Responsible AI in the OSS: Reconciling Innovation with Risk Assessment and Disclosure

Mahasweta Chakraborti[1], Bert Joseph Prestoza[1], Nicholas Vincent[2], Vladimir Filkov[1], Seth Frey[1]
[1]University of California Davis, USA
[2]Simon Fraser University, CA
{mchakraborti, sethfrey, vfilkov}@ucdavis.edu, nvincent@sfu.ca, bertjosephprestoza@gmail.com

Abstract

Ethical concerns around AI have increased emphasis on model auditing and reporting requirements. We thoroughly review the current state of governance and evaluation practices to identify specific challenges to responsible AI development in OSS. We then analyze OSS projects to understand if model evaluation is associated with safety assessments, through documentation of limitations, biases, and other risks. Our analysis of 7902 Hugging Face projects found that while risk documentation is strongly associated with evaluation practices, high performers from the platform's largest competitive leaderboard (N=789) were less accountable. Recognizing these delicate tensions from performance incentives may guide providers in revisiting the objectives of evaluation and legal scholars in formulating platform interventions and policies that balance innovation and responsibility.

Introduction

Artificial intelligence (AI) has been widely adopted for individual, collective, and public use, including in education, business, and research domains. This widespread implementation has been argued to boost productivity, enhance manufacturing, accelerate development, and facilitate the provisioning of critical services and infrastructure at unprecedented levels and reach (see e.g. Chapter 4 of the AI Index Report for an overview of specific claims along these lines (Maslej et al. 2024)). The scalability offered by these technologies could revolutionize numerous industries and promote further investment in AI innovation for improved service delivery across diverse domains.

Our work explores current evaluation practices and developer accountability in open source AI development. OSS is extremely valuable (Wright, Nagle, and Greenstein 2024; Hoffmann, Nagle, and Zhou 2024), and open source AI applications are seeing an increasing market footprint (Liesenfeld and Dingemanse 2024). Nascent technologies often face scrutiny before they earn public trust (Floridi et al. 2018). AI usability can be threatened by the quality of training data (Hutchinson et al. 2021) and algorithmic limitations (Dwork et al. 2012; Weidinger et al. 2021). AI applications are also highly prone to inappropriate uses (Contractor et al. 2022; McDuff et al. 2024). Therefore, potential risks in AI should be addressed at every stage of development, and such concerns have led to calls for ethical training and regulatory oversight for providers and deployers. Yet OSS faces unique challenges in promoting developer ethics and actuating responsible development.

Model evaluation is a central component of the AI development and deployment cycle. In its most common form, evaluation involves testing models on held-out datasets to understand how well a model has learned. Such benchmarking is essential for assessing novelty against the state-of-the-art, deciding whether the model is suitable for widespread use, and informing future design choices. Today, competitive benchmarking against other models is a popular form of evaluation, and high performers garner legitimacy from users and investors, enjoy market visibility, and even steer development (Dehghani et al. 2021; Ethayarajh and Jurafsky 2020; Raji et al. 2021).

With rising stakes, developers are increasingly expected to use evaluations not only to assess model capabilities but also to *recognize specific limitations* (Rauh et al. 2024; Lam et al. 2024). Given product objectives and target uses, vulnerabilities can arise from development aspects such as biased training data (Hutchinson et al. 2021) (content, source, representative quality, etc.), limited robustness or generalizability (Li et al. 2023), and other associated design choices. Therefore, holistic evaluation should go beyond simply reporting gross accuracy and involve probing for edge cases, measuring biases in predictions for specific domains and vulnerable subpopulations, and stress testing for possible failure modes (Liang et al. 2023; Bommasani et al. 2022). Importantly, these assessments should be documented and reported in an interpretable, accessible form that can empower developers, investors, and ordinary users to make informed choices over model application and mitigate liability and harm.

We conduct an in-depth model documentation study of metadata on Hugging Face (HF), one of the largest open source AI hosting platforms. In particular, we focus on collecting data and conducting quantitative analyses on the evaluation and risk documentation practices among projects and modeling the relationship between the two.

Our contributions include the following:

- First, we share our detailed review of documentation practices, project organization, information manage-

ment, monitoring, and other forms of platform support on Hugging Face.

- We conduct extensive data collection to explore the evolution of the platform in terms of scale of models, their applications, developer contribution patterns, and documentation practices.
- Among 456,545 usable, service-ready projects, around 15.9% and 2.2% models contained evaluation and risk-related documentation from the developer.
- Next, we filter our dataset to examine 7902 models that 1. carried sufficient model information 2. satisfy both OSI (Open Source Initiative 2024) and FSF (Free Software Foundation 2024) conceptions of open source AI. Based on a thorough quantitative analysis with controls, we find a strong, positive association between the practices of model evaluation and risk documentation practices.
- Lastly, we further look only at 789 of these projects participating in HF's Open LLM leaderboard (Beeching et al. 2023). Here, higher-performing models were found to be less likely to provide documentation on risks and limitations, highlighting the need for technical governance tailored to OSS ecosystems.

Motivation

Here, we discuss background on Open Source AI and key areas of research that motivated our dataset documentation effort and accompanying quantitative study.

Ethics of OSS AI

Open sourcing has led to remarkable progress in science and technology, and notable AI projects also started out as OSS to inspire and support collective innovation. At the same time, there have been multiple instances of biased or improperly curated training data and artifacts that can compromise regular applications (Birhane, Prabhu, and Kahembwe 2021; Birhane et al. 2023; Lee et al. 2024), or OSS applications being reportedly appropriated for nefarious uses (DoJ 2024; AIID 2024; Marchal et al. 2024; Mellor 2022).

AI has seen oversight from both academia and governments. Researchers and ethicists came together early on to specify and advocate documentation guidelines for every link in the AI development pipeline. Model cards, (Mitchell et al. 2019; Crisan et al. 2022; Arnold et al. 2019), datasheets (Gebru et al. 2018; Bender et al. 2018), and other factsheets are crucial sociotechnical governance tools that keep stakeholders informed and define the scope of AI consumer applications, ultimately contributing to more transparent and accountable development practices. These outline permissible use cases (possibly extending to licensing (Contractor et al. 2022)), caution against out-of-scope applications, and disclose other anticipated use risks. Examples include predictive biases across ethnicity or gender-profession skew in generated outputs, robustness to adversarial attacks, corner cases (e.g., whether a medical diagnostic system underdiagnoses rare conditions), etc. Recent work has empirically evaluated the effect of responsible AI artifacts such as Model Cards (Berman, Goyal, and Madaio 2024; Kawakami, Wilkinson, and Chouldechova 2024; Deng et al. 2022; Pistilli et al. 2023).

Formal governance systems are rapidly shaping both at the state and global level (Arnold et al. 2019; Floridi et al. 2018; Olteanu et al. 2019; Jobin et al. 2019) and have also heavily incorporated these governance tools. The EU AI Act, one of the first major steps towards formalized AI governance (European Parliament and Council of the European Union 2024), strongly endorses datasheets and model cards for OSS. However, as a result of significant deliberation on the regulation vs innovation tradeoff, open source AI has been kept largely exempt from any strict evaluation and reporting standards so far (Law and Krier 2023; Liesenfeld and Dingemanse 2024). Therefore, to enjoy user reception and sustain projects, open source developers must still strive towards responsible practices through information sharing and institute contingencies to address risks and harms.

Fostering and standardizing best practices is a challenge in the OSS, as unlike for-profit corporations, projects here have been historically informal and decentralized (Benkler 2006; Crowston et al. 2012; Coleman 2013). Developer motivations range from altruism to popularity to professional development, which can lead to varying governance structures, cultural codes, and project objectives (Li et al. 2021; Chakraborti et al. 2024; Yin et al. 2022). This may complicate consensus building on standards and best practices, as well as enforcement and monitoring. Therefore, responsible AI in OSS would require an approach beyond centralized top-down management and more large-scale collaboration to monitor evaluation, documentation standards, and track malpractice (Franzen 2024; Ethayarajh and Jurafsky 2020; Balloccu et al. 2024). Depending on the criticality of the application (e.g., medical diagnostics or defense), evaluation may demand nuanced expertise, where the developer needs to invest in understanding methodologies, conducting different tests, and selecting appropriate metrics. As a result, comprehensive evaluations (Liang et al. 2023) may also bring additional costs and effort to projects. This may be especially untenable for small teams comprising volunteers or fleeting contributors. Lastly, developers may be apprehensive about retaining their user base with stringent risk disclosure requirements.

