AI as a Sport: On the Competitive Epistemologies of Benchmarking

Will Orr
orrw@usc.edu
University of Southern California
Los Angeles, California, USA

Edward B. Kang
New York University
New York, USA
e.kang@nyu.edu

ABSTRACT

Artificial Intelligence (AI) systems are evaluated using competitive methods that rely on benchmark datasets to determine performance. These benchmark datasets, however, are often constructed through arbitrary processes that fall short in encapsulating the depth and breadth of the tasks they are intended to measure. In this paper, we interrogate the naturalization of benchmark datasets as veracious metrics by examining the historical development of benchmarking as an epistemic practice in AI research. Specifically, we highlight three key case studies that were crucial in establishing the existing reliance on benchmark datasets for evaluating the capabilities of AI systems: (1) the sharing of Highleyman's OCR dataset in the 1960s, which solidified a community of knowledge production around a shared benchmark dataset, (2) the Common Task Framework (CTF) of the 1980s, a state-led project to standardize benchmark datasets as legitimate indicators of technical progress; and (3) the Netflix Prize which further solidified benchmarking as a competitive goal within the ML research community. This genealogy highlights how contemporary dynamics and limitations of benchmarking developed from a longer history of collaboration, standardization, and competition. We end with reflections on how this history informs our understanding of benchmarking in the current era of generative artificial intelligence.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies**; • **Applied computing**;

KEYWORDS

Machine learning benchmarks, Machine learning competitions, History of benchmarking, Benchmarking for generative AI, Benchmark datasets.

ACM Reference Format:
Will Orr and Edward B. Kang. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3630106.3659012

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659012

1 INTRODUCTION

LLM – Detect AI Generated Text: Identify which essay was written by a large language model

"This competition challenges participants to develop a machine learning model that can accurately detect whether an essay was written by a student or an LLM. The competition dataset comprises a mix of student-written essays and essays generated by a variety of LLMs." [29]

This is the description for one of the over 600 machine learning (ML) "competitions" hosted on Kaggle, a data science competition platform hosted by Google. At the time of writing in December 2023, this particular competition's participation metrics boasted over 2,789 competitors across 2,516 teams, and over 38,900 entries. It is scheduled to award over $110,000 in prize money to the winners and is funded by major foundations including the Bill & Melinda Gates Foundation, Schmidt Futures, and the Chan Zuckerberg Initiative. These sorts of ML competitions have now become a taken-for-granted infrastructure of ML practice: how else would machine learning systems be evaluated, and technological progress measured?

An indispensable component of these competitions is the "competition dataset." This essentially refers to the standardized dataset that all competitors work with to develop and test their models. Within the context of the competition, this dataset becomes the central artifact around which the parameters of the task to be represented and measured are defined. It might be thought of along a similar vein as a professional basketball court for the National Basketball Association (NBA) – the lines on the court determine important boundaries that distinguish between in-bound vs out-of-bounds, 2-points vs 3-points, in-paint vs out-of-paint etc. The standardization of these courts and the rules that follow allow for competitive games to be played, in which a win between two teams in Toronto during the Fall becomes comparable to another win between two different teams in New Orleans during the Spring. As Bowker and Star [8] have written about at length, such a *standard*, understood as a "set of agreed-upon rules for the production of (textual or material) objects" [p. 13], allows for a community of practice to form that can persist across space and time.

We begin our paper with a discussion of competitions and the standards that make them possible because we're interested in querying the emergence of leaderboards, accuracy scores, and competition datasets and how they became accepted features for evaluating the progress of machine learning. In other words, how did machine learning become a sport? Here, we understand 'sport' as a domain where participants compete amongst a community under a standardized set of rules and objectives, striving for recognition within a structured framework. To address this question, we zoom

out from specific competition datasets, and instead, examine the role of benchmark datasets in ML practice, writ large. Defined as a "particular combination of a dataset or sets of datasets (at least test data, sometimes also training data), and a metric, conceptualized as representing one or more specific tasks or sets of abilities, picked up by a community of researchers as a shared framework for the comparison of methods" [48, p. 2], the benchmark dataset as a standardized framework for comparison has become a core element of the broader culture of machine learning. Indeed, it is near impossible to conceive of a recognized machine learning task today outside of the benchmarks that have come to effectively define and represent these tasks, such as ImageNet [13] for object detection in computer vision or GLUE [57] for tasks such as paraphrasing and textual entailment in natural language processing (NLP). The benchmark dataset has, in this way, become the standard that both brings together particular subcommunities of ML researchers, as well as further enables their 'progression' through the competitive and iterative comparison of ML models.

In this paper, we trace how this came to be. Specifically, we argue that there were three key moments in the development of pattern recognition and eventually machine learning research that positioned the benchmark dataset as an indispensable feature of machine learning culture: (1) the distribution of Highleyman's dataset among pattern recognition researchers working separately on optical character recognition (OCR) in the 1960s, (2) the emergence of the Common Task Framework (CTF) in the 1980s that set standards supported by the US federal government for how to share data and evaluations in AI research, and (3) the Netflix competition of the 2000s which incentivized machine learning researchers to compete for a million dollar prize in developing the best recommendation algorithm using a dataset provided by Netflix. We select these three cases because they each illustrate a pivotal shift in how AI systems were evaluated. More specifically, these historical moments together trace the gradual development of (1) a community of knowledge production, (2) standardized practices for evaluating progress, and (3) a platform for competition. By focusing on these three pivotal moments – each roughly two decades apart – we trace a genealogy [15] of how organizing and sharing data, and bounding a community around an accessible standard eventually evolved and established the grounds for the competitive computational culture we see in machine learning today. While these three cases are by no means the only important historical cases of AI benchmarking, we present them as particularly striking moments that represent the progression we document. Ultimately, we argue that this competitive computational culture continues to provide a form of discursive power in what is now being called the "AI race" [55] between major technology companies such as Google and Microsoft, especially in their chase for more powerful generative AI systems.

2 BENCHMARKING PRACTICES

Benchmark datasets in machine learning aren't rendered "benchmarks" from their inception. Typically, they begin as just another dataset that formalizes a certain technical challenge or task into a set of input and output pairs, which can then be used to evaluate the performance of a particular ML model. This arrangement allows for the production of "accuracy scores" representing the performance of the ML model on the particular task represented by the dataset. In the context of a platform like Kaggle, these scores are then compiled into public-facing leaderboards that serve to represent "technical progress" on the denoted challenges. In this way, the ambiguous concepts of "task," "performance," and "progress" are translated into tangible material artifacts: the dataset, the accuracy score, and the leaderboard.

