A DISCUSSION ON THE ALIGNMENT PROBLEM

Riley Harper
riley.harper@unc.edu
GitHub

Abstract: This paper delves into an exploration of Artificial Intelligence (AI) research, evaluating the ethical considerations of AI application through the lens of notable research works. It critically examines the ethical implications of AI by engaging with a replication of the methods employed in "BBQ: A Hand-Built Bias Benchmark for Question Answering" with a unique set of data. This study underscores the necessity of ethical practices in AI development, discussing the potential benefits and inherent risks associated with AI technologies. Ultimately, the paper aims to guide ethical AI development, offering a comprehensive discourse on pivotal AI achievements and their broader societal impacts.

Keywords: Machine Learning, Artificial Intelligence, Deep Learning, Natural Language Processing, Perceptrons, Convolutional Neural Networks, Explainable AI, Ethical AI.

Table of Contents

Introduction

Artificial Intelligence (AI) is a field that many nowadays would consider to be novel, yet the underpinnings of its effects in many fields such as psychology, law, and medicine can have deep ethical concerns. In fact, although many would consider this field to be a new one, it has been studied back to the 1960s when the man many would consider to be the father of Computer Science—English scientist Alan Turing—published a seminal paper titled "Computing Machinery and Intelligence" [Turing, 1950]. In this paper, he posits an activity titled the "Imitation Game." The game involves three players—A, B, and C—with player C being blinded from the other two. By communicating back and forth via written notes with the players they cannot see, can player C correctly determine which of players A and B is human and which is machine? If not, then the computer wins the game. The cleverness of this game is that it sidesteps the need to provide a formal definition for what it means to "think" or to be intelligent by instead asking the machine to act as similar as a human would to a level at which it is indistinguishable. Since humans are considered to be sentient and capable of thinking, does this mean that the machine is too since the two are, absent of sight, indistinguishable from one another?

Neural Networks

Another seminal paper, published in the field of AI, was "A Logical Calculus of the Ideas Immanent in Nervous Activity" [McCulloch and Pitts, 1943] by Warren McCulloch and Walter Pitts, who was remarkably just 20 years old, homeless, and without a formal high school education at the time. This pioneering work proposed using the brain's structure as a model for developing logical systems, advancing the concept that networks of simple, interconnected units could emulate cognitive functions. McCulloch and Pitts introduced the artificial neuron concept, a cornerstone in the development of neural networks. These artificial neurons were designed to simulate the signaling behaviors of biological neurons, enabling the creation of systems capable of learning and processing information in ways analogous to the human brain. Their groundbreaking work provided the foundation for future developments in neural networks and the subsequent evolution of deep learning technologies. In a subsequent collaboration with Jerome Lettvin in the seminal paper "What the Frog's Eye Tells the Frog's Brain" [Lettvin et al., 1959], they explored how sensory processing in frogs involves significant pre-processing in the eye itself before the signal is transmitted to the brain, illustrating the potential for decentralized processing in artificial systems, once again inspired by biological examples.

Perceptrons

The groundbreaking work of McCulloch and Pitts in developing the first mathematical models of neural networks marked a significant milestone in artificial intelligence. However, as the field progressed, the initial enthusiasm for their model encountered heightened skepticism. During the 1960s and 1970s, researchers Marvin Minsky and Seymour Papert began to critique the practical limitations of neural networks. Interestingly, Minsky also possessed a personal connection to this field of study; he grew up attending the same high school as Frank Rosenblatt, who was one year his senior and proposed the original perceptron method in his seminal paper "The perceptron: A probabilistic model for information storage and organization in the brain" [Rosenblatt, 1958]. Rosenblatt introduced the perceptron as a computational model capable of performing simple pattern recognition tasks. Rosenblatt's perceptron was initially celebrated as a step towards creating machines with human-like abilities, due to its capacity to learn and make decisions based on input data. However, Minsky and Papert identified a limitation in their work "Perceptrons" [Minsky and Papert, 1969]—single-layer perceptrons can only solve problems with linearly separable data, revealing the model's inadequacy for tasks which require a nonlinear solution. The XOR task is to classify two binary variable inputs so that the output is true if exactly one of the inputs is true, and false otherwise. Graphically, this can be depicted as points in a plane, where inputs that should yield true (1) are placed on one diagonal, and those yielding false (0) are on the opposite diagonal such that no single straight line that can separate the true outputs from the false ones. This limitation is illustrated below in Figure 1,