Given the contemporary state of affairs, it is imperative to closely understand OSS developer aspirations to formulate incentives that promote responsible development while preserving creative freedom.

Evaluation and Responsible AI

Evaluation and testing are integral to the development of emergent technologies. The accuracy of model performance serves to inform design decisions and investment. However, performance is just one of the axes of evaluation, the other being user safety and governance (Mökander et al. 2023; Falco et al. 2021; Rauh et al. 2024). Ideally, even a model exhibiting 100% accuracy should explain tests it undertook to determine (or rule out) risks and guarantee safe deployment. We review current approaches to AI evaluation that inform how we analyze our collected data and interpret our findings.

A bid for greater accuracy through standardization led to the rise of evaluation benchmarks. As benchmarks became more widely adopted, they evolved into popular, competitive leaderboards (Wang et al. 2019, 2018; Rajpurkar et al. 2016; Srivastava et al. 2022; Bojar et al. 2017; Deng et al. 2009; LeCun et al. 1998) that generally rank models by performance. These initiatives are of particular interest, as they draw significant attention and participation. They are a predominant mode of evaluation today and serve as a recognition-driven incentive for continuous, incremental innovation.

Despite their usefulness as a general estimation of model capabilities, accuracy on a specific test or set of tests does not necessarily quantify or guarantee generalizability, adversarial robustness, or risk tolerance. That is, the error rate on a standard test may not always reflect the population error rate (Jia and Liang 2017; Zhang et al. 2020; Ethayarajh and Jurafsky 2020; Dehghani et al. 2021; Arora and Zhang 2021; Blum and Hardt 2015), and developers need to anticipate and probe failure modes despite a high performance. Moreover, these ranks, being generally based on models' aggregate predictive accuracy over a collection of static tasks and datasets, often conceal racial and gender biases (Bordia and Bowman 2019; Manzini et al. 2019; Rudinger et al. 2018; Blodgett et al. 2020). This lack of transparency and granularity is well known (Ethayarajh and Jurafsky 2020; Raji et al. 2021).

Several researchers have reported other issues, including errors in test data, model contamination, and other vulnerabilities, that can compromise their validity in general (Dehghani et al. 2021; Gema et al. 2024; Bowman and Dahl 2021; Sainz et al. 2023). Due to the visibility enjoyed by top performing developers, leaderboards have been known to be gamed through multiple submissions and tweaks that marginally boost performance without actual meaningful, feature contributions (Hardt 2017). Infact, leaderboards are often dominated by highly over-parameterized, complex, overfitted models (Ethayarajh and Jurafsky 2020; Hardt 2017; Arora and Zhang 2021; Dehghani et al. 2021; Blum and Hardt 2015). All in all, *competitiveness and focus on performance alone tend to undermine the role of evaluation in detecting risks and testing for application safety*. With the legitimacy popular leaderboards enjoy (Raji et al. 2021), high performance may easily lead developers and users to assume generalizability of their models and undermine the need for additional efforts into nuanced evaluations for runtime risks. This is especially dangerous and necessitates reform in current developer practices.

With increasing recognition of the inadequacy of current evaluation practices, scholarship in AI safety has seen a notable push towards holistic approaches with expanding definitions, objectives and approaches beyond simply predictive accuracy (Liang et al. 2023; Bommasani et al. 2023, 2022; Mehrabi et al. 2021). These include robustness against malicious or adversarial inputs (Nie et al. 2020), explainability of the model's decision-making and intermediate processes (Liao et al. 2021; Ehsan et al. 2021), generalizability on out-of-sample data (Bartlett, Montanari, and Rakhlin 2021), and granular testing across demographic subgroups for biased behavior. There have also been notable strides in designing evaluations to measure model fairness (Zhao et al. 2018; Nadeem, Bethke, and Reddy 2021; Raji and Buolamwini 2019).

Related Empirical Studies

With rapid interest and growth of AI in open source contributions, researchers have been exploring the potential of data-driven research through repository mining on AI-centric development platforms and services (Jiang et al. 2023a,c; Ait, Izquierdo, and Cabot 2023; Ait, Cánovas Izquierdo, and Cabot 2024). Software engineers have been particularly interested in modularity and artifact reuse from pre-trained model repositories or "PTMs" (Jiang et al. 2023b; Gong et al. 2023; Taraghi et al. 2024), and accompanying security risks and vulnerabilities (Jiang et al. 2022; Kathikar et al. 2023). Several studies have explored model documentation (Pepe et al. 2024; Gong et al. 2023; Yang, Liang, and Zou 2023). Gong et al. focus on usage documentation across multiple platforms (Gong et al. 2023), while Liang et al. analyzed different sections across Hugging Face modelcards and how comprehensive documentation may improve model popularity (Liang et al. 2024). Castano et al. studied carbon footprint reporting (Castaño, Martínez-Fernández, and Franch 2024). Hugging Face's internal study found that among all model card components, respondents found the risks sections the longest and most challenging to complete (Ozoani, Gerchick, and Mitchell 2022). Osborne analysed licensing and collaboration patterns, finding positively skewed patterns in contribution, engagement, and model usage (Osborrne, Ding, and Kirk 2024).

Research Questions

After careful perusal of repositories spanned by our literature review, we base our exploration and empirical analysis on Hugging Face, the most popular of contemporary PTM repositories (Gong et al. 2023) hosting over 0.7M projects at the time of the study. Hugging Face is increasingly appealing to AI developers, even over long-standing platforms like GitHub (Ait, Izquierdo, and Cabot 2023; Ait, Cánovas Izquierdo, and Cabot 2024). This is especially true as projects scale, requiring greater storage and computing resources. Hugging Face is a PaaS exclusively for AI/ML development, offering tooling, storage for large artifacts, and remote servers for training, testing, and hosting apps, all under a single roof. While some contemporary model directories provide official base model releases for specific libraries and frameworks (e.g., Nvidia CUDA, Tensorflow, or Pytorch model directories) or vetted research projects (e.g., ModelZoo), HF spans models from these categories alongside vast numbers of amateur submissions, community contributions, as well as public and private institutions. Therefore, it provides a large enough representative sample to observe development and model adaptation practices as is and gauge developer ethics in the wild.

RQ1: How is documentation of risks, limits, and biases among projects related to model evaluation? Evaluation

and Risks are core components of standard model cards. Hugging Face's guided annotation template (HuggingFace 2024b) encourages developers to select appropriate testing procedures and benchmarks to evaluate model performance and document the results. It also recommends that such evaluation involve testing for potential usage limitations, vulnerabilities, and biases in the model to aid sociotechnical experts in comprehensive risk documentation. Therefore, proper motivation, understanding, and proficiency in evaluation are expected to inculcate cognizance of responsible development practices. We frame the research question as follows:

$$Risks \quad and \quad Biases \quad Documented \quad \sim \quad Project\ Covariates + Evaluation\ Reported$$

RQ2: Do models reporting high accuracy also tend to be more thorough and transparent about their limitations? We may expect developers of highly accurate models to be more proficient and well-rounded, and therefore likelier to be able to thoroughly probe and document risks, biases, and other limitations. Yet, we explain how current trends may undermine the validity of benchmarking or even downplay the need for holistic evaluations above and beyond accuracy. RQ2 can be modeled as follows:

$$Risks \quad and \quad Biases \quad Documented \quad \sim \quad Project\ Covariates + Model\ Accuracy$$

We answer RQ2 by looking at submissions to Hugging Face's first edition of the Open LLM leaderboard, which ran from May 2023 to June 2024. It drew remarkable levels of participation across different project types. Importantly, it observed rigorous community monitoring for contamination (Balloccu et al. 2024) and other evaluation malpractices and reproducibility checks to substantiate self-reported performances, thus strengthening the validity of measurements and analysis.