Denton et al. [14] show in their genealogical tracing of ImageNet, the de facto benchmark for computer vision, that the transition from 'dataset' to 'benchmark' is often an informal and mercurial process. For instance, a team developing a model may select a particular dataset – out of convenience, availability, their own subjective preferences etc. – to test the performance of their model on the task represented by the chosen dataset. If this model becomes successful or highly cited, future teams developing similar models that seek to compare and establish "state-of-the-art" (SOTA)[1] performance are expected to use the same dataset for evaluation. As such, certain benchmarks become widely cited and circulated within the ML subcommunities that form around particular tasks, with these datasets being viewed as rigorous yardsticks due to being "*implicitly community vetted*" by frequent use and impressive citation counts [41]. As Jaton [26] documents in his ethnography with a team of researchers developing a new ground truth dataset for saliency detection in images, more challenging or differently arranged datasets are often developed that may address the limitations of existing benchmarks. Despite this, these new attempts still often have to position themselves alongside the existing benchmarks and often exist as supplementary to the original highly cited datasets that continue to be used as standards for the community [48]. These practices of "peer-washing" naturalize benchmarks and their limitations as authoritative proxies for specific tasks or domains [41]. Indeed, such is the character of standards: they "have significant inertia and can be very difficult and expensive to change" [8, p. 14].

This concretization of standardized benchmarks *within* particular ML subcommunities also results in the adoption of these benchmark datasets *across* subcommunities, resulting in these datasets being reframed and adopted for tasks beyond their original domain [30]. This phenomenon is exacerbated by the fact that creating a dataset can be a task of significant human, financial, and temporal investment [42], which renders their revision and construction a particularly inaccessible endeavor. For instance, a dataset designed for evaluating e-commerce recommendation algorithms is commonly used for evaluating the 'sentiment' of written reviews [35]. In fact, Koch and colleagues [30] found that over 70% of benchmark datasets used in prominent computer vision papers were appropriated from datasets that were originally developed in different domains.

Existing critical ML research has highlighted that these moments of naturalization and uncritical adoption occlude the inherent limitations of benchmark datasets [14, 41]. As proxies for the tasks they're meant to evaluate, benchmark datasets are never complete nor can they ever amount to a comprehensive representation of reality [12, 27, 28, 48]. Rather, benchmark datasets provide one perspective on examples that are deemed most pertinent, defined by the

[1] State-of-the-art (SOTA) performance by a machine learning model refers to the "correct" prediction of the outputs of a popular benchmark. This is generally taken to indicate technical progression in the field [48]

instances collected in the dataset. This means that the performance of a model on that benchmark is not necessarily representative of a broader more generalized "domain" such as 'computer vision,' but a more specific and localized "task" such as 'figure-ground distinction' that is defined in direct relation to the arrangement of the dataset. While benchmarks can be useful apparatuses for evaluating ML models, their practical use in existing ML practice renders them as normative instruments that perpetuate particular epistemological perspectives about how the world is ordered. Indeed, the grandiose rhetoric that accompanies popular benchmarks such as ImageNet and GLUE matches the obsession with scale that has come to symbolize the AI community. ImageNet, for instance, claims to "map out the entire world of objects" [14], while GLUE presents itself as a benchmark for evaluating models on their "General Language Understanding" of the English language [57]. The large scale and scope of benchmark datasets, however, do not necessarily mean they represent meaningful technological progress in AI research. As Raji et al. [48] have emphasized, benchmarks that claim to be all-encompassing occlude the inherently closed worlds and humble capabilities of these artifacts.

In addition to concerns around construct validity – i.e., whether a dataset is actually representative of the task it serves as a proxy for – the culture of competitive benchmarking has also started to receive pushback regarding the reliability of accuracy metrics as a proxy for *technical* performance. As benchmark datasets are more frequently used and circulated, models may be able to exploit statistical patterns within the data. As such, models produce impressive accuracy scores on certain benchmarks while being unable to solve a simple problem that is not found within the dataset [60]. In a process known as overfitting, these models essentially 'game the system' while not actually improving their technical performance at a given task. For instance, a dataset commonly used to train large language models (LLMs) was found to include evaluation examples from GLUE, which would mean that the LLM would perform exceptionally well on that benchmark thus inflating the perceived capability of the model [47]. Overfitting also obscures any meaningful progress that can be made on a given benchmark as it is difficult to distinguish between models that have solved a problem "correctly" versus those that have exploited statistical shortcuts.

Despite these epistemological and technical limitations, benchmarking remains the primary apparatus for evaluating ML systems and technical progress within the field. In the sections that follow, we highlight three historical moments in the development of machine learning as a field of research that have been formative in the transformation of benchmarking from processes of informal standardization to naturalized and institutionally accepted metrics of success. In so doing, we document a "history of the present" [15, p. 31] of how the benchmark dataset became central to the practice of machine learning, and how these dynamics are shifting – or not – in the current era of generative AI.

3 FORMATION OF A COMMUNITY: HIGHLEYMAN'S DATASET (1960s)

The late 1950s and early 1960s was an influential period in the early cultural history of machine learning. This was the era of pattern recognition, in which the prime focus among ML practitioners and researchers was optical/object character recognition (OCR) – i.e., the automatic recognition of handwritten letters and numbers by digital computers.[2] Multiple labs around the United States proposed computational methods for translating handwritten alphanumeric characters into machine code, but these methods had only been evaluated by their creators and not by other researchers [23], which thus raised questions about scientific replicability. Indeed, without agreed-upon evaluation metrics or standards, there was little consensus within this community of researchers as to which methods were superior.

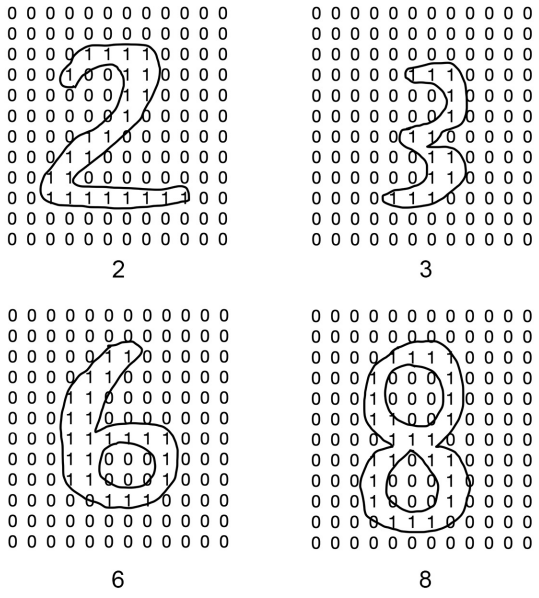

Figure 1: Examples of the quantized forms of hand-printed numbers. Adapted and redrawn by authors from [21, p. 1511]

To address this, two researchers from Bell Labs, Highleyman and Kamentsky, created a scanner that translated handwritten symbols into punch cards to be read by a computer. They recruited 50 participants to handwrite the 26 letters of the alphabet and ten digits, which resulted in a dataset of 1800 alphanumeric characters rendered onto punch cards. In so doing, they hoped to "facilitate a systematic study of character-recognition techniques and an evaluation of methods prior to actual machine development" [23, p. 291]. Indeed, in a paper published 3 years later, Highleyman [20] was able to evaluate the OCR techniques proposed by the different teams of researchers using this dataset as a common framework for analysis.

