfactual decisions made by algorithms alone [6, 85], radiology screening algorithms outperform radiologists [86], and human-AI systems might be less fair than algorithms alone [54]. Such a viewpoint led Cass Sunstein to propose to "governance by algorithm", absent any human biases (2021).

Rebuttal. This argument only highlights that centaurs' performance is task-dependent (an observation that [24] 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 [35], call centers [12], entrepreneurs [59], and material scientists Toner-Rodgers [75].

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 4 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 [70]. Roughly ten years later, there is no benefit to centaurs compared to computers alone, according to [29].

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 ("machine"), there will be no productive centaur in the middle because there is no technology for this stage of automation. With the current culture of machine learning evaluation, we see low performance in stages 2 to 4 while waiting for stage 5, at which point large inequality in wealth and power arises. We believe presuming that centaurs can perform is a societally beneficial assumption.

6.2 Centaur Evaluations Have Insufficient Statistical Power

Argument. Benchmarks are the core of machine learning methodology (see Kolter [46], Rahimi and Recht [63]), so we should be careful with changes to evaluation. Centaur benchmarks, at their core, are glorified RCTs. They suffer from the same issues these have: brittleness and dependence on experimental details [77, 78], noisy data, and high sample complexity even for moderately tight comparisons of models.

Rebuttal. On this point, we first point out that centaur evaluations are more than RCTs, see Sections 5.2 and 5.3.

On the other points, first, on brittleness. The argument cites literature from the behavioral sciences (in particular, social psychologists). Note that centaur evaluations do not face the same challenges as behavioral sciences, as they do not involve human choice as an expression of their preference. In particular, in a centaur benchmark, results are scored as correct or not.[4] It might be that a particular interaction model or a particular way to communicate the task leads to task improvements. This, however, is not an instance of the brittleness of results but an integral part of what centaurs optimize for. Small changes that make systems work better with humans are part of the design space in centaur evaluations.

Second, on sample complexity. Results might indeed be noisy and/or conditional on a high number of covariates because humans are very heterogeneous. We agree that sample complexity might be substantial (or as high as in other studies involving humans). This, however, does not run counter to our call for more centaur benchmarks. Currently, there are, to our knowledge, no centaur benchmarks for large language models. We believe that *some* centaur benchmarks will be worth paying for sufficiently many human samples.

7 SUMMARY

Evaluations are crucial for machine learning methodology. Current benchmarks consider machine learning 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 consist of a task, an interaction module, a scoring function, and a transcript module. They can be run based on existing infrastructure for crowdsourcing, RCTs, and machine learning competitions.

Centaur evaluations allow us to identify those 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.

We call on the machine learning community to evaluate systems using centaur evaluations, where humans and AI jointly solve a task in a shared environment.

[4] If humans make the choice for how to score outputs, such as in LM Arena Chiang et al. [19], the criticism from the behavioral sciences made here applies. Issues of human scoring are not the focus of this position paper.

REFERENCES

[1] Daron Acemoglu and Simon Johnson. 2023. *Power and progress.* PublicAffairs.

[2] Daron Acemoglu and Simon Johnson. 2023. Rebalancing AI. *International Monetary Fund* (2023).

[3] Daniel A. Ackerberg, Kevin Caves, and Garth Frazer. 2015. IDENTIFICATION PROPERTIES OF RECENT PRODUCTION FUNCTION ESTIMATORS. *Econometrica* 83, 6 (2015), 2411–2451. http://www.jstor.org/stable/43866416

[4] Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schäffer, Kazuhiro Kosuge, and Oussama Khatib. 2017. Progress and prospects of the human–robot collaboration. *Autonomous Robots* 42, 5 (Oct. 2017), 957–975. https://doi.org/10.1007/s10514-017-9677-2

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

[6] Victoria Angelova, Will Dobbie, and Crystal S Yang. 2024. Algorithmic Recommendations When the Stakes Are High: Evidence from Judicial Elections. In *AEA Papers and Proceedings*, Vol. 114. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 633–637.