Data

Here, we describe 1. the variables of interest (presence of risk/limitations in documentation) and 2. other project-level covariates we consider in our empirical data analysis and explain their inclusion, i.e., how they may motivate documentation habits and accountability among projects. Table. 1 lists details of our multi-source data collection.

Our review of prior studies and the HF Hub codebase and documentation revealed how crucial project information is distributed across the model landing page, repository, metadata, index tags, and finally, the model card markdown files. HF Hub uses semantic tags to index models and facilitate search, which are often auto-detected or parsed from the YML component of model cards. To access repository records or model tags, we use the Hugging Face API. A total of 700,072 repositories were hosted on Hugging Face as of 06/15/2024. We exclude gated repositories whose file contents or commit history are private. Since AI auditing and documentation particularly apply to service-ready models and AI applications (Arnold et al. 2019; Law and Krier 2023), we identify deployable models among these (See. Appendix), narrowing down to 456,545 projects.

Project use, Developer activity, and Community engagement: Controlling for model age lets us account for documentation practices as a function of evolving development standards, conception of ethical practices, and regulatory oversight. We measure this as the period between the project initiation (first commit) and the data collection timestamp. For developer and community engagement around a model, we measure the total number of commits, pull requests, and all other discussions (including issues) on each repo. Developers seeking greater exposure and usage of their projects are expected to practice better documentation (Gong et al. 2023). We use total likes from users as a cumulative measure of a model's popularity. Hugging Face allows model porting to build derived applications called Spaces, similar to GitHub's forking. We measure the uptake of the models through the number of spaces they support.

Application type: HF tracks information on the modalities and tasks performed for most service-ready models. These span six major types: Natural Language Processing, Computer Vision, Audio, Reinforcement Learning (RL), Tabular Data, and Multimodal. Note that a particular AI application may qualify under multiple categories, e.g., vision language models.

Developer attributes: The growing importance and evolving sophistication of documentation benefits from multiple contributors and distributed responsibilities (HuggingFace 2024b). In addition, information management may also depend on the type of developer or provider. In particular, commercial entities anticipating regulatory purview may ideally conduct more thorough risk assessments to meet regulatory requirements as well as avert potential liabilities from failures and misuse. We scrape developer profiles of respective models for information on team strength and affiliation, such as a for-profit company releasing 'freemium' models, an educational institution (university or classroom), or a nonprofit. For all developers, we also include the total number of models they contributed as a measure of experience.

Model Scale: In the context of AI, scaling refers to improving learnability and performance by developing highly parameterized and data-intensive models. Due to high investment and commercial uptake associated with them, large Foundation ('Frontier') models have received particular attention from ethicists and policy oversight bodies (Heim and Koessler 2024; Bender et al. 2021). Recent regulatory proposals, e.g., SB 1047 in California, have begun exploring graded requirements by model size. To inform ethical practice and test hypotheses around compliance behavior, scale variables control for emerging legal and social motivations from model valuation that may also influence evaluation and disclosure standards.

We represent scale through model size (number of parameters) and training data volume (number of samples). Non-uniform file nomenclature and frameworks complicate the automated parsing and tracking of model size (HF-Discuss 2024). Safetensors (HuggingFace 2024d) and GGUF (HuggingFace 2024a) are two popular tensor formats promoted and tracked by HF. Models and training checkpoints correctly stored in these formats display verified details on their pages, including the number of parameters (HuggingFace

2023). We obtained the parameter count for the 140,783 models stored in these formats.

Model index tags contain links to training data (if released), and Hugging Face provides size and other structured information on nearly all datasets it hosts. After selecting models that had released their training data on HF and screening out models directed to invalid dataset repositories, we obtained training data sizes for 17,260 models. *Overall, 7093 models had information on both model size and training data size*. One sample was excluded before analysis for being an outlier (Cook's D > 1).

Knowledge Domain: In the context of ML, this refers to the specific cases and tasks a model has learned to perform. Models are generally trained with data samples from their target domain, and high-stakes/ critical applications may command greater developer accountability. For example, minor diagnostic errors can significantly increase liabilities and derail the applicability of AI in medicine. HF tracks multiple popular training data domains, including medical, finance, code generation, etc, through the link training datasets (if available). Domain tags were also collected for these 7093 models.

Compliance Information: Based on project objectives and intended model use, developers should choose appropriate tests and metrics to quantify model performance. Our first hypothesis test relationships between documentation of model risk and limitations and the frequency of model evaluation. Model cards are the default face of a model's landing page on HF, rendered from the repository's 'README.md' file. They are designed to present several sections for technical information, such as data provenance, development specifications, performance, legal/copyright aspects, and social implications of the model's use. A repository lacking a "README.md" model card will display a blank landing page. We measure such projects as non-compliant in documenting evaluation, model risks, or environmental impact.

RQ1 focuses on two major sections. Based on the official HF Annotation guide, the Evaluation section requires the model developer to specify testing objectives, protocols, and performance results. Ideally, these should be selected to test both the accuracy of performance and the safety of use (e.g., fairness across relevant user groups, security risks, and foreseeable error contexts). The findings of all safety assessments should be filled in under the "Bias, Risks, and Limitations" (HuggingFace 2024b).

HF's society and ethics team recently developed a regulatory tool to scan model cards to check if certain sections have been filled out. We scanned all model cards to detect whether evaluations or risk assessments have been filled out. Some integrated libraries, e.g., Autotrain[1], initialize default model cards based on the HF template but only contain placeholder text (such as "More information needed"). To preserve analytical integrity, we detect and screen out these nonsubstantive texts through pattern matching[2].

CO_2 footprint assessments are another crucial sociotechnical component of model cards (Lacoste et al. 2019; Strubell, Ganesh, and McCallum 2020) and are often included by conscientious developers in general. These reports are also controlled for due to potential confounding with risk and limitations disclosures. We parse model card metadata followed by string matching to detect valid CO_2 emission entries under designated sections in the model card text.

Competitive Benchmarking: Over time, various leaderboards have been created to test different AI applications. Leaderboard ranks are vied, and high performers enjoy considerable visibility and popularity. The Open LLM Leaderboard[3] is the most prominent and active leaderboard on Hugging Face, with its first edition running from May 2023 to June 2024. It was mainly geared towards language technologies and ranked submissions on aggregate performance across six extremely popular benchmarks (Cobbe et al. 2021; Sakaguchi et al. 2021; Clark et al. 2018; Zellers et al. 2019; Hendrycks et al. 2021; Lin, Hilton, and Evans 2022). With 7173 unique, complete submissions, the leaderboard encourages performance validation while supporting community-informed model selection.

We use leaderboard archives to collect details on participating models. Submissions were ranked by their aggregate performance across these six popular benchmarks.

- **AI2 Reasoning Challenge (ARC)** (Clark et al. 2018): Grade-school science questions (25-shot)
- **HellaSwag** (Zellers et al. 2019): Commonsense inference, challenging for SOTA models but easy for humans (10-shot)
- **MMLU** (Hendrycks et al. 2021): Multitask accuracy across 57 tasks, including mathematics, history, law, and more (5-shot)
- **TruthfulQA** (Lin, Hilton, and Evans 2022): Measures model's propensity to reproduce common online falsehoods (0-shot)
- **Winogrande** (Sakaguchi et al. 2021): An adversarial benchmark for commonsense reasoning (5-shot)
- **GSM8k** (Cobbe et al. 2021): Diverse grade school math word problems to test multi-step mathematical reasoning (5-shot)

Developers often submit multiple entries to report incremental increases in test accuracy, while effectively inducing overfitting (Ethayarajh and Jurafsky 2020; Hardt 2017). Such may be a marker of competitiveness rather than sustainable development. For our analysis, we only consider the best performance for each model, controlling for the number of attempts and the floating point precision of the tested model version. We additionally control for evaluation malpractice through a binary predictor indicating flagged models.

Developers or platform moderators often assign 'not-for-all-audience' and 'NSFW' tags to certain projects inappro-

[1]https://huggingface.co/autotrain

[2]Following a survey of placeholder texts generated by HF supported libraries (Appendix) and their removal from data, we find a much lower frequency of actual risk documentation than previously reported (Liang et al. 2024), adjusted as (model cards documenting risks/total models carrying model cards) for even comparison

[3]https://huggingface.co/datasets/open-llm-leaderboard-old/results

Project Aspect	Variables	Description	Type	Source
Model Features	Model Size	Number of Model Parameters	Numeric	Model Page
	Training Resources	Data samples used to train model	Numeric	Training Data metadata (HF API)
	Modalities	Modalities served e.g. Computer Vision	Categorical	Model Card metadata (HF API)
	Domain	Specific fields of application model is trained, e.g., code analysis, medical applications	Categorical	Training Data metadata (HF API)
Model Developer	Team Size	Community Strength	Numeric	Linked Developer Profile
	Total Models	Development experience of contributor	Numeric	Linked Developer Profile
	Entity Type	If contributor is a for-profit or non profit, research projects, etc	Categorical	Linked Developer Profile
User Engagement	Likes	Total Likes from HF Users	Numeric	Model Page
	Deployed Apps	Number of apps on HF using model	Numeric	Model Page
Project Activity	Age	Repository Age in days	Numeric	Git History (HF API)
	Total Commits	Development activity on repository	Numeric	Git History (HF API)
	Pull Requests	Feature Additions and Contributions received	Numeric	Git History (HF API)
	Discussions	Community feedback and engagement with repo	Numeric	Git History (HF API)
Compliance	Performance Evaluation	Developer's evaluation objectives, protocols selected and results	Categorical	HF Model Card scanner and API
	Risks, Limitations and Biases	Foreseeable harms, vulnerabilities and limitations	Categorical	HF Model Card scanner
	CO_2 Emissions	Model training footprint on environment	Categorical	HF Model Card scanner and API
Competitive Benchmarking	Accuracy	Best aggregate results reported on the Open LLM Leaderboard	Numeric	Leaderboard Archives
	Attempts	Number of leaderboard submissions for a single model	Numeric	Leaderboard Archives
	Precision	Precision used in testing e.g. 8 Bit, BF16 etc	Categorical	Leaderboard Archives

Table 1: Data collection across Hugging Face: Variables with description.

priate for general use, such as ones trained on or meant for sexual content generation[4]. By virtue of violating the fundamental premise of ethical AI, they are expected to show limited compliance. High risk applications were incorporated as a categorical control in our analysis.

Exploratory Study: Trends on Hugging Face

Before answering our quantitative RQs (covered in the next section), we conduct a preliminary analysis of development trends on Hugging Face and share key findings. Hugging Face was launched through an open source implementation (Wolf et al. 2020) of the seminal transformer architecture (Vaswani 2017) for Natural Language Processing. The first version of their client library was released in December 2020[5] to foster remote collaboration, artifact storage, reuse, and sharing.

Examining models by modality and application category tags, we find Natural Language Processing dominates, followed by Reinforcement Learning, Computer Vision, and Audio applications. The democratization of innovation through collaborative platforms and rapid progress in AI have also paved the way for training libraries and solutions that greatly ease and accelerate development. As of the time of the study, HF supported over 20 libraries and offered

[4]https://huggingface.co/content-guidelines

[5]https://pypi.org/project/huggingface-Hub/

off-the-shelf options ranging from industrial-scale foundation models to easy, low-compute customization. The empowerment of individual developers was evident, as around 87.57% of all the 456,545 service-ready projects were contributed by individuals, and a staggering 86.34% of them were built without receiving any collaborative input through pull requests. At the same time, only 5.46% of all these projects saw any downstream use in application development.

We obtained verifiable model sizes (in parameters) and training data samples for 140,783 and 17,260 of the service-ready projects, respectively. Examining the temporal evolution of model sizes and training data requirements among post-2020 uploads, we observed a clear upward trajectory in development scale, favoring more sophisticated and data-intensive models.

Yet, transparency among service-ready models was generally low. Around 15.9% of the 456,545 models carried any form of evaluation, while risks and limitations were found among 2.2%. Finally, CO_2 emissions saw the least reporting at 0.7%. Only around 0.1% models complete all three sections. These findings broadly agree with trends seen in prior work (Liang et al. 2024; Yang, Liang, and Zou 2023) and call for greater attention to documentation and comprehensiveness across all reporting requirements.

Main Analysis

Sample Selection: What counts as Open Source AI?

The definition of open source AI is contentious (Liesenfeld and Dingemanse 2024). Unlike traditional software, which is mostly code, AI systems comprise multiple components. These include code, training data, model weights, documentation, and different projects exercise varying degrees of openness concerning each. Liesenfeld et al. particularly document "open-washing," whereby for-profits promote AI systems as open source by releasing only code or weights, while withholding critical components like training data or specific training approaches. This promotes models through unrestricted use ("freemium"), helps avail compliance exemptions generally reserved for OSS, but limits experimentation, reproducibility, and reverse engineering, which have been core to the OSS movement (Free Software Foundation 2024). IP (Hutchinson et al. 2021) and security (Li et al. 2023) are other reasons behind the withholding of training data.

In view of the fast-changing landscape, we carefully studied the latest deliberations from both the Open Source Initiative (Open Source Initiative 2024) and the Free Software Foundation (Free Software Foundation 2024), the two main authoritative OSS bodies. Both highlight data transparency as core to open source AI. While OSI expects detailed documentation, data provenance, and disclosure of data sources, FSF insists on the full release of data alongside weights and code. Finally, we took the conservative approach of pursuing our questions through projects that span both interpretations. This left us with 7092 samples for RQ1, a small subset of the service-ready models, whose data and other essential model specifics (including number of parameters) were known. Around 23.19%, 7.86% and 2.04% among these had provided evaluation, risk assessments, and CO_2 emission data, respectively.

RQ2 is based on a subset of RQ1, i.e. 789 models participating in the open LLM leaderboard and only comprises NLP models. In this particular subset where every model has been evaluated (at least through benchmark participation), we again find higher than average risk (8.7%) and carbon emission (1.5%) reporting. While 7173 models completed the full set of six benchmarks, we test RQ2 on a smaller subset of 789 models that satisfy both OSI and FSF criteria.

Multivariate Hypothesis Testing for RQ1 and RQ2

We frame our RQs as binary prediction modeling to determine if risk documentation is significantly associated with 1. rates of performance evaluation and disclosure, and 2. absolute mean performance on a set of very popular benchmarks used by the OpenLLM Leaderboard. For both cases, we model the likelihood of risk assessment in model cards using binomial logit models, where the frequency of evaluation (RQ1) or reported performance (RQ2) are the main regressors of interest, adjusting for crucial project-level covariates. We set the significance level of our analysis at 0.01. In RQ2, some generalist models spanned multiple domains, such as medicine and legal/financial applications, leading to the aliasing. These knowledge domains were merged into a single category and renamed "multi-domain."

Numeric covariates were log-transformed (base 10) for skew correction and comparison along the scale of different projects, followed by standardization. We check for multicollinearity using the *car* package from R, removing variables with VIF factor > 5. This excluded the model domains 'music' and 'art' from RQ1 and 'Biology,' 'Chemistry' and precision category 'Torch Float16' from RQ2. We checked for high-leverage outliers based on Cook's distance ($D > 4/N$) and standard residuals (> 3), and removed 1 data point each from both analyses. Using the Box-Tidwell approach (Box and Tidwell 1962), suitable higher-order transformations (See Table. 2 and Table. 3) were performed on some variables to ensure that assumptions of linearity between log odds and the predictors held. Compared to simpler, more interpretable models, the models with power-transformed variables were ultimately preferred for the final reporting due to greater explainability (AIC from 374.4 to 371.6 for RQ2 and from 3454.3 to 3410.9) and validity. Encouragingly, the significant effects and their directionality remain largely preserved across both approaches, confirming the robustness of the results[6].

Finally, residual tests were performed using the *DHARMa* package. Neither regression model showed significant dispersion (RQ1:$p = 0.82$ and RQ2:$p = 0.85$), presence of outliers (RQ1:$p = 0.18$ and RQ2:$p = 0.42$), or deviation from normal distributions (KS test; RQ1:$p = 0.78$ and RQ2:$p = 0.78$).

[6]All effects significant in the transformed models are also significant in the simpler models, except for the number of models built by the developer and likes in RQ1, which do not appear significant prior to transformation

	Predictor	Coefficient	p-value
	(Intercept)	-3.263988	$<$**0.0001**
Scale	Parameters [1]	0.120302	0.026334
	Data size [1]	0.179451	**0.000169**
Modality	Audio	-0.949783	**0.003028**
	Computer Vision	0.815687	0.014426
	Multimodal	-1.288032	0.209746
	Natural Language Processing	0.384321	0.019977
	RL	-13.517337	0.984074
Domain	Biology	0.068115	0.914340
	Chemistry	0.320644	0.713684
	Climate	1.646929	0.324356
	Code	-0.483826	0.135364
	Finance	-0.174213	0.796490
	Legal	0.191019	0.720080
	Medical	-1.235878	0.035409
Developer Team	Team members [1]	0.193956	**0.000149**
	Total models [2]	0.248375	$<$ **0.0001**
	Company	-0.204922	0.273141
	University	-0.005852	0.983995
	Classroom	0.623096	0.445068
	Non-profit	0.529477	0.021068
Use and Popularity	Likes [3]	0.174444	**0.001643**
	Number of Spaces [1]	-0.015841	0.713853
Project Activity	Total Commits [2]	-0.359598	$<$ **0.0001**
	Threads [1]	0.090050	0.026619
	PR [1]	0.001501	0.974375
	Repository age [2]	0.054596	0.268635
Compliance	CO_2 footprint?	2.177332	$<$ **0.0001**
	Evaluation Present?	0.913310	$<$ **0.0001**
Others	High Risk Model	-13.834954	0.968635
		N= 7092 R^2 = 0.115 AIC = 3411	

[1] Log transformed (base 10) and Standardized
[2] Log (base 10), $1/x$ and Standardized
[3] Log (base 10), $x^{0.3}$ and Standardized

Table 2: Test statistics for binomial logistic regression of limits, bias, and risks documentation rates among models based on 1. their project attributes, 2. rates of compliance with related components of the Model Card

	Predictor	Coefficient	p-value
	(Intercept)	-2.8854	$<$ **0.0001**
Scale	Parameters [2]	0.6695	**0.000803**
	Data size [1]	-0.1617	0.371708
Domain	Multi-domain	17.3876	0.987295
	Code	-16.6237	0.987853
	Medical	0.5741	0.730284
Developer Team	Team members [1]	1.0562	$<$ **0.0001**
	Total models [1]	-0.6927	**0.000257**
	Company	-1.6773	**0.003041**
	University	-0.2654	0.714144
	Non profit	-1.3751	0.173400
Use and Popularity	Likes [1]	-0.3491	0.194646
	Number of Spaces [1]	0.2701	0.155802
Project Activity	Total Commits [1]	0.8053	$<$ **0.0001**
	Threads [1]	0.1761	0.427108
	PR [1]	-0.1766	0.162182
	Repository age [1]	-0.0203	0.920262
Compliance	CO_2 footprint?	2.3698	**0.001487**
OpenLLM Evaluation	Accuracy [3]	-0.7631	**0.001124**
	Flagged	-0.2596	0.796655
	Attempts [1]	0.3128	0.038215
Precision	4 bit	-18.0137	0.993056
	8 bit	-0.3797	0.802545
	Torch BFloat16	0.3294	0.316380
Others	High Risk Model	-15.3855	0.994607
		N= 789 R^2 = 0.272 AIC = 371.636	

[1] Log transformed (base 10) and Standardized
[2] Log (base 10), $x^{4.5}$ and Standardized
[3] Standardized

Table 3: Test statistics for binomial logistic regression of limits, bias, and risks documentation rates among models based on 1. features of leaderboard models 2. competitive performance of the models

Interpretation Our analysis from RQ1 confirms a strong association between evaluation and risk documentation, with models reporting some form of evaluation being 149.2% more likely to also carry information on model risks and limits. Other positive effects come from training data size, high commit activity, documentation of CO_2 footprint, developer team size, and popularity (likes). Audio applications and models associated with high contributors (more models) are also less likely to carry risk documentation.

Meanwhile, RQ2 finds that high performers on the Open LLM Leaderboard are less likely to document risks and limitations. One standard unit increase in accuracy reduced risk reporting chances by 53.4%. Greater model size (parameters), documentation of CO_2 footprint, high number of commits, and developer team size also predict higher chances of a project carrying such documentation. At the same time, companies and high contributors are less likely to do the same. Interestingly, specific model knowledge domains do not exert any significant effect across both analyses, i.e., risk reporting rates are relatively the same across high-stake applications such as medicine or finance, niches such as code, and all other general domains.

Discussion

Evaluation is core to responsible AI. It is essential to determine model capabilities and for responsible development through understanding and acknowledgment of risks and limitations. Our analyses of OSS practitioners reaffirm that evaluation and risk assessment generally go hand in hand. However, we also observed that unidirectional focus on model performance, typical of competitive leaderboards, can undermine developer accountability.

Specific other observations were consistent across both analyses. As one might anticipate, development at scale (data-intensive training or parameterization) positively correlates with compliance, likely due to generally greater oversight. Documentation of risks is also closely associated with general ethical awareness, as evident through the strong effect from reporting of CO_2 footprint. Projects with more activity, contributions, and larger teams tended to do a better job with risk reporting. On the other hand, prolific developers appear to pay less attention to product safety and documentation. Informed by these trends, we present our recommendations for contributors, entrepreneurs, and AI hosting services. These include practices and interventions to encourage documentation overall and to improve the efficacy of evaluation protocols in informing both model strengths and weaknesses.

Recommendations

Evaluation Prerogatives: Data providers and platforms hosting leaderboards need to consider multi-faceted tasks and metrics for emerging considerations beyond simply performance. The selection of tests and metrics is generally left to the developer's discretion. While HF model cards mention that evaluation choices should address social impact, there is indeed a gap, as standards and governance mechanisms are still evolving and are yet to bridge ethical concerns and normative practice.

Platforms should strive to adapt to these guidelines as they evolve. Spreading awareness on the different dimensions of AI risks (social and environmental) and communicating ethical and regulatory expectations clearly and precisely through practitioner manuals and training modules could significantly impact OSS reporting practices. Similarly, promoting well-documented models (Liang et al. 2024), comprehensive testing suites, and bias bounties (Liang et al. 2023; Sun et al. 2024; Globus-Harris, Kearns, and Roth 2022), and messaging on the importance of quality and safety of models (over quantity) can improve overall developer accountability.

Promoting Ethical Practice: As one of the leading open source AI hosting services, Hugging Face sets several commendable precedents through its initiatives. These include steps to inculcate responsible documentation, monitor compliance[7], and to keep up with regulation (HuggingFace 2024c). Results from our empirical analyses suggest that risk documentation practices are more prevalent among large teams, while most contributions come from individual developers. Meanwhile, model card guidelines used by Hugging Face[8] and other notable institutions[9] are detailed to facilitate auditing, and usually designate specific tasks across developers, sociotechnical experts, and managers. Risk assessments involve multiple roles and can make compliance overwhelming for small teams or individual contributors. Streamlining, such as outlining priorities and accessible testing tools, may make risk documentation more approachable.

Reforming Leaderboards: HF's open LLM leaderboard is a massive undertaking, supported by collaborative monitoring labor from the community and moderators. It is notably more transparent and comprehensive than conventional leaderboards, and tracks model size, precision, libraries, and architectures of most submissions. Multi-metric leaderboards such as SNLI which displays model sizes alongside accuracy (Bowman et al. 2015; Group 2015), can guide developers and users towards efficient and sustainable choices. They support explainability, promote sustainable models, and inform judicious model selection for small-scale, less resourced developers and users.

Opportunistic overfitting can be mitigated by incorporating dynamic benchmarks (Nie et al. 2020; Dehghani et al. 2021; Kiela et al. 2021; Gehrmann et al. 2021), that are constantly updated to accommodate temporal/distributional drifts and emerging domains, newer tasks, and capabilities. Dehghani et al. and Hardt et al. further guide benchmark design to ensure judicious assessment while mitigating opportunistic submissions (Dehghani et al. 2021; Hardt 2017; Blum and Hardt 2015). Other proposed measures include confidentiality of test/hold out sets and mitigation of data leakage (Lilja et al. 2024; Deng et al. 2024) and contaminated models (Dwork et al. 2015; Balloccu et al. 2024; Magar and Schwartz 2022).

[7] https://huggingface.co/society-ethics

[8] https://huggingface.co/docs/hub/en/model-card-annotated

[9] https://ai-toolkit.xd.gov/resources/model-card-generator/

Conclusion

Through our focused study of a rising open source platform, we had the opportunity to observe a diverse range of AI/ML applications and development practices. Our analyses empirically probe open source AI trends in the backdrop of increasing concerns over their potential to transform or affect society, and consequent legal and ethical oversight. As we situate our investigation amidst the interests of these various stakeholders, we discover promising trends of concurrent compliance of evaluations and risk assessments. At the same time, our large sample study produces evidence supporting long-standing observations and calls for fundamental reforms and greater rigor in AI evaluation. As AI continues to grow, these lessons emphasize the importance of fostering a culture of responsible development and accountability across all sectors, not only commercial but also informal and non-profit undertakings. Platforms, developers, and stakeholders must work together to establish best practices and design balanced policies and standards that mutually support each other while also preserving the true spirit of innovation. This will be vital in ensuring that AI technologies are developed and deployed ethically, safely, and benefit humanity at large.

Limitations

As consensus on technical governance approaches and specific requirements for OSS continues to evolve, we do not evaluate whether the risk assessments provided in a model card are sufficient for a given model. Future research on evaluation protocols and safety standards will benefit from continued exploratory studies of practitioners, particularly how specific tests succeed in risk mitigation and promote user trust.

Hugging Face's popularity, participant sample size, and moderation make their leaderboards amenable to our research questions. Most leaderboards cater to a specific domain and set of tasks, and submissions are generally uniform in modality. The open LLM leaderboards are primarily intended for language models. Yet the patterns of developer behavior they reflect can provide crucial insight into rapidly growing technologies and can translate across platforms and modalities. We look forward to future studies on improved, up-and-coming leaderboards for further validation of our findings and to inform evaluation practices as we advance.

Appendix

Service-ready Features and Identifiers

Based on a comprehensive review of platform documentation and widgets/third-party features supporting direct or downstream applications, we identify service-ready models as ones with at least one of the following:

- Filled out Model card sections for detailed instructions, examples, and use cases: Detected using HF's Model card scanner
- **Use this Model button**: Platform-generated example scripts to guide model loading and use through recognized libraries.
- **Endpoints compatible feature**: this tag indicates the model supports Inference Endpoints and hence is scalable and production-ready
- **Pipeline tag**: denotes the specific task a model was designed for, such as "text-classification", or "object-detection", and is auto-detected or indicated by the developer.
- **Autotrain compatible**: indicates if a project is a complete pre-trained model and compatible within the HF ecosystem for downstream fine-tuning on custom data.
- **Text-embeddings-inference**: Allows generation of text embeddings at scale from compatible models.
- **Text-generation-inference**: A runtime/ web widget to handle generation queries to a model.

A total of 456,545 projects out of the 700,072 repositories scanned fulfilled this criteria.

Model Card Boilerplate Texts

- "[More Information Needed]"/ "More information needed."/"More information needed"
- "Users (both direct and downstream) should be made aware of the risks, biases and limitations of the model"
- "More information needed for further recommendations."
- "Users (both direct and downstream) should be made aware of the risks, biases and limitations of the model. More information needed for further recommendations."

Ethics Statement

We largely followed prior work for our data collection and used the public Hugging Face API for model cards and open repository data. Beyond the API, we collected some limited public-facing numeric data from model landing pages, which are intended for public reference and sharing with no expectation of privacy (Table. 1). We did not add any features to our data collection code for specifically parsing personally-identifying information, nor was such information required for our analysis. The only identifiers used were public developer usernames, which are also part of the model path in the HF web indexing. Finally, we note that some models carry "Not Safe For Work" and "Not for all Audiences" warnings from developers or the platform moderators (See Hugging Face Content Guidelines). We retain these labels in our dataset so that any researchers wishing to use this data can be fully informed about the potential for some model metadata to contain content inappropriate for some settings. Proprietary LLM-based language editing services, Grammarly and Anthropic Claude, were used to a limited extent to correct misspellings, grammar, and consistency of composition, and the resulting manuscript was thoroughly verified and updated by all the authors over multiple iterations.

Acknowledgments

This study was supported by the National Science Foundation under the Growing Convergence Research Scheme (Grant No #2020751).

References

AIID. 2024. Artificial Intelligence Incident Database - Discover — incidentdatabase.ai. https://incidentdatabase.ai/apps/discover/?is_incident_report=true&s=open%20source. [Accessed 11-09-2024].

Ait, A.; Cánovas Izquierdo, J. L.; and Cabot, J. 2024. HF-Community: An extraction process and relational database to analyze Hugging Face Hub data. *Science of Computer Programming*, 234: 103079.

Ait, A.; Izquierdo, J. L. C.; and Cabot, J. 2023. On the Suitability of Hugging Face Hub for Empirical Studies. ArXiv:2307.14841 [cs].

Arnold, M.; Bellamy, R. K.; Hind, M.; Houde, S.; Mehta, S.; Mojsilović, A.; Nair, R.; Ramamurthy, K. N.; Olteanu, A.; Piorkowski, D.; et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5): 6–1.

Arora, S.; and Zhang, Y. 2021. Rip van Winkle's Razor: A Simple Estimate of Overfit to Test Data. *arXiv preprint arXiv:2102.13189*.

Balloccu, S.; Schmidtová, P.; Lango, M.; and Dušek, O. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 67–93.

Bartlett, P. L.; Montanari, A.; and Rakhlin, A. 2021. Deep learning: a statistical viewpoint. *Acta numerica*, 30: 87–201.

Beeching, E.; Fourrier, C.; Habib, N.; Han, S.; Lambert, N.; Rajani, N.; Sanseviero, O.; Tunstall, L.; and Wolf, T. 2023. Open LLM Leaderboard (2023-2024).

Bender, E. M.; Bender, E. M.; Friedman, B.; and Friedman, B. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. Virtual Event Canada: ACM. ISBN 978-1-4503-8309-7.

Benkler, Y. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press. ISBN 9780300110562.

Berman, G.; Goyal, N.; and Madaio, M. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery.

Birhane, A.; prabhu, v.; Han, S.; Boddeti, V.; and Luccioni, S. 2023. Into the LAION's Den: Investigating Hate in Multimodal Datasets. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 21268–21284. Curran Associates, Inc.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Blum, A.; and Hardt, M. 2015. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, 1006–1014. PMLR.

Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huang, S.; Huck, M.; Koehn, P.; Liu, Q.; Logacheva, V.; et al. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, 169–214.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2022. On the opportunities and risks of foundation models. *ACM Computing Surveys*, 55(5): 1–166.