Following their analysis, several labs around the US requested a copy of the dataset to develop and test novel methods. Highleyman [22] thus offered to mail the dataset to anyone who requested it "for the nominal charges of reproduction and shipment" [p. 136].

[2] See Mendon-Plasek [37] for a comprehensive examination of this period.

He elaborated that "Since it appears that this data is being used commonly, it may, therefore, serve as an unintended, incomplete, yet interesting, available and temporary *standard* by which workers in the field may compare their results with those of others [emphasis added]" [p. 136]. It is important to note, here, that Highleyman was clear in identifying the limitations of the dataset, stating that a formal standard for evaluating models should be "well thought out and certainly more complete than this data of mine" [p. 136].

Despite these shortcomings in design, the *sharing* of the dataset itself was decidedly useful. "Many published works tend to be ambiguous as to the quality or sources of their data" [p. 136], so a shared dataset allowed for a level of control in the evaluation of methods. Indeed, by standardizing evaluation, Highleyman and Kamentsky were able to identify methodologies that reported accuracy scores that didn't align with their own evaluations. For instance, while a method proposed by Bledsoe and Browning [7] recorded an accuracy of 78.4% in recognizing hand-printed characters, Highleyman and Kamentsky's [24] duplication of their methods on their dataset achieved only 19.6% accuracy (see also [49]). Though there was disagreement and controversy regarding the success of individual methods, it was only by evaluating models on Highleyman's data that these techniques gained legitimacy. As such, the constructed accuracy scores, which were inextricably linked to Highleyman's dataset, became community-accepted proxies for models' capabilities. The standardization of pattern recognition evaluation allowed these metrics to be taken as "matters of fact" [51].

These metrics also gained legitimacy through the public record and reproduction. As Shapin and Schaffer [51] outline in their canonical text *Leviathan and the Air-Pump*, the development of the scientific method as a means of producing scientific knowledge was intertwined with its visibility. The ability of members of the scientific community to witness experiments allowed them to verify the findings and conclusions. In instances in which physical witnessing was not possible, extensive documentation of methods undertaken, as well as candid admissions of missteps, served as a proxy or "virtual witnessing" of these experiments. Contemporary peer-reviewed scientific research papers can be seen as modern iterations of this kind of virtual witnessing. These practices of visibility were crucial for the collective agreement and understanding of experimental phenomena, wherein clear documentation of experimental practices encouraged the perception that these experiments could be replicated and thus valid [51]. In the case of the pattern recognition researchers, Highleyman's sharing of the dataset in 1963 constituted the first time that a community of researchers could experiment separately and legitimately compare their results. It addressed the previous problem of scientific replicability, which then solidified the idea of the "benchmark dataset" as a central object and viable scientific tool that made model evaluation across time and space possible in the OCR community.

The circulation of Highleyman's dataset was not only instrumental in forming a scientific community of knowledge production, but also in establishing specific methods for the evaluation of machine learning models. Indeed, a consensus regarding the comparison of models necessitates both an agreed apparatus for analysis (the dataset) *as well as* established practices for making use of this tool [51]. The legitimacy of scientific findings is grounded in a knowledge community's understanding and belief in the methods of the experiments. Perhaps most influentially, in replicating his method on Highleyman's dataset, Bledsoe [6] trained his model on the characters of 40 writers, leaving the characters written by the remaining ten writers to later test the model. This is likely the first documented instance of a 'training-test split' in a dataset, which is now of course an indispensable practice within benchmarking in machine learning [49]. The development of standardized and community-accepted practices for evaluation allowed for the competitive culture of model testing to emerge, which would soon become synonymous with technical progress.

The story of Highleyman's dataset demonstrates the utility of a standardized artifact around which a scientific community of knowledge production can form. It is also exemplary, however, of the arbitrariness and haphazardness of how a dataset *becomes* a benchmark, as well as how that arbitrariness is easily forgotten once it becomes accepted or "usable" as a standardized tool for comparison. Indeed, as Mulvin [38] writes, "Standardization is a process of forgetting" [p. 6], in which we begin to lose the communal, intentional, institutional, practical, and material work that goes into turning an arbitrary "thing" into a standard. Highleyman himself explicitly emphasized the limitations and incompleteness of the dataset, but a synthesis of the dataset's early introduction, availability, and eventual adoption allowed the OCR community to overlook those limitations and use it as a de facto benchmark for character recognition. In this way, its significance as a benchmark was not necessarily "technical," but *cultural*: it allowed a collaborative-competitive community to form and knowledge to be produced through a standardized way of comparison and *doing science*.

4 STANDARDIZATION: DARPA'S COMMON TASK FRAMEWORK (1980s)

The trajectory of AI research has also been profoundly shaped by the ebbs and flows of public interest and funding. Despite recent surges in AI headlines and claims of innovation, the field has historically experienced seasonal funding cycles oscillating between periods of fervent growth and advancements known as "AI summers," and disillusionment and skepticism, known as "AI winters."

The first AI winter occurred in the 1970s following a series of overhyped state-funded AI projects that failed to meet their lofty expectations [54]. During the 1960s, the Defense Advanced Research Projects Agency (DARPA) provided grants of millions of dollars to AI research labs across the US, often with few restrictions or accountability requirements such as the need for timelines or justifications [54]. Similar to the excitement around AI today, the hype surrounding AI research created immense public expectations regarding the possibilities of the technologies; the hype was similarly matched with exaggerations and false predictions by experts which further inflated expected capabilities. As John Pierce [44], inventor of the transistor and Vice president of Bell Labs, argued, AI researchers acted "not like scientists, but like mad inventors or untrustworthy engineers". Immense time and financial resources were spent on "grandiose aims" [33], resulting in negligible returns. Put simply, "To sell suckers, one uses deceit and offers glamour,"

encapsulating the prevailing sentiment of the time [44]. The AI winter of the 1970s was thus not just a crisis of funding, but also a crisis of trust and credibility.