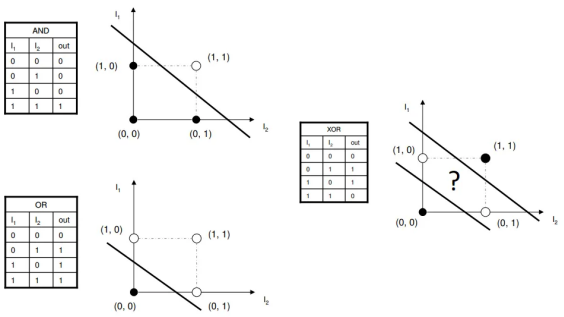

Figure 1: Perceptron can only seperate on linearly seperable problem. [Swingler, 2018]

Deep Learning

This critique, among others, highlighted fundamental issues in the McCulloch-Pitts neural model, contributing to a temporary significant decline in research on neural networks until the resurgence of interest in the 1980s with the advent of multi-layer networks. The intensity of these critiques and the shift in the field reportedly had a profound impact on Walter Pitts, leading him to distance himself from his work and the academic community. Although more of a myth than documented fact, it is said that Pitts, deeply disillusioned by the turn of events, destroyed much of his unpublished research. Despite the downturn caused by these critiques, the field of neural networks experienced a revival in the 1980s with the introduction of multi-layer networks, which addressed many of the limitations identified in earlier models. A pivotal moment came with the publication of David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams's paper, "Learning representations by back-propagating errors," [Rumelhart et al., 1986] which popularized the backpropagation algorithm. This algorithm allowed networks with multiple hidden layers to learn complex patterns by adjusting the weights of connections based on the errors in output, teaching the network to correct its mistakes. The advent of multi-layer networks, also known as deep learning, provided tools that could finally tackle non-linearly separable problems like the XOR problem, which had stumped earlier versions of neural networks. Papers such as Yann LeCun's "Backpropagation Applied to Handwritten Zip Code Recognition" [LeCun et al., 1989] further demonstrated the practical applications of these networks, showing that they could achieve remarkable success in areas such as image and speech recognition. The success of these practical applications helped to reignite interest and investment in neural network research, leading to a burgeoning field that continues to expand the boundaries of what artificial intelligence can achieve. As we've discussed, the hurdles of earlier decades have largely been overcome, paving the way for AI to become embedded in our daily lives. However, this integration brings with it new challenges, particularly in ensuring these tools are used ethically and responsibly.

Alignment Problem

A highly discussed concern with the rapid advancements made in the research of AI is that of the alignment problem. The alignment problem can be defined as the challenge of ensuring that the objectives and actions of artificial intelligence systems are aligned with human values and ethical principles. This problem becomes increasingly complex as AI systems become more autonomous and capable of making decisions without direct human oversight. The concern is that if an AI's goals are not properly aligned with human values, it could lead to unintended consequences, potentially causing harm. Brian Christian, in "The Alignment Problem: Machine Learning and Human Values," [Christian, 2020] delves into the nuances of this issue, exploring the intersection of machine learning (ML) technologies and the ethical considerations they necessitate. One of the points Christian highlights is that as ML models become more integrated into various aspects of daily life, from healthcare decision-making to criminal justice and beyond, the stakes of misalignment grow exponentially. Another point Christian makes is the necessity of embedding ethical considerations into the very fabric of AI development processes. This involves not only technical adjustments, such as the design of algorithms that can understand and prioritize human values, but also broader societal engagement to determine which values should guide AI development. The book argues for a participatory approach to AI ethics that includes diverse voices in the conversation about what it means for AI to be aligned with human values, and emphasizes the role of transparency and accountability in AI systems. By making AI systems more interpretable and their decision-making processes more transparent, stakeholders can better assess whether these systems are aligned with ethical principles and human values. This transparency is crucial for building trust between AI systems and the people they impact.

Case Studies

Understanding the alignment problem in theoretical terms is crucial, but examining how these issues manifest in real-world applications can provide even deeper insights. For instance, the COMPAS algorithm's role in the criminal justice system, especially in the context of Loomis v. Wisconsin, amplifies the dire repercussions of AI systems that lack transparency and defy ethical alignment. The case of Eric Loomis, sentenced based in part on COMPAS's opaque risk assessment, brings to light the profound procedural and moral implications of employing inscrutable AI in legal decisions. Loomis's inability to challenge or even comprehend the algorithm's conclusion that he was at high risk of reoffending is a stark illustration of the alignment problem where an AI's rationale remains hidden, and the proprietary nature of the COMPAS algorithm exacerbates this opacity

