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

Mahasweta Chakraborti,[1] Bert Joseph Prestoza,[1] Nicholas Vincent,[2] Seth Frey,[1] Vladimir Filkov[1]

[1]University of California Davis, USA
[2]Simon Fraser University, CA

{mchakraborti, sethfrey, vfilkov}@ucdavis.edu, nvincent@sfu.ca, bertjosephprestoza@gmail.com

Introduction

Emerging technologies often face skepticism and scrutiny before earning public trust. As AI grew, researchers and ethicists came together early to establish comprehensive documentation guidelines. Model cards, (Mitchell *et al.* 2019; Crisan *et al.* 2022; Arnold *et al.* 2019) and data sheets (Gebru *et al.* 2018; Bender *et al.* 2018) are crucial sociotechnical governance tools that ensure transparency by defining the scope of AI consumer applications. Model cards solicit permissible use cases, out-of-scope applications, and other anticipated risks and practical challenges for any given AI system.

Open-source AI is a rising player with a considerable market presence and increasing corporate adoption. While prized for rapid innovation through crowd-sourced contributions (Crowston *et al.* 2012; Coleman 2013; Benkler 2006), OSS may still face specific unique challenges in actuating responsible development. Firstly, their governance and power structures are often more informal and decentralized than corporations, with stark differences in goals, governance, and culture between communities (Li *et al.* 2021; Chakraborti *et al.* 2024; Yin *et al.* 2022; Shah 2006). This may complicate agreement over requirements and standards. Further, accountability and standardization would necessitate collaborative monitoring for sufficient documentation, inadequate evaluations (Franzen 2024; Ethayarajh and Jurafsky 2020; Balloccu *et al.* 2024), vulnerabilities (Birhane *et al.* 2021; 2023; Lee *et al.* 2024), and other forms of downstream misuse (DoJ 2024; PAI 2024; Marchal *et al.* 2024; Mellor 2022). Finally, strict regulation in OSS can impede innovation (Law and Krier 2023). Open source is primarily exempted by major AI legislations, including the EU AI act (European Parliament and Council of the European Union 2024). Yet, the act strongly endorses model cards and data sheets for open source developers "to accelerate information sharing along the AI value chain, promoting trustworthy AI systems in the Union". Therefore, we must closely understand OSS aspirations to inform the design of guardrails that foster mindfulness and social responsibility while preserving developer freedom.

Evaluation is core to AI development and generally involves testing models on held-out data to gauge their accuracy. Benchmarking is an essential yardstick for innovation against the state-of-the-art, guiding improvement and informing investment and deployment. Competitive leaderboards are often spun off from established benchmarks, where participation and high performance afford greater visibility (Dehghani *et al.* 2021; Ethayarajh and Jurafsky 2020; Raji *et al.* 2021). Besides the market-centric significance, evaluations are also core to AI governance. Developers are increasingly expected to use evaluations not only to assess model capabilities but also to *recognize their limitations*. Compliance through risk documentation necessitates probing edge cases, measuring predictive biases across specific domains and vulnerable subpopulations, and identifying other corner cases and failure modes (Liang *et al.* 2023; Bommasani *et al.* 2022; 2023; 2022; Mehrabi *et al.* 2021). Depending upon the criticality of the application (e.g., medical diagnostics or defense), judicious choice of evaluation suites and metrics can provide useful indicators and confidence levels, whose implications are explained through a dedicated 'Risks and Limitations' sections in model cards in a manner comprehensible to experts and users alike.

Research Questions

We explore current evaluation practices and accountability through documentation among OSS projects on Hugging Face (HF), a PaaS exclusively for AI/ML development, currently hosting the most extensive collection of AI repositories (Gong *et al.* 2023; Ait *et al.* 2023; 2024). Their guided annotation template (Face 2024) encourages developers to evaluate their models, including testing for potential usage limitations, vulnerabilities, and biases. Therefore, we may expect model evaluations and explanations of constraints to appear together. This should especially hold for proficient developers contributing high-value models that are highly accurate and have ideally been tested for potential run-time risks.

