Mitigating Downstream Model Risks via Model Provenance

Keyu Wang
Mila, Quebec AI Institute
McGill University

Abdullah Norozi Iranzad

Scott Schaffter
Google

Meg Risdal
Google

Doina Precup
Mila, Quebec AI Institute
McGill University

Jonathan Lebensold
Mila, Quebec AI Institute
McGill University

Abstract

Research and industry are rapidly advancing the innovation and adoption of foundation model-based systems, yet the tools for managing these models have not kept pace. Understanding the provenance and lineage of models is critical for researchers, industry, regulators, and public trust. While model cards [25] and system cards [14] were designed to provide transparency, they fall short in key areas: tracing model genealogy, enabling machine readability, offering reliable centralized management systems, and fostering consistent creation incentives. This challenge mirrors issues in software supply chain security, but AI/ML remains at an earlier stage of maturity. Addressing these gaps requires industry-standard tooling that can be adopted by foundation model publishers, open-source model innovators, and major distribution platforms. We propose a machine-readable model specification format to simplify the creation of model records, thereby reducing error-prone human effort, notably when a new model inherits most of its design from a foundation model. Our solution explicitly traces relationships between upstream and downstream models, enhancing transparency and traceability across the model lifecycle. To facilitate the adoption, we introduce the *unified model record* (UMR) repository, a semantically versioned system that automates the publication of model records to multiple formats (PDF, HTML, LaTeX) and provides a hosted web interface. This proof of concept aims to set a new standard for managing foundation models, bridging the gap between innovation and responsible model management.

1 Introduction

Despite the incredible performance of frontier and foundation models [3] and their increasing use in production applications [18], the lack of transparency regarding the datasets and upstream models used during training remains a critical issue [4]. Without clear documentation, downstream developers risk inheriting harmful biases or flaws from upstream models, sometimes surfacing months or even years after a model's release [17]. However, our goal is not to compel model publishers—especially those offering closed-source, API-based models like GPT-4 [27]—to disclose proprietary training details. Instead, we propose an open-source contribution addressing the following question:

How do we warn downstream model providers of upstream risks?

Socially Responsible Language Modelling Research (SoLaR) workshop at the 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

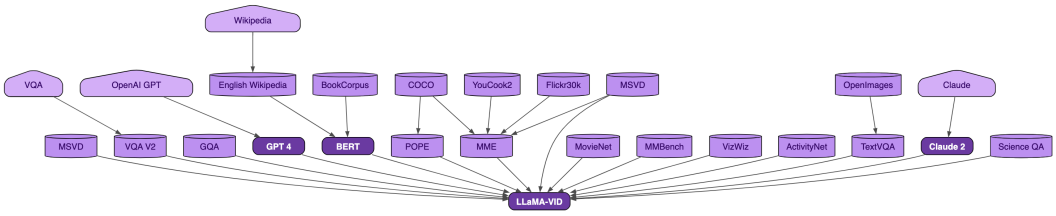

Figure 1: The provenance graph for the Llama-VID Short Video, a video-to-text captioning model built using a number of open-source foundation models.

Increasingly, foundation models rely on Internet-scale datasets in addition to pre-trained models to bootstrap their development [9, 5]. A number of common techniques for improving model performance, particularly in specialized tasks, involve composing multiple models or doing some form of model fine-tuning. For example, many vision language models (VLMs) are the result of a number of existing auto-encoders, classifiers and embedding models which have their own complex provenance graph (see fig. 1 for example) [26].

Unfortunately, many technical reports (such as model cards and system cards [25, 14]) do not report upstream dependencies in a format which can be easily interrogated. The consequence is that the community of researchers, system developers and users can inherit risks from the upstream models long after a foundation model is published. To address these risks, our contributions are as follows.

- We illustrate model provenance risk in the healthcare domain.
- We identify the properties needed to create early model warning systems.
- Finally, we propose an open-source, community-led system for tracking model provenance.

1.1 Data Poisoning and Regulatory Risks

Foundation models are defined by key artifacts: code, training and retrieval datasets, and final model parameters. Mostly for foundation models, only the model parameters are made available, while the code and datasets often remain undisclosed. For instance, ChatGPT [27] only provides API access, while models like Gemma [12] release parameters without sharing training details.