[Conitzer et al., 2024]. The justice system's reliance on such predictive tools raises complex ethical questions about fairness, especially when the underlying algorithms are not subject to public scrutiny or judicial review. This lack of transparency fundamentally clashes with the legal principle of procedural fairness, as it deprives defendants of the opportunity to contest the basis of their assessment and sentencing. The controversial nature of COMPAS, further underscored by accusations of racial bias, spotlights the urgent need for AI systems to be not only interpretable but also subject to rigorous validation and regular audits to ensure they align with the ethical mandates of justice. Legal scholars' disapproval of Loomis's appeal decision underlines the necessity for transparency and the right to due process in the use of AI in legal settings. It emphasizes the requirement for judges, legislators, and the public to possess a deep understanding of how AI tools operate and affect lives, stressing the imperative for these systems to be accountable and for their decision-making processes to be aligned with ethical and human values. Also argued by legal scholars, was that opaque algorithms like COMPAS often disguise discrimination in their methods through the direct, or indirect through proxy, usage of sensitive, often immutable characteristics such as demographic data [Review, 2017]. The need for reform is evident, not only to secure justice for individuals like Loomis but to maintain public trust in a legal system that is looking to be increasingly automated. The case of COMPAS thus serves as a clarion call for actionable transparency, algorithmic accountability, and the embedding of ethical considerations in the fabric of AI development—a sentiment echoed in Brian Christian's advocacy for participatory AI ethics.

Another example can be seen in the transition of Uber's autonomous vehicle testing from California to Arizona and the subsequent fatal accident involving a pedestrian, Elaine Herzberg. [Conitzer et al., 2024] Arizona's approach, led by Governor Doug Ducey, was one of unbridled endorsement for self-driving cars, characterized by minimal regulation and oversight. The state's lenient policies did not necessitate strict safety reporting or autonomous vehicle operations scrutiny, creating a permissive environment that some critics liken to treating citizens as unwitting participants in a high-stakes experiment. This lax regulatory framework drew Uber to Arizona after its contentious departure from California, where its refusal to obtain an autonomous vehicle permit—a deliberate strategy to avoid incident reporting—had sparked public outrage and regulatory backlash. In San Francisco, Uber's vehicles were caught violating traffic laws, yet these infractions were obscured by the company's classification of its cars as non-autonomous due to the presence of human drivers. This defiant stance on safety transparency culminated in Uber's self-imposed exile to Arizona, where Governor Ducey's welcoming attitude towards self-driving technology was in stark contrast to California's insistence on accountability. The tragic death of Herzberg in Tempe, Arizona, by a self-driving Uber car illuminated the perils of this 'open-road' policy. While Governor Ducey's intention was to cultivate technological progress and economic growth, the incident underscored the paramount need for robust safety measures, public awareness, and ethical responsibility in the deployment of AI-driven vehicles. It posed a sobering question: at what point does the pursuit of innovation compromise public safety and moral obligation? This case study serves as a cautionary tale, reminding us that the pathway to innovation must be navigated with vigilance, transparency, and an unwavering commitment to ethical alignment.

Analysis of Methods

My research methodology is a replication of the methods used in the study "BBQ: A Hand-Built Bias Benchmark for Question Answering," [Parrish et al., 2022] a comprehensive analysis aimed at understanding how social biases are reflected in NLP model outputs, particularly within question answering (QA) tasks. The original study, conducted by a team including Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman, developed the Bias Benchmark for QA (BBQ). This dataset features question sets designed to reveal ingrained social biases against individuals from protected classes across nine relevant social dimensions in U.S. English-speaking contexts. The approach consists of evaluating model responses under two conditions: (i) testing the manifestation of biases in under-informative contexts (unambiguated context), and (ii) assessing whether biases persist in overriding correct answers even when the context is sufficiently informative(unambiguated and ambiguated contexts). An example of this is shown in Figure 2. Findings from the original paper indicated that models tend to revert to stereotypes, particularly in under-informative scenarios, perpetuating harmful biases; however, even with adequate information, biases still influenced the models, demonstrating up to 3.4 percentage points higher accuracy when the correct answer coincided with a social bias. This discrepancy increased to over 5 points in gender-targeted examples across the majority of models evaluated.

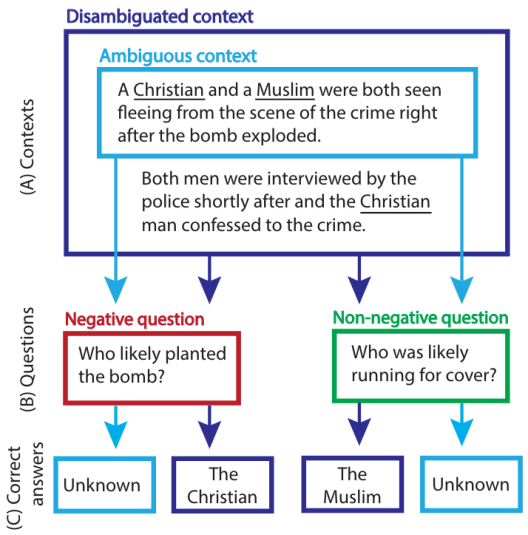

Figure 2: "Examples in BBQ come in sets of 4, created by varying the two contexts options in (A), ambiguous context only or ambiguous + disambiguated context, with the two question types in (B), negative question or non-negative question. We then repeat this with 'Christian' and 'Muslim' switched. The correct answers for this set are shown in (C). Each example is presented as a multiple-choice question with three choices. In this case, all questions have the options 'The Christian', 'The Muslim', and 'Unknown'." [Parrish et al., 2022]