Despite the eroding trust in AI research during this time, there was a resurgence of funding by DARPA only a decade later in 1986. This can largely be attributed to DARPA project manager Charles Wayne, who introduced the Common Task Framework (CTF), "a virtuous cycle involving shared objectives, data and evaluations" [32, p. 1]. This framework was characterized by detailed task definitions and explicit, quantitative success metrics, wherein evaluations of AI technologies were conducted by a 'neutral' entity, the National Institute for Standards and Technology (NIST). In a marked departure from the earlier ad-hoc approach epitomized by Highleyman's dataset, NIST created and distributed test datasets specifically with the purpose of evaluating AI systems on specific tasks at scale.

The introduction of centralized and standardized performance evaluations was a strategic response to the aforementioned mistrust in AI research that brewed during the seventies. Indeed, a program with clear objectives and metrics for success was seen as the only viable way to attract and justify funding [31]. Many government agencies, however, remained skeptical of the promises of AI technologies asserting that "You can not turn water into gasoline, no matter what you measure" [11, p. 3], while AI practitioners were similarly unhappy about the loss of freedom: it was like "being in first grade again—you are told exactly what to do, and then you are tested over and over" [11, p. 3]. Despite this lackluster response from both funders and engineers, the centralized effort to standardize and measure progress provided by the CTF succeeded in providing a trusted framework for evaluating AI systems.

Positioned as "objective performance evaluations" [32], CTF measures became trusted proxies to represent a system's performance and general progression against standardized goals. Unlike in the OCR case in which benchmarking was used to construct internal validity primarily amongst the research community, the CTF made the technical work of AI development also accessible to non-experts. These metrics functioned "to create a basis for mutual accommodation in a context of suspicion and disagreement" between AI researchers and funders [45, p. 149]. Funders no longer had to rely on engineers' accounts to ascertain whether the research was on track, while researchers had quantified goals to progress towards to maintain funding, as well as to compare their progress to other labs within their field. This shift effectively delegated critical funding decisions from individuals to quantifications [45].

The standardization of benchmarking culture in AI research also made regular self-evaluation the norm within research labs. Indeed, researchers began evaluating their systems every hour on set-aside evaluation datasets [11]. This created a culture of "algorithmic hill-climbing" in which researchers could make steady and incremental progress toward their goals, allowing for cumulative advancements [32, p. 27]. This self-regulatory practice became so ingrained in the research methodology that some labs joined the benchmarking groups even without DARPA funding [11]. In 1992, DARPA and NIST cemented this culture of collaborative evaluation with the inception of the Text Retrieval Conference (TREC), an annual series of workshops focusing on common language-based tasks and datasets. This initiative marked a significant milestone in the evolution of ML research, establishing a communal culture of collaboration and comparison that would become a hallmark of the field.

As this case illustrates, what initially started as a centralized strategy to secure funding and legitimize the scientific community around AI research ultimately developed into an indispensable cultural and technical component of how progress and performance for AI systems are measured [48]. Indeed, the development of benchmarking in AI research, and specifically the CTF, is not just a feature of technical progress, but also a story of how a field adapts to external pressures and internal aspirations. As Porter [45] underscores, "the bureaucratic imposition of uniform standards and measures has been indispensable for the metamorphosis of local skills into generally valid scientific knowledge" [p. 21]. In the case of AI benchmarking, such standardized processes and metrics allowed a discipline once marred with skepticism and hyperbole to redefine itself as rigorous and trustworthy.

5 COMPETITION: THE NETFLIX PRIZE (2000s)

The Netflix Prize, launched in 2006, marked a significant shift in the culture of competition within machine learning, moving away from bureaucratic, government-led evaluations to industry-driven challenges. The Prize was a strategy to address the hurdles encountered in improving Netflix's in-house movie recommendation algorithm, Cinematch. Netflix, still a DVD rental service at the time, asked users to rate films on a traditional 5-star scale. It would then recommend films to its users to rent based on their previous ratings. Netflix viewed its future success as being largely dependent on how well this system could successfully recommend relevant content to its user base. Despite Cinematch generating approximately 30 billion predictions a day [9], Netflix engineers faced stagnation in enhancing the effectiveness of the system. To address this challenge, Netflix introduced a public competition in October 2006, offering $1 million for a model that could surpass Cinematch's accuracy by 10%.

The cornerstone of the competition was the release of a massive dataset, containing 100 million movie ratings of 17,770 movies from 480,189 customers, collected between October 1998 and December 2005. Netflix claimed that the dataset was representative of the overall distribution of user ratings [5] and that personal information had been removed from this data such that individuals could not be identified [39]. Three million user ratings were set aside for evaluating models submitted by participants. The aim here was for Netflix to implement the successful design and thus improve its recommendations to maintain a satisfied customer base.

The competition attracted an unprecedented level of global participation, drawing over 50,000 participants from 186 countries [9]. Approximately 40,000 teams, composed of PhD students, established academics, garage hobbyists, and technologists, were formed in an endeavor to claim the prize. Beyond the obvious financial draw, researchers were attracted by the wealth of data that had been made available by Netflix. In 2007, data of this scale was rarely available outside of proprietary settings. As Jackson [25] states, Netflix CEO Reed Hastings "was a tech-age Willy Wonka letting any curious hacker into his digital Chocolate Factory." Participants took this opportunity to explore using machine learning techniques the insights available from large magnitudes of data.

Models submitted for the competition were evaluated against two test sets: one accessible to researchers and a private one retained by Netflix to prevent tailoring models to the specific test data. Accuracy scores were calculated for each model and displayed on a publicly accessible leaderboard that highlighted the time of submission and progress toward the 10% improvement target. Progression on this leaderboard, of course, soon became the defining measure of success, in which the ultimate goal transformed into developing the top-ranking model on the leaderboard. It was a centralized scale and public endorsement of competition among participants, through which they could measure their progress in real-time. In the final hours of the competition, participants reported eagerly refreshing the leaderboard to check their standings [4].

The Netflix Prize illustrates the naturalization and universalization of competition through leaderboards and accuracy metrics [36, p. 6]. As Mau [36] explains, "In many fields, quantification practices are actually responsible for the enactment of competition, of a kind that revolves around numbers" [p. 6]. In this context, numbers and rankings become the dominant language for evaluating both societal and technical progress. Indeed, the Netflix Prize leaderboard didn't just track progress, it shaped the very conception of success, prioritizing numerical accuracy scores over all other qualitative and arguably more practical aspects of evaluation. For instance, on June 26, 2009, BellKor's Pragmatic Chaos, a team led by AT&T researchers, finally reached the 10% improvement threshold and were crowned the winners of the competition. However, despite the million-dollar payout, Netflix never actually implemented the team's solution. This was primarily because BellKor's model was too complex. It consisted of 104 algorithms created by multiple groups and strung together by a single neural network [10]. As Amatriain and Basilico [3] explain, "the additional accuracy gains that [Netflix] measured did not seem to justify the engineering effort needed to bring them into a production environment." In this way, while benchmarking may measure progress at one specific goal and the leaderboard may indeed be an accurate representation of progress on that goal, neither account for the additional resources necessary to construct these models.

Today, large multimodal AI models continue to be products of immense human labor, costing millions of dollars to train, a tremendous cost that is alleviated only through the exploitative strategy of tech companies like OpenAI using underpaid workers from the Global South [43, 58]. In addition to the necessary human labor required to develop these models, there is also a tremendous amount of computing power that enables their development, which also dramatically expands the carbon footprint of AI research [52]. Focusing primarily on benchmark performance and leaderboard rankings thus flattens these human and environmental costs and quantifies progress through singular metrics for success. Indeed, embedded within these leaderboards is the implicit acceptance of the criteria underpinning quantitative rankings and their resulting outcomes, which is that higher results – i.e., winning the competition – equal better performance and progress [36].

The Netflix competition also highlights how datasets are inextricably tied to their contexts of production, which means they can often become measures of irrelevant tasks that aren't representative of shifting expectations in their real-world application. For instance, when BellKor's team finally reached the threshold for victory in 2009, not only was their model financially impractical to construct, but the state of movie consumption was also very different from when the competition began. Netflix launched its immensely popular streaming service in 2007, one year after the competition started, and this substantially changed the types of data that Netflix had access to. By 2009, Netflix no longer had to rely on self-reported ratings for insights into viewers' watching habits. Netflix now collected granular data such as when users start, pause or rewind videos, whether they completed a video and went on to watch another, and the time of day they consumed content. This shift in data available to the company redefined the object of prediction: predicting consumption patterns became much more essential than predicting self-ratings [3]. The task of content recommendation thus transformed into approximating which titles could keep users engaged on a platform, as opposed to how they might rate that content. As Hallinan and Striphas [18] articulate, this paradigm shift of data usage and prediction objectives underscores how the value and relevance of predictive models are ultimately dependent on how well they actually represent a dynamic real-world problem. In this way, as Raji et al. [48] have critiqued for ImageNet and GLUE, construct validity is a major variable in determining a benchmark's relevancy that is not necessarily captured in the competitive epistemologies of leaderboards and benchmarking cultures.

Taken together, the stories of Highleyman's dataset, the Common Task Framework, and the Netflix Prize trace a genealogy of how a field of research was made scientific through an epistemology grounded in competition, for which community and standardization served as the foundational building blocks. While the Netflix Prize was foundational for establishing a culture of competition that persists in contemporary AI research through the material artifact of centralized leaderboards, it also highlighted the complexities and limitations of competitive benchmarking as a measure of technological progress. Because accuracy metrics and leaderboards have become the primary means through which capabilities are evaluated and technological progress communicated, they are exploited for the illusion of progress that is taken as an indication of technological superiority.

Benchmarking thus risks a "thinning" of the world [46], neglecting elements that cannot be easily measured or analyzed numerically. Indeed, this history sheds light on the enticing nature of competitive science – made possible through the constellation of artifacts such as benchmark datasets and leaderboards – and the often problematic ways in which the quantitative gamification of technological progress can lead practitioners astray from more pressing and practical modes of intervention. In the final section, we situate this genealogical analysis and the shifting dynamics of benchmarking in the contemporary era of generative artificial intelligence.

6 COMPETITIVE BENCHMARKING IN THE ERA OF GENERATIVE AI

Traditional benchmarking practices have faced new challenges with the advent of generative AI systems such as OpenAI's GPT 1-4 and Google's Gemini. Like the circulation of Highleyman's dataset among OCR researchers, publicly accessible generative AI systems

have fueled a groundswell of community-driven experimentation and evaluation. Yet, unlike conventional ML models in which accuracy is measured against a correct label, generative AI systems operate in a realm where there is often no single 'right' answer [27]. Indeed, these systems are tasked with generating content that must be coherent and contextually relevant across varied prompts and situations.

The subjective nature of evaluating outputs such as text and images further complicates the matter. Users' artistic interpretations or specific contextual needs play a significant role in determining the quality of a generative AI model's output, making the evaluation process inherently subjective. For instance, a popular online application of the AI image generator Midjourney has been to produce new styles of popular franchises such as "Harry Potter by Balenciaga" [1] or "The Great Hogwarts Rave of 1996" [40]. While the results showcased online are undoubtedly impressive, it is important to also note that the impressiveness of these models is qualitatively different from what might have been considered state-of-the-art in traditional ML models. The extraordinary quality of 'DJ Voldemort' depicted in the "The Great Hogwarts Rave of 1996" collection [17], for example, is not necessarily a function of scientific accuracy, but rather creative interpretation – it is not that this particular black and white image of the 'dark lord' is 'accurate' but 'creative' that makes it exciting. Indeed, the character could have easily been portrayed as an outdoor festival DJ with a backdrop of colorful neon lights and fireworks, as opposed to the more grungy underground techno DJ depicted, and viewers could presumably have been equally as captivated. This expectation of diversity in responses adds an additional layer of complexity: a generative AI system is expected to produce varied images even when given the same prompt.

This qualitative aspect of evaluation has now become integral to the development and training of these systems, and while some generative AI systems such as Midjourney and Google's MusicLM engage users in the evaluation process by allowing them to select and rank outputs, such evaluation datasets often remain proprietary due to competitive pressures in the industry. In light of these challenges, we ask: how then does one systematically evaluate the performance of a generative AI model? The challenge of evaluating generative AI outputs echoes the shift during the AI winter of the 70s, wherein qualitative judgments of 'usefulness' and social impact began to complement pure accuracy metrics.

Despite these challenges of comprehensively benchmarking generative models, researchers rely on benchmarks to communicate progress of generative models. This was particularly highlighted by the launch of Google's family of multimodal models, Gemini, which underscored the continued relevance of benchmarking. According to the introductory paper accompanying Gemini, this system was evaluated against over 50 benchmarks as a "holistic harness" to assess its capabilities [53], and it explicitly claims state-of-the-art performance on 30 benchmarks covering diverse domains spanning image and video understanding, audio processing, coding, reading comprehension, math, and machine translation tasks. This declaration of progress in the AI race [55], and the claim of outperforming OpenAI's GPT-4, underscores the continued reliance on benchmarking to communicate advancements in AI.[3]

Of particular note is that the Gemini team [53] reports state-of-the-art performance on the Massive Multitask Language Understanding (MMLU) benchmark [19], surpassing 'human performance' for the first time. Composed of over 57 subject areas across STEM, the humanities, the social sciences, and more, this benchmark aims to evaluate knowledge acquired during pretraining in zero-shot and few-shot settings, including professional exams. While it is contestable that standardized tests such as the Graduate Record Examinations (GRE) or the Law School Admission Test (LSAT) are effective predictors of an individual's potential to be a successful graduate student or a lawyer, they have been undeniably effective in producing hype for generative large language models (LLMs) such as ChatGPT: "GPT-4 beats 90% of humans in world's toughest exam" [34]. Here we are reminded again of Pierce's [44] words regarding hype cycles of automated technologies: "To sell suckers, one uses deceit and offers glamour." In this case, however, benchmarking statistics are used to generate hype, interest, and investment.

The MMLU benchmark thus reflects current trends of employing standards typically used to evaluate the performance of human actors on machines. As Wong [59] reports, however, this conflating of human intelligence with machine intelligence, which is often (problematically) framed as a step towards what has been loosely referred to as 'Artificial General Intelligence' (AGI), can be understood as a "really sophisticated PR" strategy that "makes the product seem more powerful." Indeed, it is not necessarily that professional exams or other human benchmarks themselves indicate intellectual prowess, but rather, that the same benchmarks can be used for machines and humans that serve as the technological spectacle through which companies such as OpenAI and Google can market their technologies. The emphasis on outperforming benchmarks on standardized tests risks a similar trap to the Netflix Prize, where a narrow focus on metrics can overshadow real-world utility and broader social implications.

Large releases like Gemini are significant within machine learning research as they come to shape the evaluation of the next generation of systems. As mentioned, for future systems to claim innovation and progress within the field, they are expected to be evaluated against the same benchmarks as previous state-of-the-art models. As such, the Gemini release may entrench the power of benchmarking within the current generation of generative AI systems.

The emergence of generative AI models has also prompted a reevaluation of the sorts of qualities that should be reflected in future benchmarks. Given the influential role of benchmarks in standardizing goals and driving competition, the selection of benchmark datasets and tasks becomes critical. For instance, Samsi et al. [50] propose benchmarks to measure the environmental impact of AI systems, focusing on the energy costs of performing prompts on

[3]Academic researchers have also utilized benchmark datasets for systematic comparisons between models like OpenAI GPT and Google Gemini [2] The Gemini team [53], however, acknowledges the possibility of data contamination in which benchmarking data is included within large-scale scraped training datasets, which would influence results. Concerns around contamination will remain an ongoing challenge of evaluating AI systems that rely on scraped and uncurated training datasets.

various language models. Similarly, Guo et al. [16] evaluate LLMs on a suite of benchmarks including testing ethical alignment, bias, toxicity, truthfulness, and safety, expanding the scope beyond traditional knowledge and capability testing. These initiatives suggest a shift towards more holistic and socially responsible benchmarking practices in AI research. More importantly, however, they also consider whether such factors can be adequately captured by numerical metrics, to be then co-opted by the established dynamics of competition. While we underscore the need for new forms of assessment to encompass the scope of impacts of emerging generative systems, we also warn against the totalizing frame of competition that quantitative benchmarks produce. As Vincent [56, p. 290] argues, "an obsession with measurement above all else will distort, distract and destroy what we claim to value."

Given the nascency of generative AI research, it is still too early to holistically grasp how benchmarking practices and cultures will shift. As is evident with the communication accompanying the release of Google's Gemini, however, benchmarking will likely continue to serve as a framework through which technological and scientific progress is articulated. Here, it will be even more critical to query whether this competitive epistemology is truly conducive to this goal. At a time in which Artificial Intelligence is becoming increasingly intertwined with traditionally subjective areas such as creativity, art, style, and taste, the quantitatively-focused methodologies of competitive benchmarking and evaluation metrics will inevitably require some sort of transformation. Whether this transformation will be formal, discursive, or conceptual is still up for question, which is why we believe this marks a pivotal moment for critical scholars, researchers, and practitioners to think collectively and creatively about future directions.

7 CONCLUSION

The historical evolution of benchmarking in AI research underscores a critical trajectory from the arbitrary beginnings of shared datasets to the institutionalized and competitive benchmarks that currently dominate the field. This development, as marked by Highleyman's dataset in the 1960s, the Common Task Framework in the 1980s, and the Netflix Prize in the 2000s, reveals the interweaving of scientific community formation, standardization of evaluation methods, and a shift towards a competitive culture in AI research.

Initially, benchmark datasets, exemplified by Highleyman's OCR dataset, played a key role in community building among AI researchers. This period highlighted the importance of shared datasets in fostering collaboration and comparative analysis within the scientific community, despite their arbitrary and unintended standardization. The introduction of the CTF by DARPA marked a significant shift towards more structured benchmarking practices, which was instrumental in establishing legitimacy and trust in AI research. However, this also ushered in a culture where quantitative metrics became the primary, sometimes sole, criterion for assessing AI advancement. The Netflix Prize further entrenched this competitive benchmarking culture, emphasizing numerical accuracy, leaderboards, and winners. While the competition showcased the potential of crowdsourcing and global participation in AI research, it also highlighted the disconnect between progress on benchmarks and practical utility.

The sporting world has also experienced the tensions that strict evaluation parameters pose. During a 1981 one-day cricket match, New Zealand, needing a six to tie a cricket match, was denied the chance when Australia's bowler, Trevor Chappell, rolled the ball along the ground for the final bowl of the match. While this bowl was technically legal, it was considered not within the spirit of the game, and underarm bowling was subsequently banned. This event highlights how strict adherence to rules, or benchmarks, can sometimes undermine the spirit of fair play and innovation. In AI benchmarking there is a risk that developers might overly optimize models to perform well on specific benchmarks without genuinely advancing the technology in meaningful or responsible ways. This "gaming" of the system can result in AI models that perform exceptionally on benchmark tests but fail to address real-world complexities, much like the technically legal but widely criticized underarm bowl. As AI has increasingly adopted dynamics akin to competitive sports, such transgressions are likely to occur. As the parameters of sports adapt, so too much benchmarking.