Deep learning models are prone to memorizing training data, leading to privacy concerns [6], even when accessed through public interfaces (i.e. in a black-box setting) [33]. This highlights the importance of managing training artifacts and ensuring transparency in the evaluation and post-training methods like RLHF [28]. Ideally, models should include reproducible evaluation protocols to promote accountability and ethical use.

Surprisingly, even models used to embed multi-modal data–such as CLIP–can be attacked to reveal training details [24]. As models are trained on internet-scale datasets, it becomes increasingly difficult to filter out problematic content. Furthermore, there are now well-known datasets, such as Common Crawl [23], and LAION-5B [32] which have been used without any pre-processing to produce models that risk reproducing undesirable samples [32]. The solution that we propose complements a growing ecosystem of related efforts described in the following section.

2 Related Work

The ecosystem for model and dataset specification, model benchmarking and open efforts are already actively engaged in developing tools to help increase transparency and surface potentially undesirable behaviour during model deployment. The *unified model record* we propose complements a number of existing initiatives.

Model and dataset specifications. Model cards [25] are designed to summarize a model's intended use, performance characteristics, ethical considerations, and other relevant details in a structured format. However, they lack the capability to model upstream relationships in a machine-readable way. Datasheets focus on how a dataset was collected while remaining agnostic to their application in different model training pipelines [10]. More recently, Croissant [1] has closed this gap by providing

a machine-readable format for defining datasets for machine learning research but stopped short of capturing a metadata description of the models themselves. System cards [14] have attempted to capture a more holistic description of a machine learning model and its deployment environment, and more recently, the ecosystem graph [4] was developed to try and relate models to their upstream dependencies. These efforts are closest to our proposal. Many of these efforts have been partially adopted by some model publishers, however, they are often under-specified or vulgarized in practice [22]. For example, at the time of this writing, Hugging Face renders a free-text README file in markdown format, denoted as a model card. More recently, they have begun tracking downstream models in a limited manner called model trees.

Benchmarking platforms. Government agencies have also begun establishing regulatory bodies focused on AI safety[36]. To aid these initiatives, several benchmark datasets targeting issues such as bias [39], toxicity [35], psychological harm [19], privacy [34] and safety [20, 2] have been developed. For instance, the ML Commons group is working on an open model evaluation platform, inspired by the HELM [22] and HEIM [21] frameworks [39]. However, these benchmarking platforms require significant computational resources and currently assess only a limited number of models [22]. These efforts build upon earlier projects like the LMSYS Chatbot Arena [8], which pioneered holistic model evaluation.

Open source efforts. On one hand, good science requires reproducible experiments, and it is clear that open-source benchmarks, models and code improve reproducibility [30]. Reproducibility checklists and standardized disclosures further aid in evaluating scientific claims. On the other hand, while web platforms like Papers with Code, Kaggle, and Hugging Face provide essential infrastructure, they do not enforce standardization across scientific contributions.

3 A Case Study in Healthcare

Results. We identify at least four widely published healthcare ML models that rely on upstream assets which may have been compromised. The risks identified come from the Pathology Language-Image Pertaining (PLIP) [16] model, a vision-language model widely used in the medical imaging field. We found that R2T-MIL[37], Breast Cancer Tumor and Immune Phenotypes Predictor [11], VLM-CPL [42], and PathLDM [40] each use PLIP. These vulnerabilities stem from two primary issues. First, PLIP potentially included sensitive and unethical content. Second, PLIP was trained on Twitter (now X) data, which recently has altered their data usage policies [7, 31]. This may have caused any downstream models to suffer from the same legal risks as PLIP itself.

Both these cases highlight a temporal risk: researchers can publish work ethically only to find that their upstream dependencies are now considered unsuitable. Proper model provenance management would add transparency and enable early warning systems for legal, CSAM and other risks.

Approach. Once a poisoned model was identified, each downstream model required at least two hours to review all available information. The whole process took in excess of 60 hours. We scrutinized popular datasets used in medical imaging, focusing on those known to have ethical concerns or potential contamination.

Examining downstream dependencies for vulnerabilities. Having found PLIP, we then began a new search for research papers, training data documentation, and model/system cards (where available) which cited its use. This process involved carefully reading through numerous papers and technical reports, reported in a variety of formats, and with varying degrees of detail regarding model architectures and training procedures.

3.1 Threats and Implications

Sensitive images in large datasets. The use of LAION-5B [32] in PLIP's training poses significant risks. In December 2023, the Stanford Internet Observatory discovered that LAION-5B contained hundreds of known child sexual abuse material (CSAM) images [1]. Although PLIP developers used only a subset of 32,000 images, they failed to rigorously filter out CSAM, meaning it may have

[1] See Stanford Internet Observatory (SIO): identified hundreds of known images of CSAM in LAION-5B

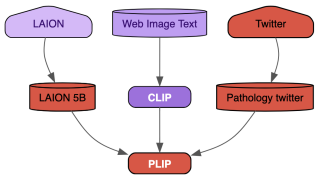

Figure 2: Model provenance graph of PLIP with its upstream dependencies

been part of the training data. Research indicates that models like PLIP, which do not generate new images, can still retain and potentially reconstruct training samples, propagating unethical and illegal content to downstream models [24]. Beyond CSAM, datasets like DukeMTMC [41], Labeled Faces in the Wild (LFW) [15], and MS-Celeb-1M [13] have encountered ethical issues, such as privacy violations and lack of informed consent [29]. Despite these problems and some data retractions, their downstream byproducts remain online. For instance, the LFW dataset, despite its flaws, saw over 3,000 downloads last month on platforms like Hugging Face.

Evolving data usage policies. The use of social media data, especially from platforms like X, complicates AI model development due to shifting policies. Over time, access to social media datasets has tightened – earlier datasets, such as the 2009-2010 Cheng-Caverlee-Lee Twitter Scrape [7], openly shared user geolocation data with full-text, while current datasets like GeoCoV19Tweets [31] limit access to tweet IDs, requiring users to retrieve full-text through the X API Tools. A key concern for researchers using PLIP is whether they must comply with updated licensing policies. Do these derivative models obtain the necessary licensing? Do they recognize the potential risks of using Twitter data and verify current privacy policies upon publication? These questions remain hotly contested by lawyers and regulators [38]. These findings stress the urgent need for better tracking of model and dataset provenance. Tools like our proposed *unified model records* can systematically track these model dependencies. Implementing robust provenance management will improve transparency, guide better model selection, and reduce risks tied to compromised or ethically questionable datasets.

4 Properties of Unified Model Records

Community support. To be successful *unified model records* requires a large amount of information about models to be curated, managed and shared in an accessible manner. Such problems are more effectively addressed using open-source communities given there is joint value derived from contributing to a common repository that helps the ecosystem in a non-extractive manner. UMRs are accordingly following this model, similar to what we find in other package management systems such as PyPI or NPM.

Web accessibility. Providing a singular source of truth to validate UMRs will also be crucial. While we can aggregate and index information about models in a distributed manner, a central index will need to be maintained to ensure data integrity across the ecosystem. This is similar to how NPM or other package managers need to be served by one or more providers, generally with one core provider, that can be augmented on a per-project or repository basis. UMR will follow a similar pattern and will provide a hosted location for model data.

Support for privately hosted unified model repositories. Much like closed-source software, private models and datasets may be produced which rely on public upstream assets. Many companies or initiatives will have needs that require managing internal-only model information. A solution exists in the open source community where *private packages*[2] can coexist with open source efforts. Similarly, UMR provides extensibility to enable the hosting of private model data that can be merged with the publicly managed set of UMR data. This will be critical in understanding and minimizing ecosystem risks.

Early warning and disclosures. Such an approach can leverage the model provenance graph to highlight impacted model owners when a given model is known to present a risk or a change in

[2]See Python: Installing private packages.

licensing or legal frameworks makes model usage no longer tenable in certain circumstances. This is similar to how a package management can be rolled back or audit a given project's dependencies for risks on an ongoing basis[3]. UMR will enable similar abilities to highlight and detect changing issues in a given model's dependency graph and enable automated early warning to address potential risks. This becomes particularly critical as model dependencies grow in size and complexity, and managing this manually ceases to be tenable.