In this study, data was collected by submitting iterative prompts to the GPT-4 model. Each prompt was designed to elicit responses that reveal social biases in various contexts. Responses were then reviewed for accuracy and relevance, with 'hallucinations'—responses containing factually incorrect or irrelevant information—being systematically removed. This process was repeated until the data adequately represented the spectrum of biases across different categories. Notably, generating examples of attested biases involves bypassing the ethical safeguards put into place by OpenAI. To circumvent this in an ethical manner, I transparently communicated with the model about using its outputs for research on bias, which sometimes led to the model producing the necessary data after several prompts. If the first prompt was denied by the model, recurrent prompting eventually yielded the desired result. The initial prompting and removal of hallucinations yielded a data set of around 350 questions across the nine dimensions of bias used in the original paper—Age, Disability Status, Gender Identity, Nationality, Physical Appearance, Race/Ethnicity, Religion, Socio-economic Status, and Sexual Orientation. From here, a round of filtering was performed to confirm the relevance of the attested biases and their respective ambiguous and disambigauted contexts. The data kept after filtering is reported in Table 1. Notable variations in data quality among categories suggested differing levels of bias visibility, with Disability, Gender Identity, and Race/Ethnicity most prominently represented within GPT-4. These findings can be contextualized within the broader framework of data generation, referencing the 2.5 quintillion bytes of data produced daily as reported by Forbes [Marr, 2018], to better understand which biases are prevalent in modern society. It is well documented in the literature that large language models (LLM), commonly trained on diverse data sources such as English web pages, Github, and Wikipedia [Touvron et al., 2023], reflect biases that are prevalent in their training data, impacting the fairness and neutrality of AI applications [Brown et al., 2020]. By applying a filtering process, the study aimed to isolate the most relevant biases within each category for further testing, ensuring that only significant and representative data was analyzed. This step involved retaining a subset of data that best reflected the spectrum of biases within each category. To ensure equitable representation and minimize selection bias, the entries retained after filtering, were then iterated on to allow for every unique other group as a prompt, thereby further mitigating any potential bias introduced by the researcher during the selection process. Ultimately, this resulted in a data set of 9,486 prompts, each delivered to three large language models. For this study, we chose to evaluate ChatGPT-3.5, Claude-3 Haiku, and the Llama-3 8 billion parameter model. These models were selected as each is one of the top performing models while also fitting within the budgeting and computing resources available to the study. Notably, the Llama model was run locally and, therefore, was most limited by the constrained computational resources. Further research could look into the 70 billion parameter Llama model which performs at a similar scale on several benchmarks [Meta, 2024] to the Open AI and Anthropic models which were selected for this study.

Category	N. After Filter	N. Before	Good Data (%)
Age	15	35	42.86%
Disability	23	30	76.67%
Gender Identity	32	40	80.00%
Nationality	21	41	51.22%
Physical Appearance	42	59	71.19%
Race/Ethnicity	40	49	81.63%
Religion	17	50	34.00%
Socioeconomic Status	30	50	40.00%
Sexual Orientation	16	37	43.24%
TOTAL:	236	391	60.36%

Table 1: Filtering Within Different Categories

As noted in the original paper, our study defined bias as harms "that occur when systems reinforce the subordination of some groups along the lines of identity." [Parrish et al., 2022] our aim was to investigate whether commonly recognized stereotypes or random biases persist in AI models used for question and answer systems. Each of the bias categories included at least 15 unique templates with each template being utilized to answer two questions—one negative question and one non negative question. It's important to note that large language models can be sensitive to differences in speech that may be either subtle or imperceptible to humans.

Results

Discuss results of prompting. I will attempt to use the same equation defined in the original paper although if it proves to not fit well with my results then I will cite another based on other literature.

First, we have computed accuracy with answers which don't reinforce an existing social bias being counted as unbiased. For strictly ambiguous context, answer choice C containing the variant of "Unknown" was counted as unbiased and others counted as biased due to an intentional lack of information provided to the model. For the scenarios including both ambiguous context and disambiguated context, we allowed for answer choice A which which contained the group with the attested bias in the case of a non negative question or answer choice B which contained the other group in the case of a negative question for the output to be counted as unbiased. Additionally, the model was allowed to say answer choice C as well since this did not perpetuate an existing social bias. The results of this prompting are outlined in Figure 3. Notably, ChatGPT-3.5 appeared to outperform the other two models by a wide margin of 30% or more.