[7] David H Autor. 2019. Work of the Past, Work of the Future. In *AEA Papers and Proceedings*, Vol. 109. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 1–32.

[8] 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. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862 [cs.CL] https://arxiv.org/abs/2204.05862

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

[10] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. arXiv:1606.01540 [cs.LG] https://arxiv.org/abs/1606.01540

[11] Erik Brynjolfsson. 2022. The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. *Daedalus* 151, 2 (05 2022), 272–287. https://doi.org/10.1162/daed_a_01915 arXiv:https://direct.mit.edu/daed/article-pdf/151/2/272/2060604/daed_a_01915.pdf

[12] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. 2023. *Generative AI at Work.* Working Paper 31161. National Bureau of Economic Research. https://doi.org/10.3386/w31161

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

[14] Vannevar Bush. 1945. As we may think. *Atlantic Monthly, July* (1945).

[15] Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, Zheng Cheng, Zifeng Zhao, Linfeng Zhang, and Guolin Ke. 2024. SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis. arXiv:2403.01976 [cs.CL] https://arxiv.org/abs/2403.01976

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

[17] 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. 2024. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. arXiv:2410.07095 [cs.CL] https://arxiv.org/abs/2410.07095

[18] 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. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[19] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132* (2024).

[20] Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. 2024. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604* (2024).

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

[22] Y Combinator. 2024. One Million Jobs 2.0. https://www.youtube.com/watch?v=BAeBkS2gBpo Accessed: 2025-01-22.

[23] Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez Rodriguez. 2021. Classification Under Human Assistance. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 7 (May 2021), 5905–5913. https://doi.org/10.1609/aaai.v35i7.16738

[24] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4573321

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

[26] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. arXiv:2202.08821 [cs.CY] https://arxiv.org/abs/2202.08821

[27] Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. 2024. WorkArena: How Capable are Web Agents at Solving Common Knowledge Work Tasks?. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 11642–11662. https://proceedings.mlr.press/v235/drouin24a.html

[28] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 2368–2378. https://doi.org/10.18653/v1/N19-1246

[29] Doug Emerson. 2013. Computers Are Now Beating Humans at 'Advanced Chess'. *Business Insider* (2013). https://www.businessinsider.com/computers-beating-humans-at-advanced-chess-2013-11 Accessed: 2025-01-30.

[30] Douglas C. Engelbart. 1962. *Augmenting Human Intellect: A Conceptual Framework.* Technical Report. Stanford Research Institute, Menlo Park, CA. https://www.dougengelbart.org/pubs/augment-3906.html Last accessed: January 22, 2025.

[31] Douglas C. Engelbart. 1968. Demo: The Augmented Knowledge Workshop. https://dougengelbart.org/content/view/209/ Last accessed: January 22, 2025.

[32] European Union. 2024. 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. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng Accessed: 2025-01-29.

[33] 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. 2024. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI. arXiv:2411.04872 [cs.AI] https://arxiv.org/abs/2411.04872

[34] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (Oct. 2020), 139–144. https://doi.org/10.1145/3422622

[35] Marie-Pascale Grimon and Christopher Mills. 2022. The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial. *Job market paper* (2022).

[36] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. arXiv:2308.11462 [cs.CL] https://arxiv.org/abs/2308.11462

[37] James J. Heckman and Richard Robb. 1985. Alternative methods for evaluating the impact of interventions. *Journal of Econometrics* 30, 1–2 (Oct. 1985), 239–267. https://doi.org/10.1016/0304-4076(85)90139-3

[38] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*. https://openreview.net/forum?id=d7KBjmI3GmQ

[39] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=7Bywt2mQsCe

[40] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[41] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770* (2023).

[42] Guido W. Imbens Joshua D. Angrist and Donald B. Rubin. 1996. Identification of Causal Effects Using Instrumental Variables. *J. Amer. Statist. Assoc.* 91, 434 (1996), 444–455. https://doi.org/10.1080/01621459.1996.10476902 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1996.10476902

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

[44] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. 2022. Designing Closed Human-in-the-loop Deferral Pipelines. arXiv:2202.04718 [cs.HC] https://arxiv.org/abs/2202.04718

[45] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[46] J. Zico Kolter. 2024. Is this really science? A lukewarm defense of alchemy. Talk at NeurIPS 2024 Workshop: Scientific Methods for Understanding Neural Networks. Accessed online: https://neurips.cc/virtual/2024/107918.

[47] Eric Krotkov, Douglas Hackett, Larry Jackel, Michael Perschbacher, James Pippine, Jesse Strauss, Gill Pratt, and Christopher Orlowski. 2018. The DARPA robotics challenge finals: Results and perspectives. *The DARPA robotics challenge finals: Humanoid robots to the rescue* (2018), 1–26.

[48] Przemyslaw A. Lasota, Terrence Fong, and Julie A. Shah. 2017. A Survey of Methods for Safe Human-Robot Interaction. *Foundations and Trends® in Robotics* 5, 4 (2017), 261–349. https://doi.org/10.1561/2300000052

[49] Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Koduvayur Subbalakshmi, Guojun Xiong, Jimin Huang, Lingfei Qian, Xueqing Peng, Qianqian Xie, and Jordan W. Suchow. 2024. INVESTORBENCH: A Benchmark for Financial Decision-Making Tasks with LLM-based Agent. arXiv:2412.18174 [cs.CE] https://arxiv.org/abs/2412.18174

[50] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* HFE, 1 (1960), 4–11.

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

[52] David Madras, Toniann Pitassi, and Richard Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. arXiv:1711.06664 [stat.ML] https://arxiv.org/abs/1711.06664

[53] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models. arXiv:2407.01725 [cs.CL] https://arxiv.org/abs/2407.01725

[54] Bryce McLaughlin, Jann Spiess, and Talia Gillis. 2022. On the Fairness of Machine-Assisted Human Decisions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 890. https://doi.org/10.1145/3531146.3533152

[55] Cade Metz. 2025. A.I. Poses Humanity's "Last Exam." Are We Ready? *The New York Times* (23 January 2025). https://www.nytimes.com/2025/01/23/technology/ai-test-humanitys-last-exam.html

[56] Hans P Moravec. 1990. *Mind children*. Harvard University Press, London, England.

[57] Hussein Mozannar and David Sontag. 2021. Consistent Estimators for Learning to Defer to an Expert. arXiv:2006.01862 [cs.LG] https://arxiv.org/abs/2006.01862

[58] Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. 2021. Differentiable Learning Under Triage. arXiv:2103.08902 [stat.ML] https://arxiv.org/abs/2103.08902

[59] Nicholas G Otis, Rowan P Clarke, Solene Delecourt, David Holtz, and Rembrand Koning. 2023. The Uneven Impact of Generative AI on Entrepreneurial Performance. https://doi.org/10.31219/osf.io/hdjpk

[60] Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Wanxiang Che and Ekaterina Shutova (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 327–337. https://doi.org/10.18653/v1/2022.emnlp-demos.33

[61] Tal Perry. 2021. LightTag: Text Annotation Platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Heike Adel and Shuming Shi (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 20–27. https://doi.org/10.18653/v1/2021.emnlp-demo.3

[62] Arc Prize. 2025. OpenAI's GPT-4o and the Next Frontier in AI Research. https://arcprize.org/blog/oai-o3-pub-breakthrough Accessed: 2025-01-29.

[63] Ali Rahimi and Benjamin Recht. 2017. Reflections on Random Kitchen Sinks. https://archives.argmin.net/2017/12/05/kitchen-sinks/ Accessed: 2025-01-26.

[64] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*. https://openreview.net/forum?id=Ti67584b98

[65] Romer, David. 2018. *Advanced Macroeconomics* (5 ed.). McGraw-Hill Education, Columbus, OH.

[66] Tomáš Rouček, Martin Pecka, Petr Čížek, Tomáš Petříček, Jan Bayer, Vojtěch Šalanskỳ, Daniel Heřt, Matěj Petrlík, Tomáš Báča, Vojěch Spurnỳ, et al. 2020. 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*. Springer, 274–290.

[67] D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's Curse? On Pace, Progress, and Empirical Rigor. https://openreview.net/forum?id=rJWF0Fywf

[68] Rainforest Scully-Blaker. 2016. *Re-curating the accident: Speedrunning as community and practice*. Ph. D. Dissertation. Concordia University.

[69] Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. 2025. Collaborative Gym: A Framework for Enabling and Evaluating Human-Agent Collaboration. arXiv:2412.15701 [cs.AI] https://arxiv.org/abs/2412.15701

[70] Marc Sollinger. 2018. Garry Kasparov and the Game of Artificial Intelligence. https://theworld.org/stories/2018/01/05/garry-kasparov-and-game-artificial-intelligence Last accessed: January 22, 2025.

[71] 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. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=uyTL5Bvosj Featured Certification.

[72] Cass R Sunstein. 2021. Governing by algorithm? No noise and (potentially) less bias. *Duke LJ* 71 (2021), 1175.

[73] 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. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *ACL (Findings)*. 13003–13051. https://doi.org/10.18653/v1/2023.findings-acl.824

[74] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. https://github.com/HumanSignal/label-studio Open source software available from https://github.com/HumanSignal/label-studio.

[75] Aidan Toner-Rodgers. 2024. Artificial Intelligence, Scientific Discovery, and Product Innovation. arXiv:2412.17866 [econ.GN] https://arxiv.org/abs/2412.17866

[76] Alan M. Turing. 1950. Computing Machinery and Intelligence. *Mind* LIX, 236 (1950), 433–460. https://courses.cs.umbc.edu/471/papers/turing.pdf Last accessed: January 22, 2025.

[77] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124 arXiv:https://www.science.org/doi/pdf/10.1126/science.185.4157.1124

[78] Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211, 4481 (1981), 453–458. https://doi.org/10.1126/science.7455683 arXiv:https://www.science.org/doi/pdf/10.1126/science.7455683

[79] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. 2022. Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions. arXiv:2107.07015 [cs.AI] https://arxiv.org/abs/2107.07015

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

[81] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (Eds.). Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

[82] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. 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* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3381069

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

[84] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. TravelPlanner: a benchmark for real-world planning with language agents. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2246, 24 pages.

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

[86] Feiyang Yu, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar. 2024. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine* 30, 3 (2024), 837–849.

[87] Andy K. Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin W. Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Mike Yang, Teddy Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Polycarpos Yiorkadjis, Kenny Osele, Gautham Raghupathi, Dan Boneh, Daniel E. Ho, and Percy Liang. 2024. Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models. arXiv:2408.08926 [cs.CR] https://arxiv.org/abs/2408.08926 Accessed: 2025-01-30.

[88] 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. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *The Twelfth International Conference on Learning Representations*. 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. [36]). A lawyer works alongside an AI contract analysis system to identify potential risks, inconsistencies, or missing clauses in legal documents (task). The lawyer can ask questions, request clarifications, and accept or reject AI suggestions through a structured review interface (interaction). 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. [15]). A researcher 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 (task). 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 (interaction). 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. [49]). A financial planner 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 (interaction). 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. [87]). A security analyst collaborates with an AI threat detection system to solve capture-the-flag problems (task). The AI system flags suspicious activities and provides automated recommendations while the human analyst interprets, refines, and executes security measures (interaction). 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. [41]). A software engineer collaborates with an AI debugging assistant to fix Github issues (task). The AI suggests possible bug locations, offers code fixes and explains error causes, while the human verifies, modifies, or rejects suggestions (interaction). 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. [12]). A support agent uses an AI assistant to resolve customer queries more efficiently (task). The AI suggests responses, retrieves relevant documentation, and assists in troubleshooting, while the human agent makes final decisions and personalizes responses (interaction). 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. [86]). A radiologist collaborates with an AI-powered image analysis tool to diagnose medical conditions from X-rays or MRIs (task). 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 (interaction). 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).