Conversely, ostensibly high evaluation performance may easily lead developers to assume generalization and overlook risks (Raji *et al.* 2021). This is especially true since leaderboards have notably been criticized for opacity and not factoring model attributes crucial to design and use, such as compactness, fairness, inference speed, and energy footprint, etc. (Ethayarajh and Jurafsky 2020; Raji *et al.* 2021). Importantly, developers may overlook the fact that error rates on static benchmarks are not a comprehensive measure of generalizability or adversarial robustness and thus

may not 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). This is especially true as leaderboards are often dominated by highly over-parameterized, complex and over fitted submissions (Ethayarajh and Jurafsky 2020; Hardt 2017; Arora and Zhang 2021; Dehghani *et al.* 2021; Blum and Hardt 2015). In a bid for top ranks, attempts to game leaderboards through multiple submissions and other hacks are not uncommon (Hardt 2017). Moreover, as rankings are generally based on models' aggregate performance over a collection of tasks and datasets, submissions often conceal skewered performances within subgroups, including racial and gender biases (Bordia and Bowman 2019; Manzini *et al.* 2019; Rudinger *et al.* 2018; Blodgett *et al.* 2020). Popular benchmark datasets also have been found with errors and other vulnerabilities that can compromise evaluations or even obfuscate risks (Dehghani *et al.* 2021; Gema *et al.* 2024; Bowman and Dahl 2021; Sainz *et al.* 2023). Lastly, while reporting high accuracy bolsters popularity, disclosing model shortcomings brings developers apprehension of hurting the same.

With arguments supporting the potential and challenges of using evaluations for AI governance, we hereby state our research questions. First, we find whether model evaluation reports tend to be accompanied by detailed sections on risks and limitations. To further probe the delicate tensions, we examine models on a large, shared benchmark suite to test whether the odds of risk documentation are also related to leaderboard performance.

RQ1: Is risk documentation associated with model evaluation? $Risks\ and\ Biases\ Documented? \sim Project\ Covariates + Evaluation\ Reported?$

RQ2: Are better performing models more transparent? $Risks\ and\ Biases\ Documented? \sim Project\ Covariates + Model\ Accuracy$

We pursue RQ2 on Hugging Face's first edition of the Open LLM leaderboard, which ran from May 2023 to June 2024. It drew 7173 unique submissions that were ranked 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 *et al.* 2022). 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 enabling and strengthening the validity of our experiments.

Data Analysis

Hugging Face hosted 700,072 open repositories as of 06/15/2024. Since AI auditing and documentation particularly apply to service-ready models and AI applications (Arnold *et al.* 2019; Law and Krier 2023), we next identify deployable models (See. Appendix), narrowing down to 456,545 projects. Around 15.9% and 2.2% of all these models were found to contain filled evaluations and risks sections, respectively.

As controlling for model size in terms of perceived risk and stake (Heim and Koessler 2024; Bender *et al.* 2021) is critical for our primary analyses, we exclude models for which size information was unavailable. We obtained model sizes (parameter count) for 140,783 models through platform supported libraries. Finally, by the current definition of open-source AI (OSI 2024), we identify projects where training data was also released and linked to the model weights. Around 7,092 models were open source, while 788 also participated in the open LLM leaderboard. Details on data collection and explanation of variables included in the regression analyses can be found in Appendix.

We frame our RQs as binary prediction modeling to determine if risk documentation is significantly associated with 1. rates of evaluation found in model cards and 2. absolute mean performance on the Open LLM Leaderboard. We model the likelihood of risk assessment in model cards using binomial logit models at a significance level 0.01, where evaluation practices (RQ1) or performance (RQ2) are the main regressors of interest, adjusting for crucial developer and project-level covariates (See Appendix). Presence of CO_2 emissions in model cards is also included as a dichotomous control in RQ1 and RQ2. We use the API for headers and string matching to detect valid CO_2 emission entries under designated sections in the card text.

We find a strong association between evaluation practices and risk documentation (RQ1), with models reporting some form of evaluation being 149.2% more likely to contain model risks and limits. More training data size, documentation of CO_2 footprint, developer team size, commit activity, and popularity (model likes) positively correlated to odds of risk documentation. Audio applications and models associated with high contributors (more models) were, meanwhile, less likely to carry risk documentation. RQ2 found 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%. Bigger models (parameters), documentation of CO_2 footprint, high number of commits, and developer team size also predicted 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. Detailed results of our analysis are available in the Appendix.

Discussion

Through our large sample empirical analysis and audit of OSS model cards, we discover promising trends in how evaluations can be mobilized for sociotechnical governance. At the same time, we find evidence of overall low evaluation and risk assessment rates, supporting long-standing observations and calls for fundamental reforms in evaluation objectives. While HF model cards encourage assessment of social impact, there is a gap in standards, expectations, and norms. Precise guidelines and training modules from platforms, such as tests and specific risks, and promoting well-

documented models (Liang *et al.* 2024) could greatly improve overall developer accountability while fostering innovation.