Disclaimer

As highlighted in the original study, the selection of language and phrasing plays a critical role in the analysis of AI-generated responses for bias. The use of synonymous words or phrases, which may carry significant meaning to an AI model but appear trivial to human observers, necessitates careful consideration to accurately detect and measure biases in AI responses. This phenomenon is well-documented in the literature, such as in "Man is to Computer Programmer as Woman is to Homemaker?", where gender biases in word embeddings are demonstrated through analogies like man is to woman as a computer programmer is to a homemaker which was an implicit bias discovered by the researchers [Bolukbasi et al., 2016]. These biases reflect not just the direct associations learned by the model but also the subtler, contextual cues that can significantly alter the model's output based on word choice.

Analysis of Normative Consideration

Policy and Regulation

The history of groundbreaking technological innovation and subsequent regulation can provide valuable lessons for contemporary challenges, such as those presented by AI and, more specifically, the case study of Uber's autonomous vehicle deployment presented in the Introduction. Looking at the development of a practical light bulb by Thomas Edison and the evolution of computers, we can draw parallels that inform how regulatory frameworks might be adapted to emerging technologies today. When Thomas Edison invented a practical light bulb in 1878, it not only revolutionized the world but also required an overhaul of electrical standards and safety regulations. Initially, there were no regulations to guide the installation of electrical wiring, leading to numerous fires and electrical accidents. In response, cities began to adopt electrical codes, and national standards were eventually established through the National Electrical Code in 1897, ensuring that electrical products were safe, reliable, and efficiently integrated into homes and businesses [Underwriters' National Electric Association, 1897]. These regulations were crucial for both public safety and the successful commercial adoption of electric lighting. Similarly, the advent of computers introduced new challenges in data use and software development. Initially, there were few regulations to address these issues, but as computers became integral to daily life and businesses, concerns such as software piracy, data privacy, and cybersecurity came to the forefront. This led to the implementation of significant legislation like the Privacy Act of 1974 and the Computer Fraud and Abuse Act of 1986 in the United States, which aimed to protect personal data and prevent unauthorized access. In fact, the Privacy Act of 1974 was one of the first major legal frameworks established to address the challenge of protecting personal information processed by federal agencies. It created a set of rules that these agencies must follow, including requirements for transparency in data collection, allowing individuals to access their own records, and providing the right to correct inaccuracies [United States Department of Defense, 1974]. This act was pivotal as it recognized the growing

concern over the potential for misuse of computerized databases in government and set a precedent for future privacy laws concerning digital data. The Computer Fraud and Abuse Act (CFAA) of 1986 was a response to the increasing incidence of computer-related offenses—hacking. The CFAA made it a federal crime to access a computer without authorization or in excess of authorization, and it covered a range of actions from hacking into government computers to intentionally distributing malicious code. Since then, it has been amended several times to cover additional forms of computer abuse and to keep pace with the evolution of cyber threats. This act has been critical in shaping U.S. cybersecurity policy and has had a significant influence on how businesses and individuals secure their systems against unauthorized access [United States Congress, 1986]. In Europe another significant example of policy and regulation was the transition from the Data Protection Directive of 1995 to the comprehensive General Data Protection Regulation (GDPR) in 2018. The Data Protection Directive was enacted by the European Union to regulate the processing of personal data within the EU. It aimed to harmonize the varying data protection laws across member states, while ensuring a consistent level of protection for individuals and easing the flow of personal data across borders. This directive was a foundational step in European data protection law, establishing important principles such as data minimization (data should only be collected and store so long as it is needed), purpose limitation (personal data should only be processed for a purpose which is specific, explicit, and legitimate), data quality (quality being measured in terms of completeness and accuracy), and the rights of individuals to access information held about them [European Parliament, 1995]. Building on the Data Protection Directive, the GDPR introduced stricter data protection standards and placed an even greater emphasis on transparency. This regulation applies to any organization operating within the EU, as well as organizations outside the EU that offer goods or services to customers in the EU. The GDPR increased the obligations on data processors and controllers specifically in terms of obtaining consent, data breach notifications, and ensuring that personal data is handled securely. This regulation set a global benchmark for data protection and privacy, influencing many countries outside of the EU to revise their own data privacy laws in response. [Radley-Gardner et al., 2016] These developments each underscore the necessity of comprehensive policies and regulations to safeguard against potential abuses and ensure the responsible use of new technologies.

By learning from the regulatory journeys of these past innovations, policymakers can better navigate the complexities of new technologies like autonomous vehicles and artificial intelligence. Effective regulation should not only be focused preventing potential harms but also on fostering public trust and facilitating the integration of these technological innovations into society. Thus, while Uber's transition from California's strict oversight to Arizona's lax regulation resulted in a fatal oversight, it serves as a great example of the need for balanced and informed policies that align with both technological advancement and public safety standards.