In the contemporary era of generative AI, these traditional benchmarking practices are encountering new challenges. The subjective nature of outputs from generative models necessitates reevaluating what constitutes a meaningful and representative benchmark. The focus on numerical metrics and leaderboard rankings, while useful for certain comparisons, overlooks broader considerations such as societal impact, ethical concerns, and practical applicability. Reflecting on the history of competition within AI research, we highlight how benchmarking has considerable influence on the *types* of progress that are made in AI research. We thus cannot confidently state whether this means existing benchmarking practices should be reformed to fit the changing demands of generative AI, or if an entirely new paradigm of AI evaluation must be established. Such is the resilient nature of standards: they are not only resistant to change, but also nearly impossible to imagine a world without. Our view is one that understands benchmarking as a standardized evaluation process that prioritizes actionability and competition over construct validity, and therefore a standard that will increasingly become less relevant to the inherently more interpretive outputs of generative AI systems, with far-reaching social impacts. As such, we argue that generative AI prompts a need for a substantial shift towards more holistic, inclusive, and socially responsible evaluation practices that transcend the competitive epistemologies of traditional quantitative benchmarking, to encompass the variability and social impact of these systems.

Finally, we also emphasize that these cases do not represent the entire history of AI benchmarking, but rather present compelling moments that shaped the competitive dynamics of AI benchmarking today. Importantly, our analysis may have overlooked crucial histories of domain-specific evaluation and areas of AI research that do not fit neatly into competitive benchmarking paradigms such as those related to AI voice synthesis or music generation systems. Further research is needed to examine the varied impacts of benchmarking across different fields of AI, including those where community-driven or collaborative models of evaluation may offer alternative insights into the development and deployment of AI systems.

REFERENCES

[1] Zeeshan Ahmed. 2023. Harry Potter by Balenciaga: AI-generated Future of Entertainment? | The Express Tribune. https://tribune.com.pk/story/2409540/harry-potter-by-balenciaga-is-this-the-ai-generated-future-of-entertainment.

[2] Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. 2023. An In-depth Look at Gemini's Language Abilities. https://doi.org/10.48550/arXiv.2312.11444 arXiv:2312.11444 [cs]

[3] Xavier Amatriain and Justin Basilico. 2012. Netflix Recommendations: Beyond the 5 Stars (Part 1). https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429.

[4] Robert Bell, Jim Bennett, Yehuda Koren, and Chris Volinsky. 2009. The Million Dollar Programming Prize. https://spectrum.ieee.org/the-million-dollar-programming-prize.

[5] J. Bennett and S. Lanning. 2007. The Netflix Prize. In *Proceedings of the KDD Cup Workshop 2007, page 3–6. New York, ACM.* Association for Computing Machinery, New York, NY, USA, 3–6.

[6] W. W. Bledsoe. 1961. Further Results on the N-tuple Pattern Recognition Method. *IRE Transactions on Electronic Computers* EC-10, 1 (March 1961), 96–96. https://doi.org/10.1109/TEC.1961.5219162

[7] W. W. Bledsoe and I. Browning. 1959. Pattern Recognition and Reading by Machine. In *Papers Presented at the December 1-3, 1959, Eastern Joint IRE-AIEE-ACM Computer Conference (IRE-AIEE-ACM '59 (Eastern)).* Association for Computing Machinery, New York, NY, USA, 225–232. https://doi.org/10.1145/1460299.1460326

[8] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences.* MIT Press, Cambridge, MA.

[9] Eliot Van Buskirk. 2009. BellKor's Pragmatic Chaos Wins $1 Million Netflix Prize by Mere Minutes. *Wired* (2009).

[10] Mita Chaturvedi. 2021. How Useful Was The Netflix Prize Really? https://analyticsindiamag.com/how-useful-was-the-netflix-prize-really/.

[11] Kenneth Ward Church. 2018. Emerging Trends: A Tribute to Charles Wayne. *Natural Language Engineering* 24, 1 (Jan. 2018), 155–160. https://doi.org/10.1017/S1351324917000389

[12] Kate Crawford and Trevor Paglen. 2019. Excavating AI: The Politics of Training Sets for Machine Learning. https://excavating.ai.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition.* Institute of Electrical and Electronics Engineers, Piscataway, NJ, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[14] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet. *Big Data & Society* 8, 2 (July 2021), 20539517211035955. https://doi.org/10.1177/20539517211035955

[15] Michel Foucault. 1991. *Discipline and Punish: The Birth of the Prison* (reprint ed.). Penguin Books, London.

[16] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating Large Language Models: A Comprehensive Survey. https://doi.org/10.48550/arXiv.2310.19736 arXiv:2310.19736 [cs]

[17] Ed Haas. 2023. The Great Hogwarts Rave of 1996 | Facebook. https://www.facebook.com/groups/cursedaiwtf/permalink/1407296606545485.

[18] Blake Hallinan and Ted Striphas. 2016. Recommended for You: The Netflix Prize and the Production of Algorithmic Culture. *New Media & Society* 18, 1 (Jan. 2016), 117–137. https://doi.org/10.1177/1461444814538646

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. https://doi.org/10.48550/arXiv.2009.03300 arXiv:2009.03300 [cs]

[20] W. H. Highleyman. 1962. The Design and Analysis of Pattern Recognition Experiments. *The Bell System Technical Journal* 41, 2 (March 1962), 723–744. https://doi.org/10.1002/j.1538-7305.1962.tb02426.x

[21] W. H. Highleyman. 1962. Linear Decision Functions, with Application to Pattern Recognition. *Proceedings of the IRE* 50, 6 (June 1962), 1501–1514. https://doi.org/10.1109/JRPROC.1962.288194

[22] W. H. Highleyman. 1963. Data for Character Recognition Studies. *IEEE Transactions on Electronic Computers* EC-12, 2 (April 1963), 135–136. https://doi.org/10.1109/PGEC.1963.263427

[23] W. H. Highleyman and L. A. Kamentsky. 1959. A Generalized Scanner for Pattern- and Character-Recognition Studies. In *Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference (IRE-AIEE-ACM '59 (Western)).* Association for Computing Machinery, New York, NY, USA, 291–294. https://doi.org/10.1145/1457838.1457894

[24] W. H. Highleyman and L. A. Kamentsky. 1960. Comments on a Character Recognition Method of Bledsoe and Browning. *IRE Transactions on Electronic Computers* EC-9, 2 (June 1960), 263–263. https://doi.org/10.1109/TEC.1960.5219829

[25] Dan Jackson. 2017. The Netflix Prize: How a $1 Million Contest Changed Binge-Watching Forever. https://www.thrillist.com/entertainment/nation/the-netflix-prize.

[26] Florian Jaton. 2021. *The Constitution of Algorithms: Ground-Truthing, Programming, Formulating.* The MIT Press, Cambridge, Massachusetts.

[27] Edward B Kang. 2023. Ground Truth Tracings (GTT): On the Epistemic Limits of Machine Learning. *Big Data & Society* 10, 1 (Jan. 2023), 20539517221146122. https://doi.org/10.1177/20539517221146122

[28] Edward B. Kang. 2023. On the Praxes and Politics of AI Speech Emotion Recognition. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).* Association for Computing Machinery, New York, NY, USA, 455–466. https://doi.org/10.1145/3593013.3594011

[29] Jules King, Perpetual Baffour, Scott Crossley, Ryan Holbrook, and Maggie Demkin. 2023. LLM - Detect AI Generated Text. https://kaggle.com/competitions/llm-detect-ai-generated-text.

[30] Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).* Association for Computing Machinery, New York, NY.

[31] Mark Liberman. 2015. Reproducible Research and the Common Task Method. https://www.simonsfoundation.org/event/reproducible-research-and-the-common-task-method/.

[32] Mark Liberman and Charles Wayne. 2020. Human Language Technology. *AI Magazine* 41, 2 (June 2020), 22–35. https://doi.org/10.1609/aimag.v41i2.5297

[33] James Lighthill. 1972. Lighthill Report. http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm.

[34] Livemint. 2023. 90% in Bar Exam, 88% in LSAT: GPT-4 Beats 90% of Humans in World's Toughest Exam. https://www.livemint.com/news/world/90-in-bar-exam-88-in-lsat-gpt-4-beats-90-of-humans-in-world-s-toughest-exam-11678882508873.html.

[35] Manal Loukili, Fayçal Messaoudi, and Mohammed El Ghazi. 2023. Sentiment Analysis of Product Reviews for E- Commerce Recommendation Based on Machine Learning. *International Journal of Advances in Soft Computing and its Applications* 15 (March 2023), 1–13. https://doi.org/10.15849/IJASCA.230320.01

[36] Steffen Mau. 2019. *The Metric Society: On the Quantification of the Social.* Polity Press, Cambridge, UK : Medford, MA.

[37] Aaron Mendon-Plasek. 2021. Mechanized Significance and Machine Learning: Why It Became Thinkable and Preferable to Teach Machines to Judge the World. In *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, Jonathan Roberge and Michael Castelle (Eds.). Springer International Publishing, Cham, 31–78. https://doi.org/10.1007/978-3-030-56286-1_2

[38] Dylan Mulvin. 2021. *Proxies: The Cultural Work of Standing In.* The MIT Press, Cambridge, Massachusetts.

[39] Arvind Narayanan and Vitaly Shmatikov. 2007. How To Break Anonymity of the Netflix Prize Dataset. https://doi.org/10.48550/arXiv.cs/0610105 arXiv:cs/0610105

[40] Ryan Northrup. 2023. Voldemort DJs The Great Hogwarts Rave In Harry Potter Art. https://screenrant.com/harry-potter-voldemort-dj-hogwarts-rave-art/.

[41] Will Orr and Kate Crawford. 2023. The Social Construction of Datasets: On the Practices, Processes and Challenges of Dataset Creation for Machine Learning. *SocArXiv* (2023).

[42] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns* 2, 11 (Nov. 2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[43] Billy Perrigo. 2023. Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer. https://time.com/6247678/openai-chatgpt-kenya-workers/.

[44] J. R. Pierce. 1969. Whither Speech Recognition? *The Journal of the Acoustical Society of America* 46, 4B (Oct. 1969), 1049–1051. https://doi.org/10.1121/1.1911801

[45] Theodore M. Porter. 1996. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (2. print., and 1. paperback printing ed.). Princeton Univ.Press, Princeton, N.J.

[46] Theodore M. Porter. 2012. Thin Description: Surface and Depth in Science and Science Studies. *Osiris* 27, 1 (2012), 209–226. https://doi.org/10.1086/667828 jstor:10.1086/667828

[47] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. https://doi.org/10.48550/arXiv.1910.10683 arXiv:1910.10683 [cs, stat]

[48] Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).* Association for Computing Machinery, New York, NY.

[49] Ben Recht. 2023. Revisiting Highleyman's Data. https://www.argmin.net/p/revisiting-highleymans-data.

[50] Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay

Gadepally. 2023. From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. https://doi.org/10.48550/arXiv.2310.03003 arXiv:2310.03003 [cs]

[51] Steven Shapin and Simon Schaffer. 1989. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (1. paperback pr. with corr ed.). Princeton Univ. Pr, Princeton, N.J.

[52] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (April 2020), 13693–13696. https://doi.org/10.1609/aaai.v34i09.7123

[53] Gemini Team. 2023. Gemini: A Family of Highly Capable Multimodal Models. https://doi.org/10.48550/arXiv.2312.11805 arXiv:2312.11805 [cs]

[54] Amirhosein Toosi, Andrea G. Bottino, Babak Saboury, Eliot Siegel, and Arman Rahmim. 2021. A Brief History of AI: How to Prevent Another Winter (A Critical Review). *PET clinics* 16, 4 (Oct. 2021), 449–469. https://doi.org/10.1016/j.cpet.2021.07.001

[55] Jon Victor. 2023. How Google Got Back on Its Feet in AI Race. https://www.theinformation.com/articles/how-google-got-back-on-its-feet-in-ai-race.

[56] James Vincent. 2022. *Beyond Measure: The Hidden History of Measurement.* Faber, London.

[57] 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.* Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

[58] Adrienne Williams, Milagros Miceli, and Timnit Gebru. 2022. The Exploited Labor Behind Artificial Intelligence. *Noema* (Oct. 2022).

[59] Matteo Wong. 2023. AI Doomerism Is a Decoy.

[60] Xue Ying. 2019. An Overview of Overfitting and Its Solutions. *Journal of Physics: Conference Series* 1168, 2 (Feb. 2019), 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

ACKNOWLEDGMENTS

We would like to thank members of the Knowing Machine team for their support and helpful feedback on drafts of this work, particularly Kate Crawford, Mike Ananny, Jason Schultz, Jer Thorp, Melodi Dincer, Hamsini Sridharan, Christo Buschek, Vladan Joler, Sasha Luccioni, Jake Karr, and Hannah Franklin. We would also like to thank the organizers and participants of The Politics of Data beyond Representation panel at the Society for Social Studies of Science (4S) 2023 conference in Honolulu, Hawaii, for their insightful comments and suggestions.