Leaderboards in particular need to consider the emerging needs of evaluation, improve upon their shortcomings, and incorporate multi-faceted tasks and metrics (Liang *et al.* 2023; Bommasani *et al.* 2023; 2022; Mehrabi *et al.* 2021; Zhao *et al.* 2018; Nadeem *et al.* 2021; Raji and Buolamwini 2019). Opportunistic overfitting can be mitigated by dynamic benchmarks (Nie *et al.* 2020; Dehghani *et al.* 2021; Kiela *et al.* 2021; Gehrmann *et al.* 2021), that are constantly updated to accommodate data drifts and emerging domains, tasks, and capabilities. Other specific measures may include confidentiality of test/hold out sets and mitigation of data leakage (Lilja *et al.* 2024; Deng *et al.* 2024) and contamination (Dwork *et al.* 2015; Balloccu *et al.* 2024; Magar and Schwartz 2022).

Limitations

As consensus on regulation and specific requirements for OSS continues to evolve, we do not evaluate whether the risk assessments provided in model cards are sufficient for these models. Future research on evaluation protocols and safety standards will benefit from continued exploratory studies of practitioners, particularly how specific tests can mitigate risks.

Hugging Face's popularity and moderation makes their leaderboards amenable for 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 to test language capabilities of AI models. Yet the patterns of developer behavior they reflect can provide crucial governance insight for rapidly growing technologies across platforms and modality. We look forward to future studies on improved, up and coming leaderboards for further validation of our findings and to inform evaluation practices going forward.

References

Adem Ait, Javier Luis Cánovas Izquierdo, and Jordi Cabot. On the Suitability of Hugging Face Hub for Empirical Studies, July 2023. arXiv:2307.14841 [cs].

Adem Ait, Javier Luis Cánovas Izquierdo, and Jordi Cabot. HFCommunity: An extraction process and relational database to analyze Hugging Face Hub data. *Science of Computer Programming*, 234:103079, May 2024.

Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

Sanjeev Arora and Yi Zhang. Rip van winkle's razor: A simple estimate of overfit to test data. *arXiv preprint arXiv:2102.13189*, 2021.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 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)*, pages 67–93, 2024.

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

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada, March 2021. ACM.

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

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Abeba Birhane, vinay prabhu, Sanghyun Han, Vishnu Boddeti, and Sasha Luccioni. Into the laion's den: Investigating hate in multimodal datasets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 21268–21284. Curran Associates, Inc., 2023.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

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

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *ACM Computing Surveys*, 55(5):1–166, 2022.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The Foundation Model Transparency Index, October 2023. arXiv:2310.12941 [cs].

Shikha Bordia and Samuel Bowman. 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*, pages 7–15, 2019.

Samuel Bowman and George Dahl. 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*, pages 4843–4855, 2021.

George EP Box and Paul W Tidwell. Transformation of the independent variables. *Technometrics*, 4(4):531–550, 1962.

Mahasweta Chakraborti, Curtis Atkisson, Ştefan Stănciulescu, Vladimir Filkov, and Seth Frey. 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*, pages 1–26, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gabriella Coleman. *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton University Press, Princeton, NJ, 2013.

Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439, Seoul Republic of Korea, June 2022. ACM.

Kevin Crowston, Kangning Wei, James Howison, and Andrea Wiggins. Free/libre open source software development: What we know and what we do not know. *ACM Computing Surveys (CSUR)*, 44(2):1–35, 2012.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The Benchmark Lottery, July 2021. arXiv:2107.07002 [cs].

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 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*, 2024.

DoJ. Man Arrested for Producing, Distributing, and Possessing AI-Generated Images of Minors Engaged in Sexually Explicit Conduct. U.S. Department of Justice, 2024. Accessed: August 27, 2024.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online, November 2020. Association for Computational Linguistics.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending 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), 6 2024. Text with EEA relevance.

Hugging Face. Model Card Guidebook huggingface.co. https://huggingface.co/docs/hub/en/model-card-guidebook, 2024.

Carl Franzen. New open source AI leader Reflection 70B's performance questioned, accused of 'fraud', 2024. Accessed: September 12, 2024.

Timnit Gebru, Timnit Gebru, Jamie Morgenstern, Jamie Morgenstern, Briana Vecchione, Briana Vecchione, Briana Vecchione, J. Vaughan, Jennifer Wortman Vaughan, Hanna Wallach, Hanna Wallach, Hal Daumé, Hal Daumé, Kate Crawford, and Kate Crawford. Datasheets for Datasets. *arXiv: Databases*, 2018.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 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)*, pages 96–120, 2021.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. Are We Done with MMLU?, June 2024. arXiv:2406.04127 [cs].

Lina Gong, Jingxuan Zhang, Mingqiang Wei, Haoxiang Zhang, and Zhiqiu Huang. 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), May 2023. Place: New York, NY, USA Publisher: Association for Computing Machinery.

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

Lennart Heim and Leonie Koessler. Training Compute Thresholds: Features and Functions in AI Regulation, August 2024. arXiv:2405.10799 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Huggingface. 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].

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

Huggingface. Safetensors — huggingface.co. `https:`

//huggingface.co/docs/safetensors/en/index. [Accessed 09-09-2024].

Huggingface. 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].

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 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*, pages 4110–4124, 2021.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Harry Law and Sébastien Krier. Open-source provisions for large models in the AI act. *Cambridge University Science and Policy Exchange*, 2023.

Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 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, 2024. Association for Computing Machinery. event-place: Honolulu, HI, USA.

Renee Li, Pavitthra Pandurangan, Hana Frluckaj, and Laura Dabbish. Code of conduct conversations in open source software projects on github. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31, 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. Systematic analysis of 32,111 ai model cards characterizes documentation practice in ai. *Nature Machine Intelligence*, 6(7):744–753, 2024.

Adam Lilja, Junsheng Fu, Erik Stenborg, and Lars Hammarstrand. 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*, pages 22150–22159, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, 2022.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 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)*, 2019.

Nahema Marchal, Rachel Xu, Rasmi Elasmar, Iason Gabriel, Beth Goldberg, and William Isaac. Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data, June 2024. arXiv:2406.13843 [cs].

Ninareh Mehrabi, MehrabiNinareh, Ninareh Mehrabi, Fred Morstatter, Fred Morstatter, MorstatterFred, Nripsuta Saxena, Nripsuta Saxena, SaxenaNripsuta, Nripsuta Saxena, Kristina Lerman, LermanKristina, Kristina Lerman, Aram Galstyan, GalstyanAram, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 2021.

Sophie Mellor. A.i. chatbot trained on 4chan by YouTuber is slammed by ethics experts, 2022. Accessed: September 11, 2024.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Moin Nadeem, Anna Bethke, and Siva Reddy. 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)*, pages 5356–5371, 2021.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, 2020.

OSI. The open source AI definition, 2024. Accessed: November 25, 2024.

PAI. Artificial intelligence incident database - discover, 2024. Accessed: September 11, 2024.

Inioluwa Deborah Raji and Joy Buolamwini. 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*, pages 429–435, 2019.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the Everything in the Whole Wide World Benchmark, November 2021. arXiv:2111.15366 [cs].

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference res-

olution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, 2018.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 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*, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Sonali K. Shah. Motivation, governance, and the viability of hybrid forms in open source software development. *Management science*, 52(7):1000–1014, 2006. ISBN: 0025-1909 Publisher: INFORMS.

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

Likang Yin, Mahasweta Chakraborti, Yibo Yan, Charles Schweik, Seth Frey, and Vladimir Filkov. Open source software sustainability: Combining institutional analysis and socio-technical networks. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), nov 2022.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, 2019.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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)*, pages 15–20, 2018.

Service-ready Features and Identifiers

Since AI auditing and regulation through documentation are particularly applicable for service-ready models and AI applications (Arnold *et al.* 2019; Law and Krier 2023), we screen out WIP projects and dumps for deploy-ready projects. Based on a comprehensive review of platform documentation and widgets/third-party features supporting direct or downstream applications, we identify service-ready models through at least one of the following properties:

- Model cards filled with detailed instructions, examples and use cases: Detected using HF's Model card scanner
- **Use this Model**: Platform-generated example scripts to guide model loading and use through recognized libraries.
- **Endpoints_compatible**: this tag indicates the model is compatible with Inference Endpoints and hence scalable and production-ready
- **Pipeline_tag**: denote the specific task a model was designed for, such as "text-classification", or "object-detection", and are 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, sometimes also a widget, to handle text generation queries to a model.