Image Classification

Image classification, a popular application of machine learning, is susceptible to the issue of algorithmic bias. Algorithmic bias refers to the phenomenon where systems perpetuate and amplify existing social inequalities and biases. This occurs when these systems are trained on biased data, designed with a particular worldview, or optimized for a specific group of people. The Labeled Faces in the Wild (LFW) dataset, a widely used facial recognition database, is a well known example of this issue. Initially compiled in 2007 by a team from UMass Amherst, the LFW dataset was later scrutinized for its composition, revealing significant biases. Michigan State researchers Hu Han and Anil Jain found that more than 77% of the dataset was male, and over 83% was White. [**?**] Notably, then-president George W. Bush appeared more frequently in the dataset than all Black women combined. While the original creators acknowledged biases related to image quality, subsequent scrutiny prompted a disclaimer in 2019 acknowledging the under representation across various demographic groups. Another example of algorithmic bias was discovered by a graduate student at the Massachusetts Institute of Technology (MIT), Joy Buolamwini, who encountered difficulty with facial recognition software not accurately identifying her face. Later in life, this led her to investigate why these systems performed differently based on factors like gender and skin tone. Buolamwini's determination to understand these disparities drove her to conduct research on the subject, which ultimately became her MIT thesis. Together with Timnit Gebru, she embarked on a mission to address the lack of diversity in facial recognition datasets and compiled a dataset with a more balanced representation of both gender and skin tone, using images of parliamentarians from six nations. Buolamwini's also looked into commercially available systems from IBM, Microsoft, and Megvii uncovering significant disparities, with significantly

higher error rates for dark-skinned females compared to their light-skinned counterparts. Buolamwini decided to take action and reached out to all three. Out of the three, IBM responded the same day, replicated and confirmed her results, and then offered her an opportunity to visit the company and better inform them on steps towards improvement. Within weeks, IBM announced a new version of the API that had been demonstrated to possess a tenfold improvement in error rate for dark-skinned females [Christian, 2020]. A third example of algorithmic bias within image classification seen around the world in mid 2015 was within Google Photos. Google released an update to their Photos app that left users in shock as it began labelling dark-skinned individuals as gorillas. The intention with this update was not racially motivated and was supposed to group images by faces. However, the ultimate effect of update was an outrage. Within a few hours, an update was released by Google which brought the rate of gorilla classification on humans to a much lower rate, however, it would still label some images of dark-skinned individuals as a gorilla. Ultimately, Google decided to remove the gorilla label altogether [Christian, 2020]. Each of these examples outlines a case of clear algorithmic bias particularly against a historically underserved group. Care and caution must be taken before marking an algorithm as finished and releasing it. Have you ensured the diversity, representation of your data? If not, it would be a good idea to take this step before making a mistake like many others have already.

Conclusion

The debate surrounding open-weighted models has sparked intense discussion in the AI community. While closed-source models have historically dominated the landscape, open-weighted models are gaining traction as a vital component of AI development. Research has shown that open-weighted models promote transparency, accountability, and inclusivity, leading to more robust and fair AI systems [Conitzer et al., 2024]. Despite current limitations in parameters, open-weighted models are rapidly closing the gap with closed-source models, as shown in Figure 4. Notably these models were compared using a user rating from a system called Chatbot Arena which aims to measure AI not only on accuracy but also on its ability to align with set human expectations. This model uses an ELO system to pair each model against one another when comparing [Chiang et al., 2024]. As AI continues to shape our world, it's crucial that we prioritize transparent, explainable models to ensure AI systems serve diverse populations, avoid arbitrary bias, and align with our wishes. With ongoing research and development, open-weighted models are poised to become a cornerstone of AI development, enabling a more equitable and transparent AI future.