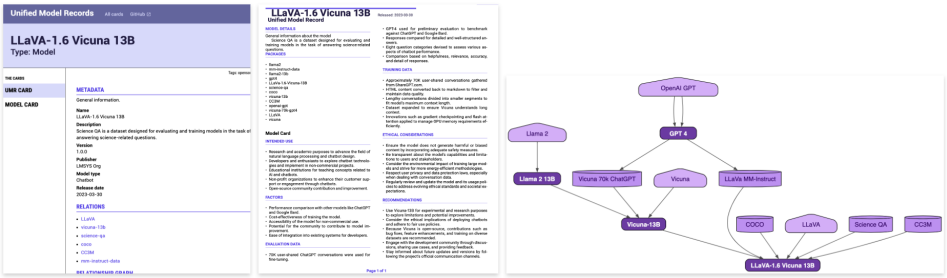

Figure 3: LLaVA-1.6 Vicuna 13B in different formats including HTML, PDF, and GraphViz

Hierarchical dependency management. To determine the provenance graph for downstream models, we claim the following properties are required. First, all the metadata must be in standardized, machine and human-readable format, such as JSON or YAML. Second, the work of developing a consistent representation needs to be undertaken by a community of contributors. Fortunately, a successful analog exists in open-source package management systems, like PyPI or NPM, where maintainers-who may or may not be the original authors-ensure metadata integrity and track dependencies. Third, metadata must be semantically versioned to adjust downstream dependencies over time. Fourth, the system must include both qualitative and quantitative evaluation results, enabling standardized reporting across models, systems, and datasets.

Standardized reporting. Standardized reporting for foundation models and their inputs will increase the overall legibility of each artifact. Much like nutrition labels, standardized reports can speed up analysis and evaluation. For example, conference reviewers will quickly be able to compare models developed as part of academic submissions, where it would be easier to compare whether the success of a new method comes from algorithmic innovation or merely the fine-tuning of larger and larger upstream models. These model records should be made available in LaTeX, HTML, PDF and GraphViz format so that they can be readily shared. Finally, industrial labs may be incentivized to improve the metadata quality of upstream open-source contributions so that they may inherit high-quality metadata in their own model records. Currently, over 50 unified model records are available on a public website[4] and Github repository [5].

5 Conclusion

The rapid advancement of foundation models necessitates robust tools for managing model provenance and mitigating downstream risks. The *unified model record* (UMR) repository will serve as an index, enabling efficient tracking of model lineage and dependencies. This system will automatically notify relevant stakeholders when upstream models are flagged for potential issues, facilitating rapid response to emerging risks.

We envision integration with platforms like Hugging Face and Kaggle, where UMRs can be automatically generated and updated alongside model uploads. Future work will focus on refining the UMR format, expanding the central repository's capabilities, and collaborating with platforms to support adoption. By providing a comprehensive solution for model provenance management, UMRs aims to foster responsible innovation and mitigate risks in the rapidly evolving landscape of AI ecosystems.

[3]See NPM: Auditing package dependencies for security vulnerabilities.

[4]See www.modelrecord.com.

[5]See Unified model records Github repository.

Acknowledgments and Disclosure of Funding

Special thanks to Adriana Romero-Soriano, Ahrav Dutta, Carole-Jean W, & Koustuv Sinha for their insightful comments and suggestions. This research was made possible by the scholarship from NSERC (Natural Sciences and Engineering Research Council of Canada).

References

[1] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, et al. Croissant: A metadata format for ml-ready datasets. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*, pages 1–6, 2024.

[2] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*, 2023.

[3] 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. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Rishi Bommasani, Dilara Soylu, Thomas I Liao, Kathleen A Creel, and Percy Liang. Ecosystem graphs: The social footprint of foundation models. *arXiv preprint arXiv:2303.15772*, 2023.

[5] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[7] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. Twitter dataset from cikm 2010. Internet Archive, 2010. `https://archive.org/details/twitter_cikm_2010`.

[8] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. `https://arxiv.org/abs/2403.04132`.

[9] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[11] Tiago Gonçalves, Dagoberto Pulido-Arias, Julian Willett, Katharina V. Hoebel, Mason Cleveland, Syed Rakin Ahmed, Elizabeth Gerstner, Jayashree Kalpathy-Cramer, Jaime S. Cardoso, Christopher P. Bridge, and Albert E. Kim. Deep learning-based prediction of breast cancer tumor and immune phenotypes from histopathology. *arXiv preprint arXiv:2404.16397*, 2024.

[12] Google. Gemma: Open foundation models. `https://blog.google/technology/developers/gemma-open-models/`, 2024. Accessed: [Insert access date].

[13] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *arXiv preprint arXiv:1607.08221*, 2016.

[14] Furkan Gursoy and Ioannis A Kakadiaris. System cards for ai-based decision-making for public policy. *arXiv preprint arXiv:2203.04754*, 2022.

[15] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Technical Report 07-49, University of Massachusetts, Amherst*, 2007.

[16] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023.

[17] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al. Position: On the societal impact of open foundation models. In *International Conference on Machine Learning*, pages 23082–23104. PMLR, 2024.

[18] Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. Real-world robot applications of foundation models: A review. *arXiv preprint arXiv:2402.05741*, 2024.

[19] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pages 1–10, 2024.

[20] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

[21] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] 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. *arXiv preprint arXiv:2211.09110*, 2022.

[23] Alexandra Sasha Luccioni and Joseph D Viviano. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*, 2021.

[24] Casey Meehan, Florian Bordes, Pascal Vincent, Kamalika Chaudhuri, and Chuan Guo. Do ssl models have déjà vu? a case of unintended memorization in self-supervised learning. *arXiv preprint arXiv:2304.13850*, April 2023. last revised 13 Dec 2023 (this version, v3).

[25] 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.

[26] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *Proceedings of Machine Learning Research*, 225:353–367, 2023.

[27] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke,

Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[28] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

[29] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *arXiv preprint arXiv:2108.02922*, 2021.

[30] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20, 2021.

[31] Usman Qazi, Muhammad Imran, and Ferda Ofli. Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information. *arXiv preprint arXiv:2005.11177*, 2020.

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[33] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[34] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*, 2023.

[35] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

[36] Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). 2023.

[37] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. *arXiv preprint arXiv:2402.17228*, 2024.

[38] The New York Times. New york times sues openai and microsoft. 2023. `https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html`. Accessed: 2024-09-02.

[39] Bertie Vidgen, Adarsh Agrawal, Ahmed M. Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Max Bartolo, Borhane Blili-Hamelin, Kurt Bollacker, Rishi Bomassani, Marisa Ferrara Boston, Siméon Campos, Kal Chakra, Canyu Chen, Cody Coleman, Zacharie Delpierre Coudert, Leon Derczynski, Debojyoti Dutta, Ian Eisenberg, James Ezick, Heather Frase, Brian Fuller, Ram Gandikota, Agasthya Gangavarapu, Ananya Gangavarapu, James Gealy, Rajat Ghosh, James Goel, Usman Gohar, Sujata Goswami, Scott A. Hale, Wiebke Hutiri, Joseph Marvin Imperial, Surgan Jandial, Nick Judd, Felix Juefei-Xu, Foutse Khomh, Bhavya Kailkhura, Hannah Rose Kirk, Kevin Klyman, Chris Knotz, Michael Kuchnik, Shachi H. Kumar, Srijan Kumar, Chris Lengerich, Bo Li, Zeyi Liao, Eileen Peters Long, Victor Lu, Sarah Luger, Yifan Mai, Priyanka Mary Mammen, Kelvin Manyeki, Sean McGregor, Virendra Mehta, Shafee Mohammed, Emanuel Moss, Lama Nachman, Dinesh Jinenhally Naganna, Amin Nikanjam, Besmira Nushi, Luis Oala, Iftach Orr, Alicia Parrish, Cigdem Patlak, William Pietri, Forough Poursabzi-Sangdeh, Eleonora Presani, Fabrizio Puletti, Paul Röttger, Saurav Sahay, Tim Santos, Nino Scherrer, Alice Schoenauer Sebag, Patrick Schramowski, Abolfazl Shahbazi, Vin Sharma, Xudong Shen, Vamsi Sistla, Leonard Tang, Davide Testuggine, Vithursan Thangarasa, Elizabeth Anne Watkins, Rebecca Weiss, Chris Welty, Tyler Wilbers, Adina Williams, Carole-Jean Wu, Poonam Yadav, Xianjun Yang, Yi Zeng, Wenhui Zhang, Fedor Zhdanov, Jiacheng Zhu, Percy Liang, Peter Mattson, and Joaquin Vanschoren. Introducing v0.5 of the ai safety benchmark from mlcommons, 2024. `https://arxiv.org/abs/2404.12241`.

[40] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. *arXiv preprint arXiv:2309.00748*, 2023.

[41] Zhimeng Zhang, Jianan Wu, Xuan Zhang, and Chi Zhang. Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project. *arXiv preprint arXiv:1712.09531*, 2017.

[42] Lanfeng Zhong, Xin Liao, Shaoting Zhang, Xiaofan Zhang, and Guotai Wang. Vlm-cpl: Consensus pseudo labels from vision-language models for human annotation-free pathological image classification. *arXiv preprint arXiv:2403.15836*, 2024.