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

Data Collection

Project use, Developer activity, and Community engagement: Controlling for time lets us account for documentation practices as a function of evolving development standards, conception of ethical practices, and regulatory oversight. We measure the age of the repository as the time between the initiation of the project (first commit) and data collection. 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 may 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, 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. 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 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 non-profit. 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. Large Foundation

('Frontier') models have received particular attention from ethicists and policy oversight bodies (Heim and Koessler 2024; Bender *et al.* 2021). Recent proposals, particularly SB 1047 in California, explore graded requirements by model value. 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 both model size (number of parameters) and training data volume (number of samples). Non-uniform file nomenclature and frameworks complicate the automated loading and parsing of model details such as size (Huggingface a). Safetensors (Huggingface c) and GGUF (Huggingface b) 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 d). We obtained the parameter count for 140,783 models. 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 with all their training data available on HF and screening out models directed to invalid dataset repositories, we obtained training data sizes for 17,260 models. Overall, 7093 models had complete model and training data size information. One sample was excluded prior to analysis for being an outlier (cook's D > 1). Knowledge Domain, in the context of ML, 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. E.g., 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.

Compliance Information: Based upon development objectives and target use, developers chose appropriate tests and metrics to quantify model performance. Our first hypothesis testing requires predicting models' risk and limitations documentation against the rate at which they evaluate model performance. 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 distinct sections for technical information such as data provenance, development specifications, performance, legal/copyright aspects, and social implications of the model's use. 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 ensure domain accuracy and demographic fairness, i.e., performance specifically tested across relevant user groups and foreseeable error contexts specific to the model's use cases. Based on such evaluation, a development team's sociotechnical expert is expected to complete sections titled "Bias, Risks, and Limitations".

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 analyze all model cards to detect whether evaluations or risk assessments have been included. 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") or are empty for most sections. We treat these as false positive, non-compliant model cards for analytical integrity.

CO_2 footprint assessments are another crucial sociotechnical component of AI development that follows closely on the heels of the social impact (Lacoste *et al.* 2019; Strubell *et al.* 2020). Presence of these reports are controlled for confoundedness with social impact disclosures. We parse model card metadata along with string matching to detect valid CO_2 emission entries under designated sections in the model card text.

Competitive Benchmarking: Over time, different leaderboards have been created to test different AI applications. Leaderboard positions are highly competitive, and the highest performers enjoy considerable visibility and popularity. The Open LLM Leaderboard 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 *et al.* 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 *et al.* 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 accuracy. While this ostensibly reflects innovation, resubmissions are often overfitted to perform (Ethayarajh and Jurafsky 2020; Hardt 2017) and may indicate the competitiveness of the participant rather than sustainable development. For our analysis, we only consider the best performance for each model, controlling for the number of attempts and the precision of the model version tested. We also control for evaluation malpractice through flagged models.

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

Project Aspect	Variables	Description	Type	Source
Model Features	Model Size	Number of Model Parameters	Numeric	Model Page
	Training Resources	Number of 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 for e.g. code analysis, medical applications	Categorical	Training Data metadata (HF API)
Model Developer	Team Size	No. of Developers	Numeric	Linked Developer Profile
	Total Models	Development experience of contributor	Numeric	Linked Developer Profile
	Entity Type	If contributor is a for 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
Developer 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.

Regression Results

	Predictor	Coefficient	p-value
	(Intercept)	-3.263988	$<$**0.0001**
Model 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
	Reinforcement Learning	-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
Model Developer	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
Repository 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
Transparency	CO_2 footprint	2.177332	$<$ **0.0001**
	Evaluation Availability	0.913310	$<$ **0.0001**
Others	High Risk Application	-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: RQ1: 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. Developers/moderators often assign 'not-for-all-audience and 'NSFW' tags to certain projects inappropriate for general use, such as ones trained on or meant for sexual content generation. These high risk applications were incorporated as a categorical control in our analysis. To satisfy modeling assumptions, suitable higher-order transformations were applied to certain predictors using the Box-Tidwell approach (Box and Tidwell 1962)

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

	Predictor	Coefficient	p-value
	(Intercept)	-2.8854	$<$ **0.0001**
Model 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
Model Developer	Team members [1]	1.0562	$<$ **0.0001**
	Profile 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
Repository 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
Transparency	CO_2 Footprint availability	2.3698	**0.001487**
Evaluation Details	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 Application	-15.3855	0.994607
		N= 788 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: RQ2: 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. Developers/moderators often assign 'not-for-all-audience and 'NSFW' tags to certain projects inappropriate for general use, such as ones trained on or meant for sexual content generation[4]. These high risk applications were incorporated as a categorical control in our analysis. Flagged models were ones that were found to be contaminated or improperly evaluated. To satisfy modeling assumptions, suitable higher-order transformations were applied to certain predictors using the Box-Tidwell approach (Box and Tidwell 1962)