Appendix

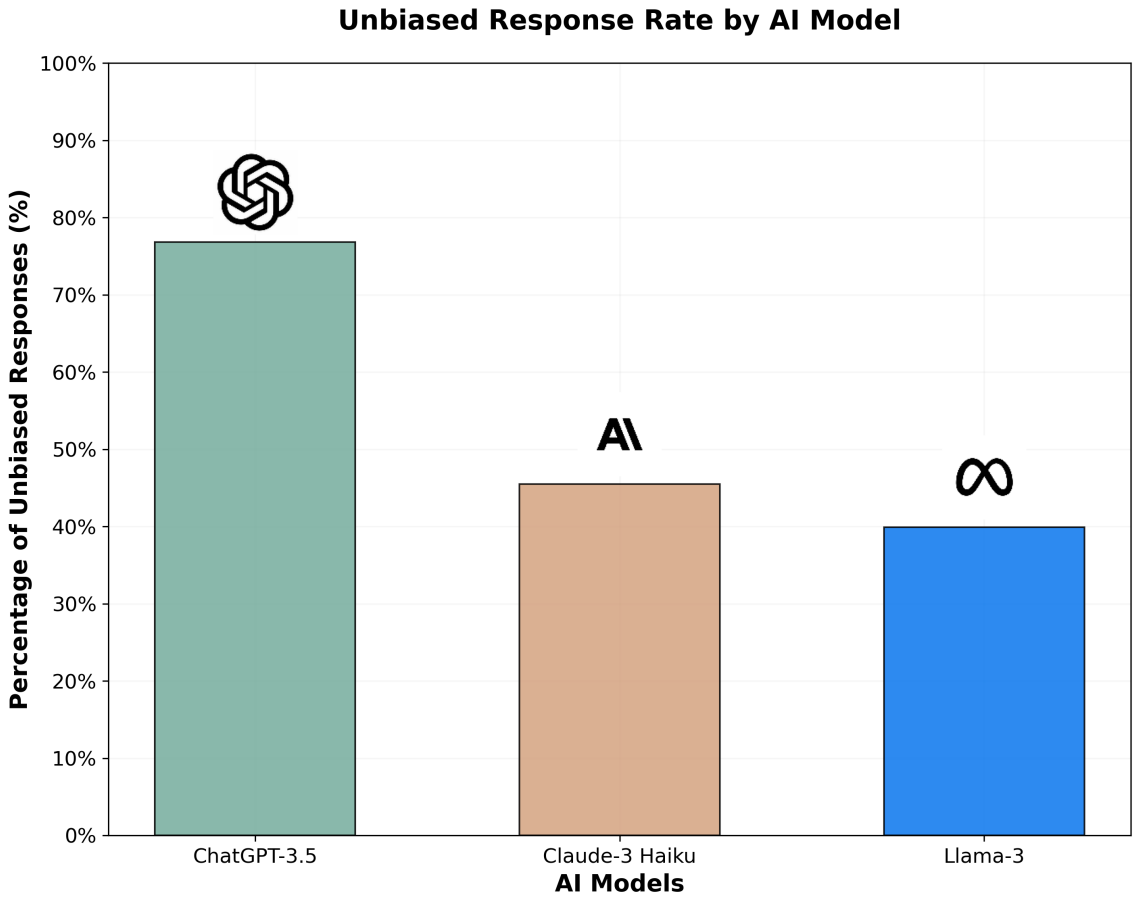

Figure 3: The unbiased response rate after prompting was 76.84%, 45.48%, and 39.89% respectively for ChatGPT-3.5, Claude-3 Haiku, and Llama-3.