Bommasani, R.; Klyman, K.; Longpre, S.; Kapoor, S.; Maslej, N.; Xiong, B.; Zhang, D.; and Liang, P. 2023. The Foundation Model Transparency Index. ArXiv:2310.12941 [cs].

Bordia, S.; and Bowman, S. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 7–15.

Bowman, S.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642.

Bowman, S.; and Dahl, G. 2021. What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4843–4855.

Box, G. E.; and Tidwell, P. W. 1962. Transformation of the independent variables. *Technometrics*, 4(4): 531–550.

Castaño, J.; Martínez-Fernández, S.; and Franch, X. 2024. Lessons Learned from Mining the Hugging Face Repository. In *Proceedings of the 1st IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering*, 1–6.

Chakraborti, M.; Atkisson, C.; Stănciulescu, Ş.; Filkov, V.; and Frey, S. 2024. Do We Run How We Say We Run? Formalization and Practice of Governance in OSS Communities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–26.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Coleman, G. 2013. Coding freedom: The ethics and aesthetics of hacking. *Princeton University Press*.

Contractor, D.; McDuff, D.; Haines, J. K.; Lee, J.; Hines, C.; Hecht, B.; Vincent, N.; and Li, H. 2022. Behavioral use licensing for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 778–788.

Crisan, A.; Drouhard, M.; Vig, J.; and Rajani, N. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 427–439. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.

Crowston, K.; Wei, K.; Howison, J.; and Wiggins, A. 2012. Free/Libre open source software development: What we know and what we do not know. *ACM Computing Surveys (CSUR)*, 44(2): 1–35.

Dehghani, M.; Tay, Y.; Gritsenko, A. A.; Zhao, Z.; Houlsby, N.; Diaz, F.; Metzler, D.; and Vinyals, O. 2021. The Benchmark Lottery. ArXiv:2107.07002 [cs].

Deng, C.; Zhao, Y.; Tang, X.; Gerstein, M.; and Cohan, A. 2024. Benchmark Probing: Investigating Data Leakage in Large Language Models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning - The Good, the Bad, and the Ugly*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, W. H.; Nagireddy, M.; Lee, M. S. A.; Singh, J.; Wu, Z. S.; Holstein, K.; and Zhu, H. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 473–484. New York, NY, USA: Association for Computing Machinery.

DoJ. 2024. Man Arrested for Producing, Distributing, and Possessing AI-Generated Images of Minors Engaged in Sexually Explicit Conduct — justice.gov. https://www.justice.gov/opa/pr/man-arrested-producing-distributing-and-possessing-ai-generated-images-minors-engaged. [Accessed 27-08-2024].

Dwork, C.; Feldman, V.; Hardt, M.; Pitassi, T.; Reingold, O.; and Roth, A. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248): 636–638.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Ethayarajh, K.; and Jurafsky, D. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4846–4853. Online: Association for Computational Linguistics.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Text with EEA relevance.

Falco, G.; Falco, G.; Shneiderman, B.; Shneiderman, B.; Badger, J.; Badger, J.; Carrier, R.; Carrier, R.; Dahbura, A.; Dahbura, A. T.; Danks, D.; Danks, D.; Eling, M.; Eling, M.; Goodloe, A. E.; Goodloe, A.; Gupta, J. P.; Gupta, J.; Hart, C.; Hart, C.; Jirotka, M.; Jirotka, M.; Johnson, H.; Johnson, H.; Lapointe, C.; LaPointe, C.; Llorens, A. J.; Llorens, A. J.; Mackworth, A. K.; Mackworth, A. K.; Maple, C.; Maple, C.; Pálsson, S. E.; Pálsson, S. E.; Pasquale, F. A.; Pasquale, F. A.; Winfield, A. F. T.; Winfield, A. F. T.; Yeong, Z. K.; and Yeong, Z. K. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence*.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28: 689–707.

Franzen, C. 2024. New open source AI leader Reflection 70B's performance questioned, accused of 'fraud' — venturebeat.com. https://venturebeat.com/ai/new-open-source-ai-leader-reflection-70bs-performance-questioned-accused-of-fraud/. [Accessed 12-09-2024].

Free Software Foundation. 2024. FSF is working on freedom in machine learning applications. *Free Software Foundation News*. Accessed: 2025-05-24.

Gebru, T.; Gebru, T.; Morgenstern, J.; Morgenstern, J.; Vecchione, B.; Vecchione, B.; Vecchione, B.; Vaughan, J.; Vaughan, J. W.; Wallach, H.; Wallach, H.; Daumé, H.; Daumé, H.; Crawford, K.; and Crawford, K. 2018. Datasheets for Datasets. *arXiv: Databases*.

Gehrmann, S.; Adewumi, T.; Aggarwal, K.; Ammanamanchi, P. S.; Aremu, A.; Bosselut, A.; Chandu, K. R.; Clinciu, M.; Das, D.; Dhole, K.; et al. 2021. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, 96–120.

Gema, A. P.; Leang, J. O. J.; Hong, G.; Devoto, A.; Mancino, A. C. M.; Saxena, R.; He, X.; Zhao, Y.; Du, X.; Madani, M. R. G.; Barale, C.; McHardy, R.; Harris, J.; Kaddour, J.; van Krieken, E.; and Minervini, P. 2024. Are We Done with MMLU? ArXiv:2406.04127 [cs].

Globus-Harris, I.; Kearns, M.; and Roth, A. 2022. An Algorithmic Framework for Bias Bounties. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1106–1124. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.

Gong, L.; Zhang, J.; Wei, M.; Zhang, H.; and Huang, Z. 2023. What Is the Intended Usage Context of This Model? An Exploratory Study of Pre-Trained Models on Various Model Repositories. *ACM Trans. Softw. Eng. Methodol.*, 32(3). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Group, T. S. N. L. P. 2015. SNLI Leaderboard. https://nlp.stanford.edu/projects/snli/. [Accessed 07-08-2024].

Hardt, M. 2017. Climbing a shaky ladder: Better adaptive risk estimation. *arXiv preprint arXiv:1706.02733*.

Heim, L.; and Koessler, L. 2024. Training Compute Thresholds: Features and Functions in AI Regulation. ArXiv:2405.10799 [cs].

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

HF-Discuss. 2024. Add sorting option by model size [New Feature Proposal] — discuss.huggingface.co. https://discuss.huggingface.co/t/add-sorting-option-by-model-size-new-feature-proposal/29085. [Accessed 09-09-2024].

Hoffmann, M.; Nagle, F.; and Zhou, Y. 2024. The value of open source software. *Harvard Business School Strategy Unit Working Paper*, (24-038).

HuggingFace. 2023. Safetensors params/precision on model page — discuss.huggingface.co. https://discuss.huggingface.co/t/safetensors-params-precision-on-model-page/67913/3. [Accessed 09-09-2024].

HuggingFace. 2024a. GGUF — huggingface.co. https://huggingface.co/docs/hub/en/gguf. [Accessed 09-09-2024].

HuggingFace. 2024b. Model Card Guidebook https://huggingface.co/docs/hub/en/model-card-guidebook.

HuggingFace. 2024c. Public Policy at Hugging Face — huggingface.co. https://huggingface.co/blog/policy-blog. [Accessed 12-09-2024].

HuggingFace. 2024d. Safetensors — huggingface.co. https://huggingface.co/docs/safetensors/en/index. [Accessed 09-09-2024].

Hutchinson, B.; Smart, A.; Hanna, A.; Denton, E.; Greer, C.; Kjartansson, O.; Barnes, P.; and Mitchell, M. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 560–575.

Jia, R.; and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jiang, W.; Jones, J.; Yasmin, J.; Synovic, N.; Sashti, R.; Chen, S.; Thiruvathukal, G. K.; Tian, Y.; and Davis, J. C. 2023a. PeaTMOSS: Mining Pre-Trained Models in Open-Source Software. ArXiv:2310.03620 [cs].

Jiang, W.; Synovic, N.; Hyatt, M.; Schorlemmer, T. R.; Läufer, K.; Lü, Y.; Thiruvathukal, G. K.; and Davis, J. C. 2023b. An Empirical Study of Pre-Trained Model Reuse in the Hugging Face Deep Learning Model Registry. *International Conference on Software Engineering*.

Jiang, W.; Synovic, N.; Jajal, P.; Schorlemmer, T. R.; Tewari, A.; Pareek, B.; Thiruvathukal, G. K.; and Davis, J. C. 2023c. PTMTorrent: A Dataset for Mining Open-source Pre-trained Model Packages. *IEEE Working Conference on Mining Software Repositories*.

Jiang, W.; Synovic, N.; Sethi, R.; Indarapu, A.; Hyatt, M.; Schorlemmer, T. R.; Thiruvathukal, G. K.; and Davis, J. C. 2022. An empirical study of artifacts and security risks in the pre-trained model supply chain. In *Proceedings of the 2022 ACM Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses*, 105–114.

Jobin, A.; Jobin, A.; Ienca, M.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*.

Kathikar, A.; Nair, A.; Lazarine, B.; Sachdeva, A.; and Samtani, S. 2023. Assessing the Vulnerabilities of the Open-Source Artificial Intelligence (AI) Landscape: A Large-Scale Analysis of the Hugging Face Platform. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 1–6.

Kawakami, A.; Wilkinson, D.; and Chouldechova, A. 2024. Do Responsible AI Artifacts Advance Stakeholder Goals? Four Key Barriers Perceived by Legal and Civil Stakeholders. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery.

Kiela, D.; Bartolo, M.; Nie, Y.; Kaushik, D.; Geiger, A.; Wu, Z.; Vidgen, B.; Prasad, G.; Singh, A.; Ringshia, P.; et al. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4110–4124.

Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lam, K.; Lange, B.; Blili-Hamelin, B.; Davidovic, J.; Brown, S.; and Hasan, A. 2024. A Framework for Assurance Audits of Algorithmic Systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1078–1092. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Law, H.; and Krier, S. 2023. Open-source provisions for large models in the AI Act. Publisher: Cambridge University Science and Policy Exchange.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lee, H.-P. H.; Yang, Y.-J.; Von Davier, T. S.; Forlizzi, J.; and Das, S. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300. Event-place: Honolulu, HI, USA.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9): 1–46.

Li, R.; Pandurangan, P.; Frluckaj, H.; and Dabbish, L. 2021. Code of conduct conversations in open source software projects on github. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–31.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2023. Holistic evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 36.

Liang, W.; Rajani, N.; Yang, X.; Ozoani, E.; Wu, E.; Chen, Y.; Smith, D. S.; and Zou, J. 2024. Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI. *Nature Machine Intelligence*, 6(7): 744–753.

Liao, Q. V.; Singh, M.; Zhang, Y.; and Bellamy, R. 2021. Introduction to Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380959.

Liesenfeld, A.; and Dingemanse, M. 2024. Rethinking open source generative AI: open-washing and the EU AI Act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1774–1787.

Lilja, A.; Fu, J.; Stenborg, E.; and Hammarstrand, L. 2024. Localization is all you evaluate: Data leakage in online mapping datasets and how to fix it. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22150–22159.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252.

Magar, I.; and Schwartz, R. 2022. Data Contamination: From Memorization to Exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 157–165.

Manzini, T.; Lim, Y. C.; Tsvetkov, Y.; and Black, A. W. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Marchal, N.; Xu, R.; Elasmar, R.; Gabriel, I.; Goldberg, B.; and Isaac, W. 2024. Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data. ArXiv:2406.13843 [cs].

Maslej, N.; Fattorini, L.; Perrault, R.; Parli, V.; Reuel, A.; Brynjolfsson, E.; Etchemendy, J.; Ligett, K.; Lyons, T.; Manyika, J.; Niebles, J. C.; Shoham, Y.; Wald, R.; and Clark, J. 2024. Artificial Intelligence Index Report 2024. *ArXiv*, abs/2405.19522.

McDuff, D.; Korjakow, T.; Cambo, S.; Benjamin, J. J.; Lee, J.; Jernite, Y.; Ferrandis, C. M.; Gokaslan, A.; Tarkowski, A.; Lindley, J.; et al. 2024. On the standardization of behavioral use clauses and their adoption for responsible licensing of ai. *arXiv preprint arXiv:2402.05979*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Mellor, S. 2022. A.I. chatbot trained on 4chan by YouTuber is slammed by ethics experts — fortune.com. https://fortune.com/2022/06/10/ai-chatbot-trained-on-4chan-by-yannic-kilcher-draw-ethics-questions/. [Accessed 11-09-2024].

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901.

Olteanu, A.; Olteanu, A.; Castillo, C.; Castillo, C.; Diaz, F.; Diaz, F.; Kıcıman, E.; Kiciman, E.; and Kiciman, E. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Social Science Research Network*.

Open Source Initiative. 2024. The Open Source AI Definition. Accessed: 2025-05-24.

Osborrne, C.; Ding, J.; and Kirk, H. R. 2024. The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub. *Journal of Computational Social Science*, 1–39.

Ozoani, E.; Gerchick, M.; and Mitchell, M. 2022. Model Card Guidebook.

Pepe, F.; Nardone, V.; Mastropaolo, A.; Bavota, G.; Canfora, G.; and Di Penta, M. 2024. How do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension*, 370–381.

Pistilli, G.; Muñoz Ferrandis, C.; Jernite, Y.; and Mitchell, M. 2023. Stronger Together: on the Articulation of Ethical Charters, Legal Tools, and Technical Documentation in ML. In *Proceedings of the 2023 ACM Conference on Fairness,*

Accountability, and Transparency, FAccT '23. New York, NY, USA: Association for Computing Machinery.

Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the Everything in the Whole Wide World Benchmark. ArXiv:2111.15366 [cs].

Raji, I. D.; and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Comanescu, R.; Akbulut, C.; Stepleton, T.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; Isaac, W.; and Weidinger, L. 2024. Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 1200–1217.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 8–14.

Sainz, O.; Campos, J. A.; García-Ferrero, I.; Etxaniz, J.; de Lacalle, O. L.; and Agirre, E. 2023. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoham, A. A. M.; Bai, X.; Gu, S.; Arora, M.; Zhou, K.; Koh, P. W.; Saxena, R.; et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

Strubell, E.; Ganesh, A.; and McCallum, A. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13693–13696.

Sun, L.; Huang, Y.; Wang, H.; Wu, S.; Zhang, Q.; et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 3.

Taraghi, M.; Dorcelus, G.; Foundjem, A.; Tambon, F.; and Khomh, F. 2024. Deep Learning Model Reuse in the HuggingFace Community: Challenges, Benefit and Trends. ArXiv:2401.13177 [cs].

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, 3266–3280.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 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*, 353–355.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; Kenton, Z.; Brown, S.; Hawkins, W.; Stepleton, T.; Biles, C.; Birhane, A.; Haas, J.; Rimell, L.; Hendricks, L. A.; Isaac, W.; Legassick, S.; Irving, G.; and Gabriel, I. 2021. Ethical and social risks of harm from Language Models. ArXiv:2112.04359 [cs].

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.

Wright, N. L.; Nagle, F.; and Greenstein, S. 2024. Contributing to Growth? The Strategic Role of Open Source Software for Global Startups. *Harvard Business School Strategy Unit Working Paper*, (24-040): 24–040.

Yang, X.; Liang, W.; and Zou, J. 2023. Navigating Dataset Documentation in ML: A Large-Scale Analysis of Dataset Cards on Hugging Face. In *NeurIPS 2023 Workshop on Regulatable ML*.

Yin, L.; Chakraborti, M.; Yan, Y.; Schweik, C.; Frey, S.; and Filkov, V. 2022. Open Source Software Sustainability: Combining Institutional Analysis and Socio-Technical Networks. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4791–4800.

Zhang, W. E.; Sheng, Q. Z.; Alhazmi, A.; and Li, C. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3): 1–41.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20.