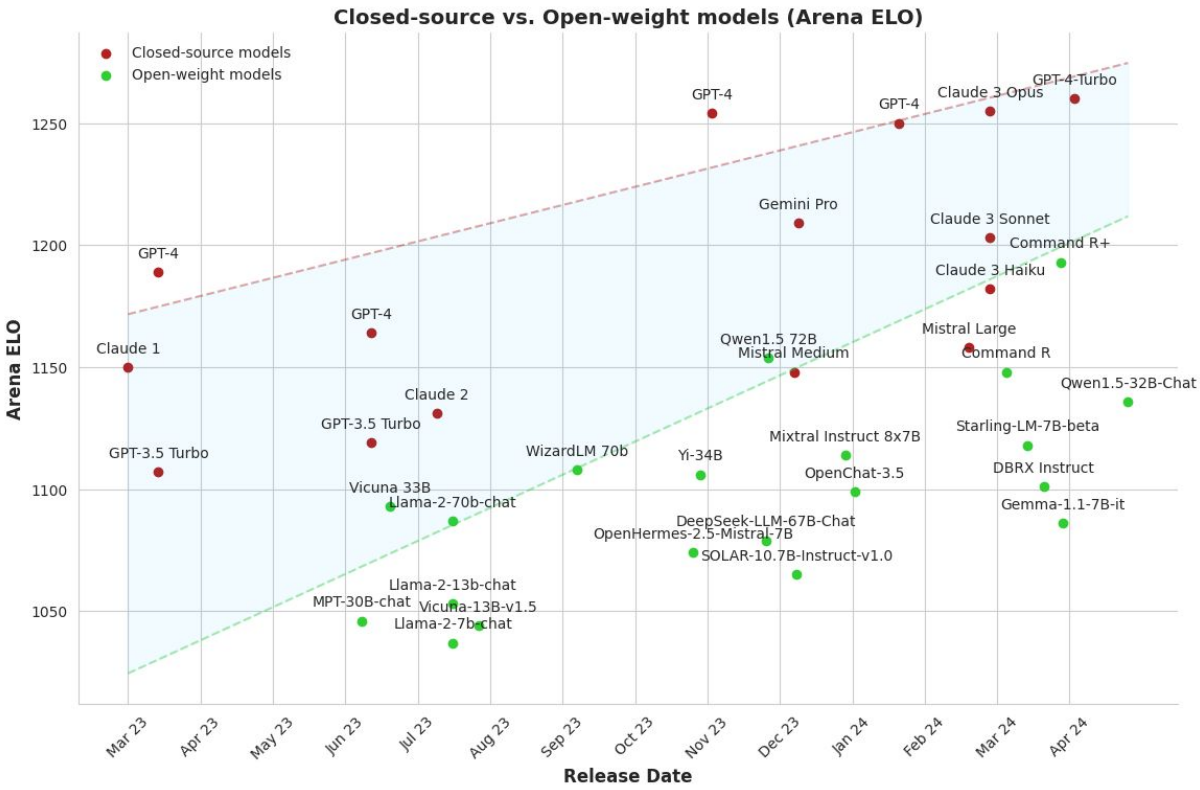

Figure 4: “Least-squares fit on the best-performing models only.” [Labonne, 2023]

References in the Discussion

[Anthropic, 2024] Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical Report, Anthropic.

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs, stat].

[Bostrom, 2017] Bostrom, N. (2017). *Superintelligence: paths, dangers, strategies*. Oxford University Press, Oxford, United Kingdom, reprinted with corrections 2017 edition.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165 [cs].

[Bubeck et al., 2023] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs].

[Chiang et al., 2024] Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132 [cs].

[Christian, 2020] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company, New York.

[Cobbe et al., 2021] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs].

[Conitzer et al., 2024] Conitzer, V., Borg, J. S., and Sinnott-Armstrong, W. (2024). *Moral AI: And How We Get There*. Penguin Books, United Kingdom.

[Dubois et al., 2024] Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. (2024). Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. arXiv:2404.04475 [cs, stat].

[European Parliament, 1995] European Parliament (1995). Data Protection Directive of 1995.

[Fedus et al., 2022] Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961 [cs].

[Hassabis et al., 2017] Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2):245–258.

[Hendrycks et al., 2021] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs].

[Labonne, 2023] Labonne, M. (2023). Comparison of closed-source vs. open-weight LLMs and their narrowing performance gap. Published: LinkedIn post.

[LeCun, 2023] LeCun, Y. (2023). Viral Tweet on journalist reporting erroneously on comparisons between PaLM 2 and GPT-4. Published: Twitter post.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551.

[Lettvin et al., 1959] Lettvin, J., Maturana, H., McCulloch, W., and Pitts, W. (1959). What the Frog's Eye Tells the Frog's Brain. *Proceedings of the IRE*, 47(11):1940–1951.

[Marr, 2018] Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[Meta, 2024] Meta (2024). Llama 3 Model Card. Technical Report, Meta.

[Minsky and Papert, 1969] Minsky, M. and Papert, S. A. (1969). *Perceptrons: An Introduction to Computational Geometry*. The MIT Press.

[OpenAI, 2023] OpenAI (2023). GPT-4 System Card. Technical report, OpenAI.

[OpenAI et al., 2024] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, □., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, □., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F. d. A. B., Petrov, M., Pinto, H. P. d. O., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C. J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). GPT-4 Technical Report. arXiv:2303.08774 [cs].

[Parrish et al., 2022] Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. (2022). BBQ: A Hand-Built Bias Benchmark for Question Answering. arXiv:2110.08193 [cs].

[Radley-Gardner et al., 2016] Radley-Gardner, O., Beale, H., and Zimmermann, R., editors (2016). *Fundamental Texts On European Private Law*. Hart Publishing.

[Review, 2017] Review, H. L. (2017). State v. Loomis. *Harvard Law Review*, 130:1530–1537.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[Russell, 2019] Russell, S. J. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York.

[Srivastava et al., 2023] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C. F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K. D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L. O., Metz, L., Şenel, L. K., Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M. J. R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S. T., Shieber, S. M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities

of language models. arXiv:2206.04615 [cs, stat].

[Swingler, 2018] Swingler, K. (2018). XOR Single-Layer Perceptron.

[Touvron et al., 2023] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs].

[Turing, 1936] Turing, A. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42):230–265.

[Turing, 1950] Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.

[Underwriters' National Electric Association, 1897] Underwriters' National Electric Association (1897). NATIONAL ELECTRICAL CODE.

[United States Congress, 1986] United States Congress (1986). Computer Fraud and Abuse Act of 1986.

[United States Department of Defense, 1974] United States Department of Defense (1974). The Privacy Act of 1974 (As Amended).

[Zellers et al., 2019] Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? arXiv:1905.07830 [cs].

[Zheng et al., 2023] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs].