nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00857-z

Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI

Received: 17 October 2023

Accepted: 21 May 2024

Published online: 21 June 2024

Check for updates

Weixin Liang[1,7], Nazneen Rajani[2,7], Xinyu Yang[3,7], Ezinwanne Ozoani[2], Eric Wu[4], Yiqun Chen[5], Daniel Scott Smith[6] & James Zou[1,4,5]

The rapid proliferation of AI models has underscored the importance of thorough documentation, which enables users to understand, trust and effectively use these models in various applications. Although developers are encouraged to produce model cards, it's not clear how much or what information these cards contain. In this study we conduct a comprehensive analysis of 32,111 AI model documentations on Hugging Face, a leading platform for distributing and deploying AI models. Our investigation sheds light on the prevailing model card documentation practices. Most AI models with a substantial number of downloads provide model cards, although with uneven informativeness. We find that sections addressing environmental impact, limitations and evaluation exhibit the lowest filled-out rates, whereas the training section is the one most consistently filled-out. We analyse the content of each section to characterize practitioners' priorities. Interestingly, there are considerable discussions of data, sometimes with equal or even greater emphasis than the model itself. Our study provides a systematic assessment of community norms and practices surroinding model documentation through large-scale data science and linguistic analysis.

The rising prevalence of AI models in various sectors has underscored the need for comprehensive model documentation[1–3]. As these models grow in complexity, their inner workings can seem increasingly obscure to those without specialized knowledge[4–6]. This creates a critical need for accurate and comprehensive documentation[7,8]. Effective model documentation serves as a vital communication bridge between developers and users, offering explicit guidance on the model's functionality, from its inputs and outputs to its range of applications[6,9]. Importantly, it reveals potential biases, errors and limitations inherent to the model[10]. This focus on transparency cultivates trust among users—a crucial component in fields in which model output has far-reaching consequences, such as in healthcare, finance and law enforcement[11]. Ultimately, model documentation serves as a pivotal tool for improving model utility and credibility across a range of stakeholders[12–16].

Model cards, inspired by the concept of food nutrition labels[17] and datasheets[18] in the electronics industry, have emerged as the standard approach to document AI models[7,19]. They are documents that provide essential information about a model in a standardized, easy-to-understand format. At the core of these cards are sections detailing model training and validation procedures, intended uses, potential limitations and usage guidance. Compared with other documentation formats such as academic papers or technical reports, model cards are increasingly becoming a preferred reference for practitioners in the AI community for a number of reasons. They offer more concise, relevant and easily understandable information about AI models, rendering models more accessible[20]. Another important aspect of model cards is their up-to-date nature, as they can be frequently updated to reflect any changes, improvements or new findings about the AI model.

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Hugging Face, Inc., Palo Alto, CA, USA. [3]Department of Information Science, Cornell University, Ithaca, NY, USA. [4]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [5]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [6]Graduate School of Education, Stanford University, Stanford, CA, USA. [7]These authors contributed equally: Weixin Liang, Nazneen Rajani, Xinyu Yang. e-mail: jamesz@stanford.edu

https://doi.org/10.1038/s42256-024-00857-z

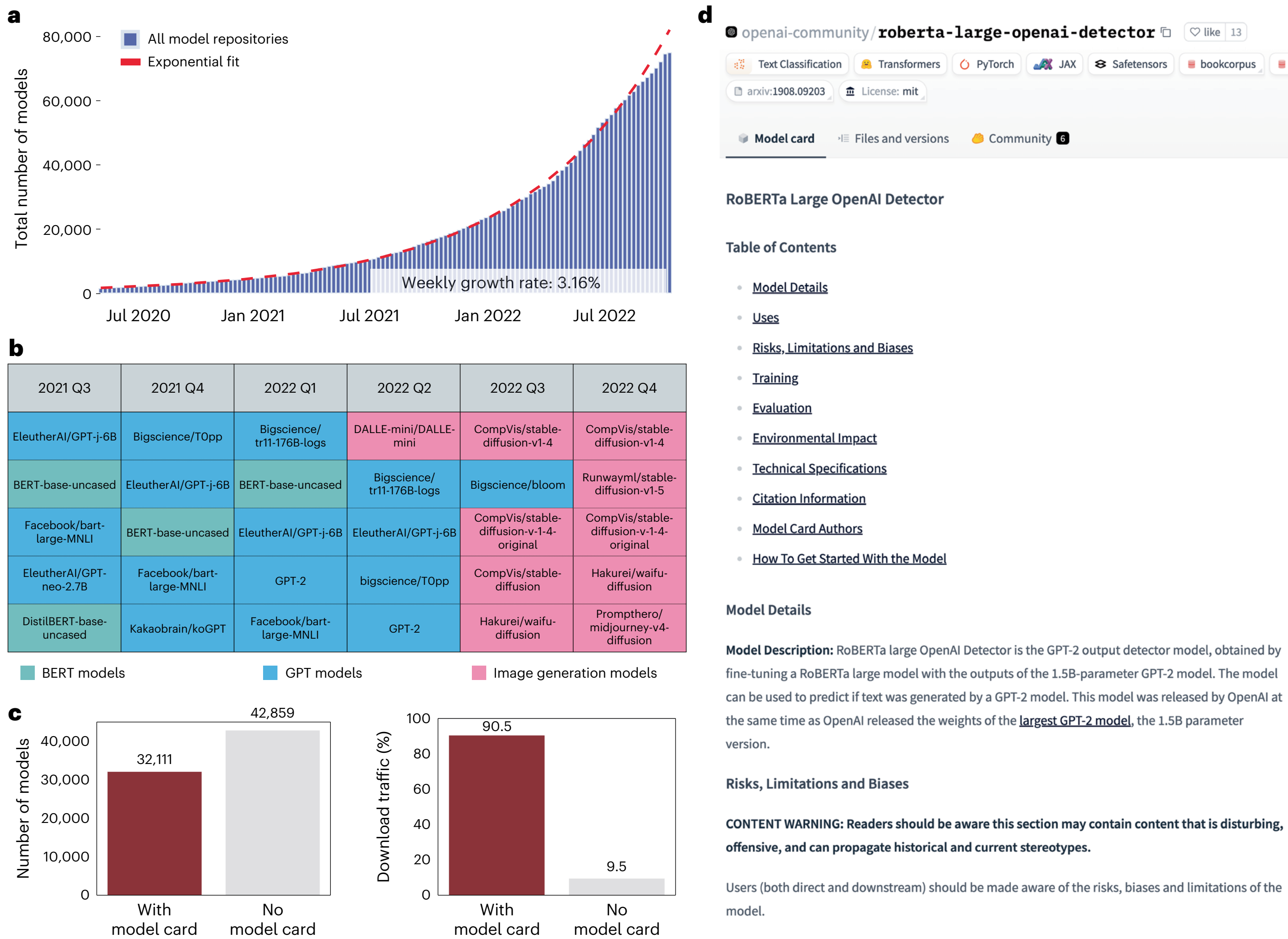


Fig. 1 | Overview of 74,970 AI models hosted on Hugging Face. a, Exponential Growth of Hugging Face model repositories. Our analysis indicates a substantial increase in the number of models hosted on Hugging Face, exhibiting a weekly growth rate of 3.16% and a doubling time of 22 weeks. As of 1 October 2022, there are 74,970 AI model repositories available on the platform. **b**, Temporal trends in model popularity. An examination of the top 5 trending AI model repositories on Hugging Face reveals dynamic shifts in community focus towards various model types. Language understanding and text generation models initially dominate, but image generative models gain prominence in later stages. **c**, Model card adoption and download traffic. Despite only 42.8% of model repositories (32,111 out of 74,970) featuring model cards, these models account for an overwhelming 90.5% of total download traffic on the platform. Our analysis focuses on these 32,111 models for further examination. **d**, Model card illustration. An exemplary model card is displayed, showcasing the essential components and information typically provided by a model card.

By contrast, academic papers, once published, may not be updated as regularly, which could result in outdated information. Furthermore, many popular model repositories, especially those originating from industry or open-source enthusiasts, do not have accompanying academic papers or technical reports[21,22]. This further accentuates the indispensable role of model cards as a comprehensive, streamlined and informative communication mechanism within the AI ecosystem.

Although many model cards are being created by researchers and developers[5–7,9,23–25], there has not been a systematic analysis of the quality and informativeness of the model cards. In particular, we do not know what has and has not been documented, how documentation practices vary across different types of models and organizations, and how documentation relates to factors such as model popularity. This is an important gap in our knowledge for several reasons. First, adherence to community norms and documentation standards is important for ensuring the ethical and responsible deployment of AI models. Comprehensive information on a model's functionality, limitations and potential biases is essential for users to make informed decisions, preventing misuse or unintended consequences. Second, understanding current documentation practices can help identify areas for improvement and guide the development of industry-wide standards. As AI becomes increasingly ubiquitous in society, there is a need for regulation to keep up with the pace of technical progress[26]. Towards this end, industry-wide AI documentation standards can help to foster innovation and mitigate potential harm[27], similar to how the electronics and automotive industries have converged toward universal technical specifications[28]. Finally, exploring the relationship between documentation and model characteristics can provide insights into the factors that influence documentation practices and help prioritize efforts to improve transparency and accountability in AI development. Without systematic analysis of current model cards, there is a risk of perpetuating inadequate documentation practices, which can obstruct efforts to ensure accountability and equitable use of AI technologies.

To address this gap in knowledge, we present a comprehensive large-scale analysis on 32,111 AI model cards created from 6,392 distinct user accounts. These model cards are uploaded by developers to Hugging Face, which is one of the most popular model repositories for hosting cutting-edge AI models across a multitude of research domains and applications[29,30]. Distinct from GitHub, which hosts general software repositories, Hugging Face offers a user-friendly, AI-specific platform that provides an integrated machine learning pipeline. Importantly, it has established community guidelines for model cards, fostering a standardization that GitHub's user-dependent README documents lack, making it an ideal choice for our study. Through this analysis we seek to characterize the extent to which the AI community has adopted and adapted model cards, the strengths and weaknesses of current documentation efforts, as well as evaluate their impact on model development and usage. Our study reveals that, although the adoption of model cards has been largely successful, considerable gaps persist in the completeness and quality of the provided documentation, with more than half of the models lacking a model card. The training section is most consistently completed, whereas sections addressing environmental impact, limitations and evaluation exhibit the lowest completion rates, indicating a need for greater emphasis on these aspects of model documentation. Furthermore, our section content analysis of four key model card sections—limitations, uses, evaluation and training—highlights a substantial focus on data, which often receivie as much, if not more, attention than the AI models themselves. Our study provides a unique outlook on community norms and practices around model documentation through large-scale data science and linguistic analysis.

Results

Data overview of AI models hosted on Hugging Face

Our analysis encompasses 74,970 AI model repositories on Hugging Face uploaded by 20,455 distinct user accounts as of 1 October 2022. The number of models exhibits exponential growth, with a weekly growth rate of 3.16% and a doubling time of 22 weeks (Fig. 1a). As a sanity check, the number of model repositories reached 145,306 by 4 March 2023, thereby confirming the exponential trend. The top five trending AI models on Hugging Face—defined by the number of likes received within a specific timeframe, according to the official Hugging Face criteria for trending models[31]—reveals shifts in community attention over time (Fig. 1b). Initially, GPT and BERT models dominated, reflecting a focus on language understanding and text generation; however, in Q3 2022, interest shifted towards image generative models (particularly stable diffusion and its variants), highlighting diverse interests beyond natural language processing. Despite widespread adoption of Hugging Face model cards, only 32,111 (42.8%, contributed by 6,392 distinct user accounts) out of the 74,970 model repositories currently include model cards as Markdown README.md files within their model repo (Fig. 1c); however, these models account for 90.5% of total download traffic, highlighting the prevalence of model cards among widely adopted and used models. Here the download traffic is measured by the number of model repositories downloads. In light of these findings, our subsequent analyses will concentrate on the 32,111 models equipped with model cards.

Completeness of model cards across various sections

Model cards provides a standardized structure for conveying key information about AI models. Grounded in academic literature[7] and official guidelines from Hugging Face[19], model cards conventionally comprise sections such as training, evaluation, uses, limitations, environmental impact, citation and how to start. As illustrated in Fig. 1d, these sections represent the essential constituents of a comprehensive model card. We parsed and evaluated the structure of model cards using a keyword-based detection pipeline for each section (for example, detecting mentions of CO_2 and its variants to identify the environmental impact section).

Our findings reveal considerable room for improvement in model card adherence to established community norms. We measured the adherence of model cards by evaluating whether the model card contained all of the six key sections recommended for model cards: training, evaluation, uses, limitations, citation and how to start, excluding the environmental impact section due to its recent introduction and low current filled-out rate of only 2%. Specifically, only 20% of the top 100 model cards and 10.2% of the top 500 model cards fully incorporate all of the recommended sections. Furthermore, our study identifies a notable correlation between a model card's adherence to community standards and the model's downloads, suggesting that higher-ranking models in terms of downloads tend to have model cards that are more closely aligned with these standards (Extended Data Fig. 1). We also analysed the variance in adherence to model card documentation norms between organizational and individual accounts on Hugging Face. Organizational accounts—which represent entities such as companies, universities or non-profits—generally exhibit better compliance with documentation standards compared with individual accounts, particularly in sections that require detailed information, such as the limitations section (Extended Data Fig. 2). A likely reason for this discrepancy could be that organizations possess more resources and may follow internal guidelines, leading to better documentation practices.

Aside from the general deficiency of model card adherence to established community norms, our evaluation indicates a considerable disparity in community attention to different model card sections, a trend that seems to be expanding over time. Across all of the model cards, environmental impact (2%, or 639 out of 32,111 model cards), citation (14.4%), evaluation (15.4%) and limitations (17.4%) sections exhibit the lowest filled-out rates, whereas the training section (74.3%) is most frequently filled out (Fig. 2a). Similar trends also hold for the top 100 and top 1,000 model cards. Sections with lower fill-out rates also tend to have a shorter average word count among the filled-out sections, indicating low community attention (Fig. 2c). For example, the environmental Impact section demonstrates both a low completion rate (3.7%) among the top 1,000 model cards and a low average word count (68 words) among the filled-out ones. By contrast, the training section exhibits the highest filled-out rate (71%) and the second highest average word count (168 words). Interestingly, despite its lower completion rate (22.1%), the limitations section tends to be among the longest (151 words on average), which hints at the complex nature of discussing model limitations. Results are consistent across the top 100 model cards and the all the model cards (Extended Data Fig. 3).

The disparities in community attention across different sections are progressively widening over time. A notable trend is the rapid increase in the filled-out rate for the training section ($P = 3.31 \times 10^{-244}$), even after accounting for model categories such as tabular and natural language processing; the fill rate of the environmental impact section is also increasing ($P = 4.64 \times 10^{-24}$). An interesting discovery is that a large majority (about 84.8%; 542 out of 639) of the environmental impact sections seem to be automatically created by AI model-building tools[32,33], which not only make the AI model, but also part of the model card. In particular, about 58.5% (374 out of 639) of these sections state 'model trained using AutoNLP,' and around 26.3% (168 out of 639) state 'model trained using AutoTrain'. Moreover, their section text aligns perfectly with the template provided by AutoNLP/AutoTrain. The adoption of these automated tools to track CO_2 emissions is a welcome change, as they increase awareness about the environmental impact of AI models. Through these tools, developers can better grasp the carbon footprint of their models, leading to more informed decisions during their creation and training process. Meanwhile, the filled-out rates for other sections are declining ($P < 0.001$; Extended Data Table 1).

The top model cards associated with the most downloaded models are distinctive from average model cards in a number of ways. One distinctive feature is that they are considerably longer (Fig. 2b and Extended Data Table 2). The top 100 model cards are, on average, 1.35 times longer than the top 1,000 model cards (521 versus 384 words) and 2.73 times longer than an average model card at the population level (191 words). The section filled-out rate also differs considerably. Take the top 100 model cards as an example. Although their training section completion rate is similar to the population level (81% versus 74.3%),

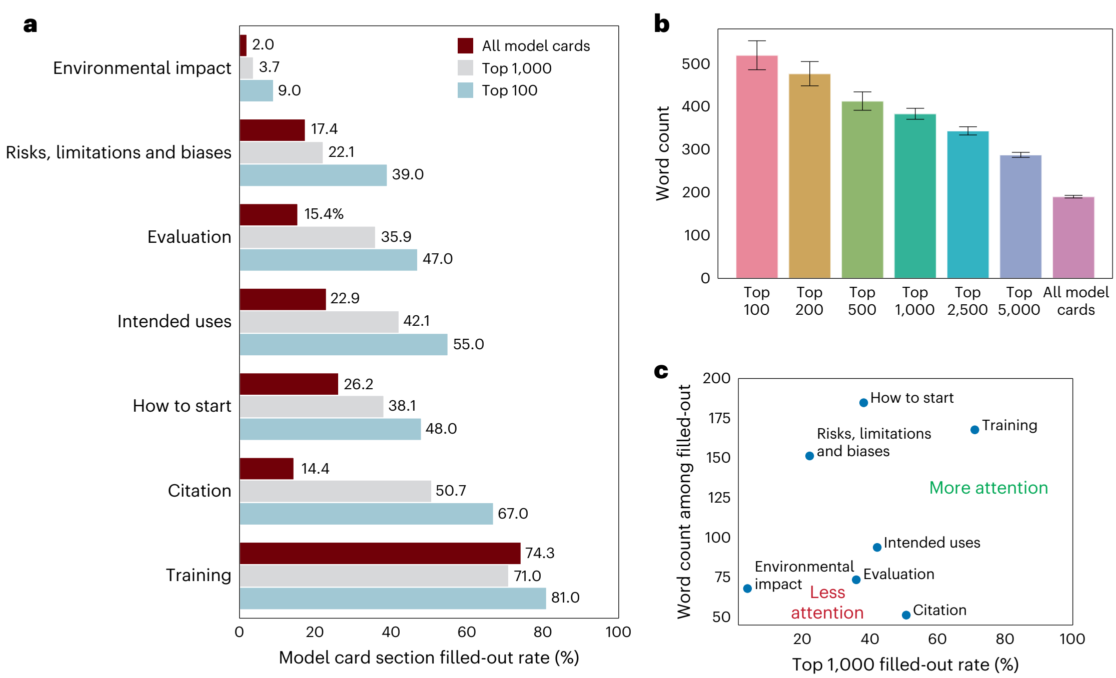


Fig. 2 | Section analysis of Hugging Face model cards. a, Low filled-out rates for model card sections. This panel presents the filled-out percentages for each section in 32,111 model cards, as well as the top 100 and top 1,000 model cards ranked by downloads. The environmental impact, limitations, and evaluation sections exhibit the lowest filled-out rates, whereas the training section is most frequently filled out. **b**, Highly downloaded models have longer model cards. Model cards in the top tier—ranked by downloads—are notably longer, suggesting a positive correlation between model card length and their usage. The total word count across all the sections is shown on the *y*-axis as mean values ± SEM. **c**, Disparate community attention patterns across model card sections: the environmental impact section demonstrates both a low filled-out rate among the top 1,000 model cards and a low average word count, indicating low community attention. Conversely, the training section exhibits a high filled-out rate and a high average word count, signifying more community attention.

they have much higher filled-out rates for the environmental impact (9% versus 2%), limitations (39% versus 17.4%) and evaluation (47% versus 15.4%) sections. The top 100 also have much higher filled-out rates for the citation section (67% versus 14.4%). These findings underscore the fact that top model cards are generally more detailed and structured, with a greater emphasis on sustainability and more thorough discussions of model performance and limitations. More specifically, we analysed the correlation between the comprehensiveness of model card sections and model downloads. We found that the limitations, how to start and evaluation sections are the top three factors associated with higher model downloads. This suggests that detailed attention to writing these sections is probably associated with a higher volume of model downloads. All sections but the training section are correlated with higher model downloads. For training, a reason could be that, among those who have model cards, most model cards have already done a good job at documenting the training sections (74.3%) and, empirically, in some cases, people write overly lengthy training sections by copying the entire training log, which might hurt readability and hence lead to a negative correlation (Extended Data Table 3). Furthermore, top models often have a strong connection to the academic research community, as shown by the high frequency of filled-out citation sections. Users may find these models more appealing because they come from scientific research, which is typically thoroughly checked for quality and accuracy by other experts in the same field.

A deep dive into model card content

To gain a comprehensive understanding of current practices and challenges in model documentation and identify areas for improvement, we conducted a content analysis of four critical model card sections on Hugging Face: limitations, uses, evaluation and training. This analysis was motivated by Hugging Face's internal user study[34], which identified the limitations and uses sections as the most challenging to write, and the fact that evaluation and training are two indispensable aspects of AI models. We employed a sentence-level topic modelling approach to accurately identify patterns and themes within the text (Fig. 3), enabling a fine-grained analysis of model card sections compared with document-level topic modelling (see Methods). We quantified the prevalence of specific themes by calculating the frequency of sections containing sentences that mention those themes, both across all model cards and within the top 100 model cards.

Limitations sections. In the limitations sections, our topic analysis uncovered a diverse array of subject matter, reflecting the myriad challenges and limitations faced by AI models. We identified three primary themes: disclaimers, data limitations and model limitations. Disclaimers emerged in 11.6% of the filled-out limitations sections, often emphasizing that the model is 'not intended for production' or 'should not be considered a clinical diagnostic tool,' particularly for medical AI models. Explicit disclaimer sentences regarding third-party usage were also noted. In the top 100 model cards, a large majority (about 89.2%) of the filled-out limitations sections included such disclaimers. This stands in contrast to the typical practice in academic research papers, in which disclaimers are less frequently seen[35,36]. This distinction is probably a reflection of the low barrier of access and deployment of AI models on Hugging Face, thereby requiring explicit cautionary notes to mitigate risks tied to potential misuse of AI models. Data and model limitations received nearly equal attention, appearing in 30.1% and 27.2% of the filled-out limitations sections, respectively. Their prevalence on the top 100 model cards is similar too. In data limitations, developers discussed the biases in training data, and the limited training data coverage. Discussions on model limitations revolved around both technical and societal aspects. From a technical standpoint, constraints, such as the maximum input length (for example, 1,024 tokens) for transformer-based models, were noted.

https://doi.org/10.1038/s42256-024-00857-z

a Limitations

Theme label	Keywords	Prevalence	Top 100 prevalence	Representative sentences
Data limitation	Dataset, model, data, bias	30.1%	51.4%	• The training data is unfiltered text from the internet and may contain all sorts of biases. • The model is limited by its training dataset and may not generalize well for all use cases. • As the training data comes primarily from biomedicine, performance on other domains may be poorer.
Model limitation	Bias, model, inherit, fair	27.2%	54.0%	• Keep this in mind when using this model, as information at the end of a text sequence longer than 1,024 tokens may be excluded from the final summary. • As the model is further pretrained on the BERT model, it may have the same biases embedded within the original BERT model. • We observed that this model could produce biased predictions that target underrepresented populations.
Disclaimer	Model, intend, purpose, response	11.6%	89.2%	• This model is intended for demonstration purposes only. • It should not be considered a clinical diagnostic tool. • Third parties who deploy or provide systems and/or services using any of these models (or using systems based on these models) should note that it is their responsibility to mitigate the risks arising from their use.

b Uses

Theme label	Keywords	Prevalence	Top 100 prevalence	Representative sentences
Intended task usage	Use, model, task, classification	58.2%	82.1%	• This model is meant to be used as a general object detector. • This finetuned model can be used to generate tweets which are related to Indian politics. • This model is intended to be used for zero-shot text classification of Italian texts.
How to use	Model, script, input, example	25.2%	64.3%	• To use the model in your Python script, just copy the following code: • When using the model make sure that your speech input is sampled at 16 khz. • Install the library via pip and activate model as below.
Out-of-scope use	Model, production, limit, intend	8.1%	44.6%	• The model should not be used to intentionally create or disseminate images that create hostile or alienating environments for people. • The model is not designed for critical decisions nor uses with any material consequences on an individual's livelihood or wellbeing. • Any production use of this model—whether commercial or not—is currently not intended.

c Evaluation

Theme label	Keywords	Prevalence	Top 100 prevalence	Representative sentences
Evaluation data	Dataset, data, corpus, split	37.8%	51.9%	• The fake-and-real news dataset contains a total of 44,898 annotated articles with 21,417 real and 23,481 fake. • The model can be evaluated as follows on the Swedish test data of common voice. • Canadian manifestos between 2004 and 2008 are used as test data.
Evaluation metrics and results	Result, accuracy, F1, test	26.9%	88.8%	• It achieves the following results on the evaluation set: loss: 0.3156 precision: 0.8332 recall: 0.8424 F1: 0.8378 accuracy: 0.9193. • The following table summarizes the F1 score obtained by ParsBERT as compared to other models and architectures. • The model achieves an overall accuracy on the test set equal to 82% the test set is a 20% split of the whole dataset.

d Training

Theme label	Keywords	Prevalence	Top 100 prevalence	Representative sentences
Training hyperparameters	Hyperparameters, optimizer, epoch, training	39.5%	72.2%	• The following hyperparameters were used during training. • The details of the masking procedure for each sentence are the following: 15% of the tokens are masked. • In total, the model went through approximately 10 epochs.
Training *data*	Dataset, data, training, token	32.7%	87.3%	• The ViT model was pretrained on a dataset consisting of 14 million images and 21,000 classes. • Convert all characters to lowercase. • Texts are preprocessed with the following rules. • The training corpus has been tokenized using a byte version of byte-pair encoding used in the original Roberta model with a vocabulary size of 50,262 tokens.
Training instruction	Script, GPU, training, Python	25.1%	51.9%	• The script used for training can be found here. • See the research paper for further details. • This model was trained on a single NVIDIA V100 GPU.

Fig. 3 | Uncovering key themes in model card sections through topic modelling analysis. a–**d**, This figure displays the outcomes of our topic modelling assessment on the contents of the limitations (**a**) uses (**b**), evaluation (**c**) and training (**d**) sections of model cards. Each panel illustrates the human-assigned theme label, prevalence, associated keywords, and representative sentences for each section. Prevalence reflects the frequency of sections containing sentences mentioning a specific theme, and top 100 prevalence is the prevalance measured on only the top 100 model cards. The keywords and representative sentences were algorithmically determined. Theme labels were manually assigned on the basis of the grouping of topic modelling sentence clusters. Our section content analysis identifies practitioners' priorities within each section, highlighting the community's focus on discussing data, sometimes with equal or even greater emphasis than the model itself.

On the societal front, concerns were raised about the biases in the AI model, as well as the potential risks that the AI model would inherit biases from its pretrained backbone models (for example, 'as the model is further pretrained on the BERT model, it may have the same biases embedded within the original BERT model.').

Uses sections. In the uses sections, our topic analysis uncovered three primary themes: designated model functionality, operational guidelines and misapplications. The most prominent theme in the uses sections was the designated model functionality, featured in 58.2% of the filled-out uses sections. Here, developers clearly laid out the specific tasks the model was built for, such as 'a fine-tuned model for generating tweets related to Indian politics.' Closely linked was the theme of operational guidelines, which was covered in 25.2% of the filled-out sections. This included practical information such as installation steps, usage examples, checkpoint details and fine-tuning instructions. Another important theme was misapplications, outlining the improper or out-of-scope use of the models. This theme, seen in 8.1% of the filled-out sections, covered topics such as malicious usage or the model's deployment in high-stakes situations. Although

Table 1 | Selected resources for promoting responsible and informative model card creation

Resource type	Name	Reference URL	Description
Model card generation tools	Hugging Face model cards writing tool	https://huggingface.co/spaces/huggingface/Model_Cards_Writing_Tool	Web application for creating model cards on Hugging Face from scratch
	US Census Bureau model card generator	https://bias.xd.gov/resources/model-card-generator/	Python-based tool for increasing transparency in government machine learning workflows
	VerifyML model card generation web tool	https://www.verifyml.com/	Web application for editing and comparing model cards online
	TensorFlow model card toolkit	https://pypi.org/project/model-card-toolkit/	Python library for automating model card generation.
	Parl.AI auto generation tool	https://parl.ai/docs/tutorial_model_cards.html	Open-source tool for documenting conversational AI research
	ChatGPT	https://chat.openai.com/	AI text generation for brainstorming limitations
	SalesForce model card creation toolkit	https://help.salesforce.com/s/articleView?id=release-notes.rn_bi_edd_model_card.htm&release=232&type=5	Tool for generating model cards on SalesForce platform
Hugging Face community resources	Hugging Face model card guidebook	https://huggingface.co/docs/hub/model-cards	Community guidelines on the structure of model cards
	The landscape of machine learning documentation tools	https://huggingface.co/docs/hub/model-card-landscape-analysis	Links to recent research on machine learning documentation
	Displaying carbon emissions for your model	https://huggingface.co/docs/hub/model-cards-co2	Guide on tracking and reporting CO_2 emissions
Education resources	About machine learning resources library	https://partnershiponai.org/about-ml-resources-library/	Multistakeholder initiative to operationalize machine learning transparency through documentation
	Code.org machine learning lesson 8: model cards	https://studio.code.org/s/aiml-2022/lessons/8/	Teaching materials on model cards
	Intro to AI ethics lession 5: model cards	https://www.kaggle.com/learn/intro-to-ai-ethics	Lesson on writing model cards
	The Hugging Face course	https://huggingface.co/learn/nlp-course/chapter4/4?fw=pt	Course on creating model cards on Hugging Face
Example model cards	Model card for Roberta large OpenAI detector	https://huggingface.co/roberta-large-openai-detector	Example model card for a fine-tuned Roberta model on Hugging Face
	Model card for pespective API	https://developers.perspectiveapi.com/s/about-the-api-model-cards	Example model card for a conversation moderation AI model
	Model cards with Google	https://modelcards.withgoogle.com/model-reports	Curated list of example Google model cards

some overlap exists between the themes in the limitations and uses sections, each emphasizes distinct aspects. For example, broad statements on commercial usage might appear under a disclaimer theme in limitations, whereas specific warnings about misuse feature in the misapplication theme in uses.

Evaluation sections. Our topic analysis of the evaluation sections in model cards highlighted two key themes: evaluation data and evaluation results. Featured in 37.8% of the filled-out sections, the first theme, evaluation data, described various datasets used for model testing. The second theme, evaluation results, appeared in 26.9% of the filled-out sections, presenting performance metrics such as F1 scores and bilingual evaluation understudy (BLEU) scores that depict a model's competence. Interestingly, these performance evaluations tend to present aggregate metrics across complete test datasets.

Training sections. In the training sections, our analysis surfaced three main themes: hyperparameter configurations, training data and training protocols. The theme of hyperparameter configurations, featured in 39.5% of the completed sections, provides important information such as the number of epochs, batch size and the selected optimizer. Equally important is the theme of training data, appearing in 32.7% of the filled-out sections. This theme provides a comprehensive description not only of the volume and characteristics of the training dataset, but also of specific data preprocessing steps such as 'convert all characters to lowercase'. Recent studies have highlighted the time-intensive nature of data work in AI development[37,38]. The third theme, training protocol, appearing in 25.1% of the filled-out sections, presents the technical steps required to replicate the training process. Taken together, these themes underscore a broader commitment within the AI community towards transparency and reproducibility by enabling researchers, practitioners, and industry stakeholders to efficiently build upon others' work[39,40].

Discussion

In this paper we present a comprehensive large-scale analysis on 32,111 AI model documentations on Hugging Face to understand the extent to which the AI community has adopted and adapted model cards; the strengths and weaknesses of current documentation efforts; and to assess their correlation with model downloads. Although our sample specifically focuses on Hugging Face, given its prominence among AI platforms, this not only mirrors AI community practices but also establishes transferable methods that could inspire further analyses on other platforms such as GitHub. Overall, our results indicate a widespread uptake of model cards within the AI community: 44.2% of the models have a corresponding model card, and these models account for 90.5% of total download traffic. Such a substantial adoption underscores the community's recognition of the importance of model cards in facilitating model comprehension and application.

However, we discovered an important dichotomy within the community's documentation practice. Although model cards have been adopted on a broad scale, there is a lack of compliance with established community standards and a striking disparity in the attention given to different sections of these cards. A majority of model cards incorporate a training section, signifying a focus on model development. Conversely, sections detailing environmental impact and evaluation—found in only 2% and 15.4% of model cards, respectively—are notably sparse. This gap extends to the limitations section, which is only present in 17.4% of the model cards. The growing divergence in the attention given to various sections, coupled with an increasing reluctance to address the limitations of models (both trends evidenced by our data; see Extended Data Table 1 and Extended Data Fig. 4) not only obstructs users from making informed decisions on model selection and usage, but also undermines trust in these AI models. Our findings echo a recognized trend in the broader scientific community: the tendency to downplay study limitations, emphasizing successes while often overshadowing potential weaknesses and negative outcomes[35,36]. Unlike traditional scientific papers which undergo peer review, model cards have no such mechanism for ensuring balanced and comprehensive documentation. This creates a unique challenge, indicating the need for novel solutions. Future work should focus on shaping strategies and standards to foster transparency and completeness in model card documentation. This is fundamental to building trust, advancing responsible AI use and equipping users—particularly those without expert knowledge—with the critical information required for making informed decisions regarding model selection and application[20].

Furthermore, our topic modelling analysis of four key model card sections—limitations, uses, evaluation and training—reveals an emerging consensus regarding the vital role of data within the AI pipeline. In the limitations sections, data limitations and model limitations received nearly equal attention. In both the training and evaluation sections, data emerge as a central theme. This emphasis on data in model documentation echoes existing literature that underscores the importance of data in AI model development. In practice, machine learning developers spend twice as much time on data as they do on the models themselves[37,38,41]. Poor-quality data can adversely affect an AI system's performance, robustness, safety and scalability[10,42]. Unfortunately, despite the criticality of data, it is widely acknowledged in the literature that data-related work is often undervalued in AI research. Data-related work is frequently perceived as operational[43] compared with the more celebrated task of building novel models and algorithms. As a result, a data-centric focus is often lacking in current AI research. Contrary to expectations, our diachronic analysis reveals that attention to data issues has not increased over time (Extended Data Table 4). Recent calls have been made by researchers to prioritize data-centric AI research[2]. Our results align with this trend and underscore the need for a more data-centric approach. By embracing and emphasizing the importance of data, the AI community can enhance the quality and reliability of models and foster advancements in responsible AI research.

In the broader context of open-source software documentation, past studies have demonstrated that the organization of repository homepages and README files can greatly influence a project's perception, sustainability and popularity, and even impact audience members' actions, such as joining the project[44–47]. To assess the causal impact of model cards on model utilization, we conducted a preliminary intervention study by adding detailed model cards to 42 popular models which have no or sparse model cards previously. We find that adding model cards is moderately correlated with an increase weekly download rates (Extended Data Fig. 5). Given our preliminary results, we believe that it is worth pursuing more extensive, larger-scale randomized studies of model cards in the future. This can shed light on the nuanced ways in which model cards influence not just model downloads, but also broader aspects of model usage and their downstream impacts. The insights from such future work will be instrumental in guiding best practices for model card design and implementation, to fully realize their potential in promoting transparency, usability and responsible AI practices.

Topics on how to create high-quality and responsible documentation should be incorporated into the AI curriculum[5,48–51]. Table 1 presents a compilation of selected resources, including tools for generating model cards with graphical user interfaces, educational materials and example model cards. These resources can serve as valuable teaching aids for students and practitioners alike. For example, green AI tools have been developed to automatically track model training CO_2 emissions and document the environmental impact[32,33]. Furthermore, the Hugging Face model cards writing tool simplifies the process of creating model cards by providing a user-friendly graphical interface, allowing individuals and teams with diverse skill sets and roles to collaborate effortlessly, even without knowledge of coding or Markdown. Documentation will be integral to the next stage of AI development, especially as we translate models from research sandbox to real-world deployment. By providing these resources, we aim to reduce barriers to entry and encourage the responsible deployment of AI models through transparent and accurate documentation.

Methods

Model cards as AI documentation framework

A plethora of documentation tools have been proposed to address different aspects of AI systems, with early efforts focused on data and later expanding to capture the AI development lifecycle more comprehensively. Data-focused documentation tools[23–25]—including datasheets for datasets[18] and data nutrition labels[17,52]—serve as a vehicle for communication between dataset creators and consumers, addressing the lack of industry standards for documenting AI datasets. These tools document the contexts and contents of datasets, including their motivation, composition, collection process and recommended uses, which facilitates greater transparency accountability and reproducibility of AI results while mitigating unwanted biases in AI models. They also enable dataset creators to be intentional throughout the dataset creation process. Datasheets and other forms of data documentation are increasingly released along with datasets, helping researchers and practitioners to choose the right dataset for their needs. These tools emphasize the data lifecycle[53], encompassing aspects such as assembly, collection and annotation. Later, efforts expanded to capture the AI development lifecycle more comprehensively, emphasizing the importance of an iterative design process to ensure accessibility to users with diverse backgrounds and goals when interacting with model cards—including developers, students, policymakers, ethicists, those impacted by AI models and other stakeholders. Model cards[7] and FactSheets[9] serve as reporting tools for trained AI models, documenting evaluation, usage, and other relevant aspects. Model cards foster transparency, accountability, comparability and user confidence, which are essential for responsible development and deployment of AI models[12–16,54]. Furthermore, model cards also have the potential to assist in regulatory compliance by offering a structured framework for documenting and communicating key information about a model's performance, training and evaluation process[6,9,55,56]. Value cards[5] educate students and practitioners about values related to AI models, while consumer labels for AI models[6] help non-experts understand key concerns. Researcher are also exploring the design of system cards for complex systems such as AI-based content ranking which often involves a pipeline of many AI models, and discussing complicated issues including use policies, access controls, and monitoring for abuse. In the broader context of open-source software documentation, previous studies have demonstrated that the organization of repository homepages and README files can greatly influence a project's perception and even impact audience members' actions, such as joining the project[44,45]. Well-structured READMEs have been linked to enhanced sustainability and popularity of projects[46,47]. These insights underscore the necessity for comprehensive and user-friendly documentation to promote user trust, engagement and adoption of AI models and tools.

Accessing and parsing model cards

We employed Hugging Face's library, 'huggingface_hub', to access the model cards on the Hugging Face hub. The 'huggingface_hub.list_models()' application programming interface grants access to all models hosted on the hub (74,970 models as of 1 October 2022). Furthermore, the library's ModelCard module allows for the loading of existing cards from the hub. By examining the CardData metadata of models, we can verify the presence of a model card, and employ 'huggingface_hub.ModelCards.load()' function to obtain the card associated with a given model.

Model cards are written in Markdown format and serve as the README file for the model repository. To extract plain descriptive text for further analysis, we used the Python package mistune (https://mistune.readthedocs.io/en/latest/) to parse the README file and remove features such as tables, codes, links and images. The structure of model cards plays a crucial role in conveying key information. Drawing from academic literature[7] and Hugging Face's official guidelines, model cards typically include sections such as training, evaluation, uses, limitations, environmental impact, citation and how to start. An example model card, shown in Fig. 1d, demonstrates the essential components and information provided by model cards.

However, due to the subjective and dynamic nature of section naming, sections that discuss the same topic may not have the same title. For example, both the training procedure and training results sections may address aspects of the training. To categorize sections with the same topic, we use a keyword-based detection method to identify similar topics. We define categories based on the Hugging Face model card template. We then applied keyword detection on both the section titles and contents to categorize the sections. It is important to note that section detection could not be solely based on the section titles. Although some section titles, such as 'Training data', may contain keywords that allow for direct categorization, other information may be hidden within the section contents. For instance, in the 'Model trained using AutoTrain' section, the topic being discussed cannot be inferred from the section title alone; however, the section contains information about the amount of CO_2 generated by AutoTrain, which can be categorized as 'Environmental impact'. By utilizing keyword detection, we achieved a coverage of categorized sections of 89.51%, which validates the reliability of our categorizing method.

Topic modelling

We leverage a sentence-level topic modelling approach to identify the patterns and themes within the section. For each section, we first employ NLP methods to split it into sentences and label the sentences with their original section, which constructs a sentence-section map. We then deploy BERTopic[57] (https://maartengr.github.io/BERTopic/) to automatically cluster the themes. BERTopic is a topic modelling technique that utilizes BERT[58] embeddings and class-based TF-IDF (c-TF-IDF) to develop compact and coherent clusters, thereby enabling the generation of easily understandable topics while preserving salient terms within the topic descriptions. It starts by converting the documents to numerical representations. The sentences are embedded by using the default embedding model, Sentence Transformer[59] all-MiniLM-L6-v2. Uniform manifold approximation and projection[60] (UMAP) was then used to reduce the dimensionality of document embedding into something easier for clustering. After reducing the embeddings, it starts clustering the data. For that, a density-based clustering technique, hierarchical density-based spatial clustering of applications with noise[61] (HDBSCAN), is used to cluster the data. It can find clusters of different shapes and has the nice feature of identifying outliers where possible. When using HDBSCAN to perform clustering, the bag-of-words representation is used on the cluster level for the purpose of topic representation, in which the frequency of each cluster can be found. That is, all documents in a cluster can be combined into a single document, and we can count how often each word appears in each cluster. From the generated bag-of-words representations, c-TF-IDF is used to extract the most important words per cluster.

$$\mathbf{W}_{x,c} = \left\|\mathbf{tf}_{x,c}\right\| \times \log\left(\mathbf{1} + \frac{\mathbf{A}}{\mathbf{f}_x}\right),$$

where $\mathbf{W}_{x,c}$ is the c-TF-IDF score of word x in class c; $\mathbf{tf}_{x,c}$ is the frequency of word x in class c; $\mathbf{f}_x$ is the frequency of word x accross all classes; and $\mathbf{A}$ is the average number of words per class. The words are then sorted on the basis of their c-TF-IDF score, and the key words are the top words. The representative sentences are selected randomly in each cluster.

By employing this approach, we have got the automatically clustered topics generated from BERTopic with their keywords and representative sentences; however, the clustered topics are too fine-grid, and we identified some topics that are acceptable to be merged. For example, in the training section, 'topic 2_used_dataset_containing_commoncrawl' and 'topic 4_bookcorpus_books_11038_tables' are both about the training data. Although we can fine-tune the parameters of BERTopic to get better results, the fine-tuning process could be time-consuming and might miss some important patterns. We therefore manually identify similar clustered topics based on the keywords and representative sentences, and use BERTopic to merge them and get the new keywords and representative sentences.

Finally, we use the sentence-section map constructed previously and map the sentences to sections, and calculate the size of sections within each topic. This allows us to identify the most prominent themes in each section and the size of each theme relative to the section.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The Hugging Face model cards data are public on Hugging Face at https://Hugging Face.co/models and can be accessed through the Hugging Face Hub API at https://Hugging Face.co/docs/Hugging Face_hub/package_reference/hf_api.

Code availability

The analysis code is publicly available at https://github.com/Weixin-Liang/AI-model-card-analysis-Hugging Face (ref. 62).

References

1. Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A. & Zou, J. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* **186**, 1772–1791 (2023).
2. Liang, W. et al. Advances, challenges and opportunities in creating data for trustworthy AI. *Nat. Mach. Intell.* **4**, 669 – 677 (2022).
3. Arrieta, A. B. et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58**, 82–115 (2020).
4. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (Lulu, 2020).

https://doi.org/10.1038/s42256-024-00857-z

5. Shen, H. et al. Value cards: an educational toolkit for teaching social impacts of machine learning through deliberation. In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 850–861 (ACM, 2021).
6. Seifert, C., Scherzinger, S. & Wiese, L. Towards generating consumer labels for machine learning models. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)* 173–179 (IEEE, 2019).
7. Mitchell, M. et al. Model cards for model reporting. In *Proc. Conference on Fairness, Accountability, and Transparency* 220–229 (ACM, 2019).
8. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (ACM, 2021).
9. Arnold, M. et al. Factsheets: increasing trust in ai services through supplier's declarations of conformity. *IBM J. Res. Dev.* **63**, 1–13 (2019).
10. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**, 1–35 (2021).
11. He, B. et al. Blinded, randomized trial of sonographer versus AI cardiac function assessment. *Nature* **616**, 520–524 (2023).
12. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should i trust you?' Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).
13. Raji, I. D. & Yang, J. About ML: annotation and benchmarking on understanding and transparency of machine learning lifecycles. Preprint at https://arxiv.org/abs/1912.06166 (2019).
14. Diakopoulos, N. Accountability in algorithmic decision making. *Commun. ACM* **59**, 56–62 (2016).
15. Selbst, A. & Powles, J. 'Meaningful information' and the right to explanation. In *Conference on Fairness, Accountability and Transparency* 48–48 (PMLR, 2018).
16. Kulesza, T., Burnett, M., Wong, W.-K. & Stumpf, S. Principles of explanatory debugging to personalize interactive machine learning. In *Proc. 20th International Conference on Intelligent User Interfaces* 126–137 (ACM, 2015).
17. Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. The dataset nutrition label: a framework to drive higher data quality standards. Preprint at https://arxiv.org/abs/1805.03677 (2018).
18. Gebru, T. et al. Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).
19. *Hugging Face Model Card Guidebook* (Hugging Face, accessed 7 May 2023).
20. Bracamonte, V., Pape, S., Löbner, S. & Tronnier, F. Effectiveness and information quality perception of an ai model card: a study among non-experts. In *Proc. 2023 20th Annual International Conference on Privacy, Security and Trust (PST)* 1–7 (IEEE, 2023).
21. Conover, M. et al. *Hello Dolly: Democratizing the Magic of ChatGPT with Open Models* (Hugging Face, 2023); https://HuggingFace.co/databricks/dolly-v1-6b
22. Chiang, W.-L. et al. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality* (Limsys, 2023).
23. Bender, E. M. & Friedman, B. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguistics* **6**, 587–604 (2018).
24. Bender, E. M., Friedman, B. & McMillan-Major, A. *A Guide for Writing Data Statements for Natural Language Processing* (Tech Policy Lab, 2021).
25. Pushkarna, M., Zaldivar, A. & Kjartansson, O. Data cards: purposeful and transparent dataset documentation for responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* 1776–1826 (ACM, 2022).
26. Brundage, M. et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. Preprint at https://arxiv.org/abs/2004.07213 (2020).
27. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399 (2019).
28. Mackiewicz, R. E. Overview of IEC 61850 and benefits. In *2006 IEEE Power Engineering Society General Meeting* 8 (IEEE, 2006).
29. Taori, R. et al. *Stanford Alpaca: An Instruction-Following llama Model* (Hugging Face, 2023); https://HuggingFace.co/models/tatsu-lab/alpaca
30. Köpf, A. et al. Openassistant conversations–democratizing large language model alignment. Preprint at https://arxiv.org/abs/2304.07327 (2023).
31. *Trending Model* (Hugging Face, accessed 12 April 2023); https://Hugging Face.co/?trending=model
32. *Displaying Carbon Emissions for Your Model* (Hugging Face, accessed 12 April 2023); https://HuggingFace.co/docs/hub/model-cards-co2
33. *AutoTrain* (Hugging Face, accessed 12 April 2023); https://Hugging Face.co/autotrain
34. *Model Card User Studies* (Hugging Face, accessed 12 April 2023); https://Hugging Face.co/docs/hub/model-cards-user-studies
35. Smith, J. J., Amershi, S., Barocas, S., Wallach, H. M. & Vaughan, J. W. Real ML: recognizing, exploring, and articulating limitations of machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2022).
36. Ioannidis, J. P. A. Limitations are not properly acknowledged in the scientific literature. *J. Clin. Epidemiol.* **60 4**, 324–329 (2007).
37. Sambasivan, N. et al. 'Everyone wants to do the model work, not the data work': data cascades in high-stakes AI. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* 1–15 (CHI, 2021).
38. *State of Data Science 2020* (Anaconda, accessed 22 May 2023); https://www.anaconda.com/state-of-data-science-2020
39. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (ACL, 2020).
40. McMillan-Major, A. et al. Reusable templates and guides for documenting datasets and models for natural language processing and generation: a case study of the Hugging Face and GEM data and model cards. In *Proc. 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)* (eds Bosselut, A. et al.) 121–135 (ACL, 2021).
41. *Kaggle ML and DS Survey* (Kaggle, accessed 27 August 2020); https://www.kaggle.com/c/kagglesurvey-2019
42. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).
43. Sculley, D. et al. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing System* Vol. 28 (NurIPS, 2015).
44. Qiu, H. S., Li, Y. L., Padala, S., Sarma, A. & Vasilescu, B. The signals that potential contributors look for when choosing open-source projects. In *Proc. ACM on Human–Computer Interaction* Vol. 3, 1–29 (ACM, 2019).
45. Vasilescu, B. et al. Gender and tenure diversity in github teams. In *Proc. 33rd Annual ACM Conference on Human Factors in Computing Systems* 3789–3798 (ACM, 2015).

46. Begel, A., Bosch, J. & Storey, M.-A. Social networking meets software development: perspectives from GitHub, MSDN, Stack Exchange, and Topcoder. *IEEE Softw.* **30**, 52–66 (2013).
47. Fan, Y., Xia, X., Lo, D., Hassan, A. E. & Li, S. What makes a popular academic AI repository? *Empirical Softw. Eng.* **26**, 1–35 (2021).
48. Fiesler, C., Garrett, N. & Beard, N. What do we teach when we teach tech ethics? A syllabi analysis. In *Proc. 51st ACM Technical Symposium on Computer Science Education* 289–295 (ACM, 2020).
49. Reich, R., Sahami, M., Weinstein, J. M. & Cohen, H. Teaching computer ethics: a deeply multidisciplinary approach. In *Proc. 51st ACM Technical Symposium on Computer Science Education* 296–302 (ACM, 2020).
50. Bates, J. et al. Integrating fate/critical data studies into data science curricula: where are we going and how do we get there? In *Proc. 2020 Conference on Fairness, Accountability, and Transparency* 425–435 (ACM, 2020).
51. Leidig, P. M. & Cassel, L. ACM taskforce efforts on computing competencies for undergraduate data science curricula. In *Proc. 2020 ACM Conference on Innovation and Technology in Computer Science Education* 519–520 (ACM, 2020).
52. Chmielinski, K. S. et al. The dataset nutrition label (2nd gen): leveraging context to mitigate harms in artificial intelligence. Preprint at https://arxiv.org/abs/2201.03954 (2022).
53. Hutchinson, B. et al. Towards accountability for machine learning datasets: practices from software engineering and infrastructure. In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 560–575 (ACM, 2021).
54. Zou, J. & Schiebinger, L. Design AI so that it's fair. *Nature* **559**, 324–326 (2018).
55. Regulation, P. Regulation (EU) 2016/679 of the European parliament and of the council. *Regulation* **679**, 2016 (2016).
56. Goodman, B. & Flaxman, S. European union regulations on algorithmic decision-making and a 'right to explanation'. *AI Magazine* **38**, 50–57 (2017).
57. Grootendorst, M. BERTtopic: neural topic modeling with a class-based TF-IDF procedure. Preprint at https://arxiv.org/abs/2203.05794 (2022).
58. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at https://arxiv.org/abs/1810.04805 (2018).
59. Reimers, N. & Gurevych, I. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing* 4512–4525 (ACL, 2020).
60. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
61. McInnes, L. & Healy, J. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* 33–42 (IEEE, 2017).
62. Liang, W. & Yang, X. *AI-Model-Card-Analysis-Hugging Face: Analysis of AI Model Cards on Hugging Face Hub* (Zenodo, 2023); https://doi.org/10.5281/zenodo.11179952

Acknowledgements

We thank D. McFarland and H. Fang for discussions. J.Z. is supported by the National Science Foundation (grant nos. CCF 1763191 and CAREER 1942926), the US National Institutes of Health (grant nos. P30AG059307 and U01MH098953), and grants from Stanford HAI and the Chan Zuckerberg Initiative.

Author contributions

W.L., N.R., X.Y. and D.S.S. designed the study framework and oversaw the systematic analysis. W.L. and X.Y. conducted the linguistic analysis of the model cards. W.L., X.Y. and J.Z. wrote the paper, with substantial input from all authors. N.R., E.O., E.W. and Y.C. contributed to data collection and preprocessing for the intervention study. J.Z. provided the overall direction and planning of the project.

Competing interests

N.R. and E.O. are employees of Hugging Face. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00857-z.

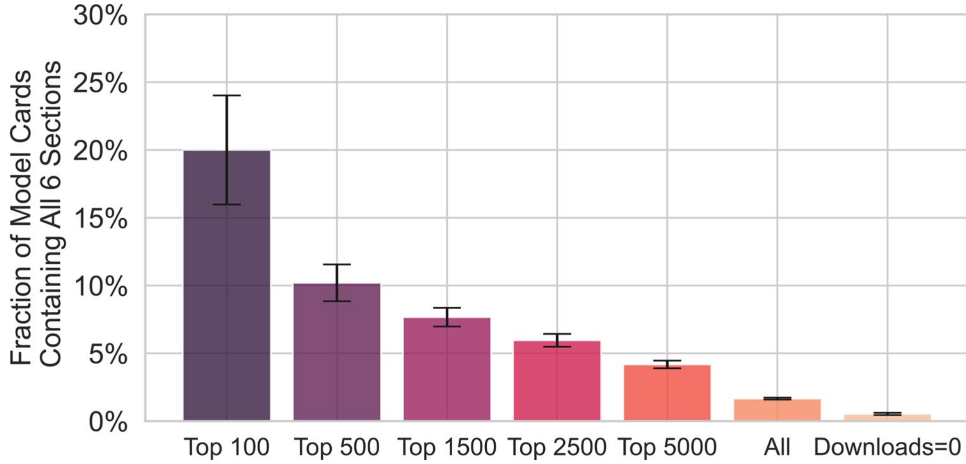
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00857-z.

Correspondence and requests for materials should be addressed to James Zou.

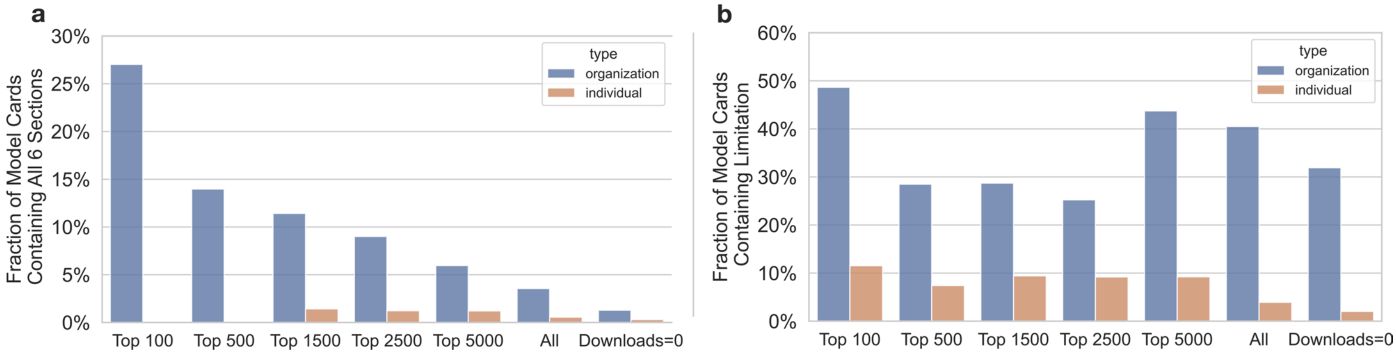
Peer review information *Nature Machine Intelligence* thanks Odd Erik Gunderson, Arvind Narayanan, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

 https://doi.org/10.1038/s42256-024-00857-z

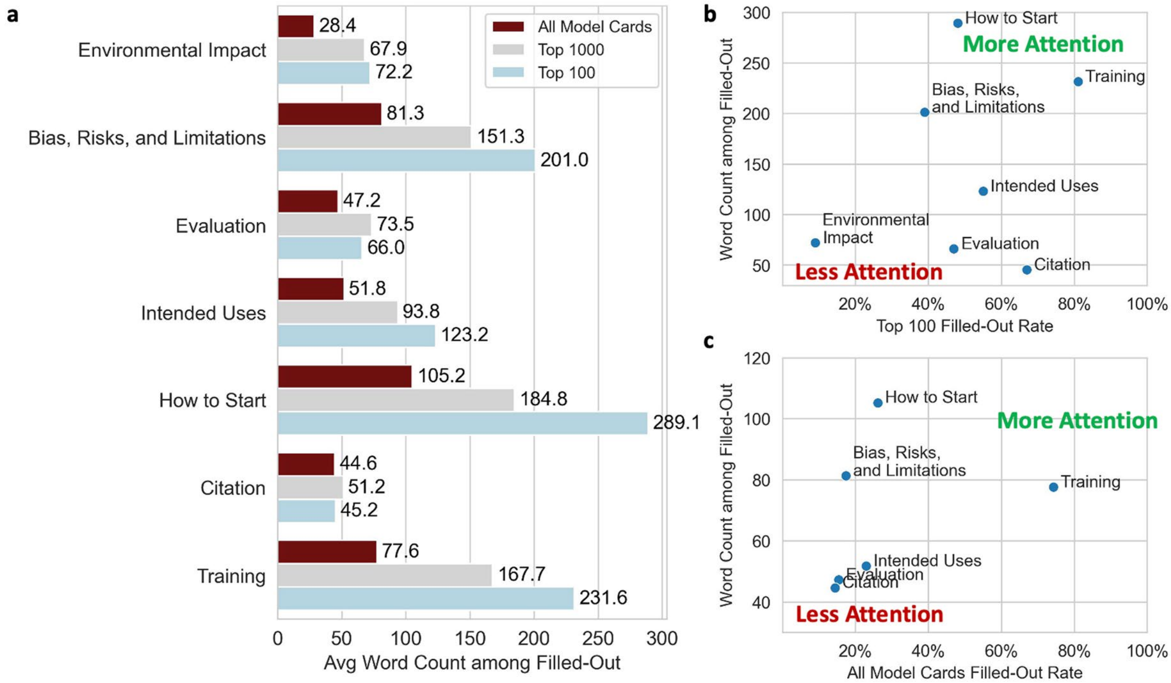


Extended Data Fig. 1 | Adherence to Hugging Face Model Card Norms. This figure reveals that there is considerable room for improvement in model card adherence to established community norms. Specifically, only 20% of the top 100 model cards and 10.2% of the top 500 model cards fully incorporate all recommended sections. Furthermore, there is a significant correlation between a model card's adherence to community standards and the model's downloads. The fraction of model cards that adhere to the community norms, grouped by download frequency, is displayed on the y-axis with error bars representing the SEM within each group.



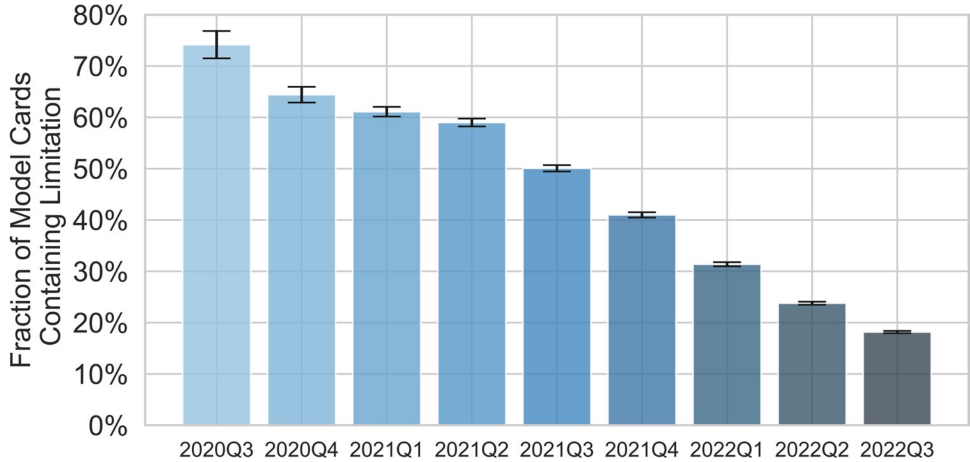
Extended Data Fig. 2 | Differences in Model Card Practices Between Organizational and Individual Accounts. This figure illustrates the disparities in model card practices between organizational and individual accounts. Download rankings are based on the entirety of model cards. For each account type, we count models of this type in every download group and calculate the percentage meeting to specified criteria. It highlights (**a**) *The Degree of Adherence to Model Card Standards* and (**b**) *The Completion Rate of the Limitations Section*. Organizational accounts show significantly greater compliance with model card norms, especially noted in their more thorough documentation of limitations across various download groups.

https://doi.org/10.1038/s42256-024-00857-z



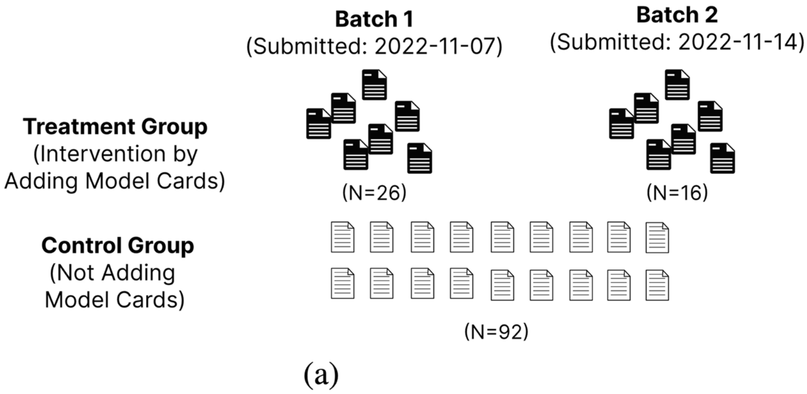
Extended Data Fig. 3 | In-depth Analysis of Section Word Counts in Model Cards. (**a**) *Comparative Assessment of Average Section Lengths in Model Cards Based on Word Count*. This figure displays the average section length, measured in word count, among completed sections for all model cards, the top 1000 model cards, and the top 100 model cards. Sections such as How to Start, Training, and Limitations are substantially longer, while Citation, Evaluation, Environmental Impact, and Intended Uses are relatively shorter. Interestingly, despite its lower completion rate, the Limitations section exhibits one of the highest average word counts (161 words in the top 1000 model cards). (**b**-**c**) *Disparate Community Attention Patterns Across Model Card Sections, Analysed for both the top 100 model cards* (**b**) *and all model cards* (**c**). The Environmental Impact section demonstrates both a low completion rate and a low average word count, indicating limited community attention. In contrast, the Training section displays high completion rates and average word counts, signifying greater community engagement.

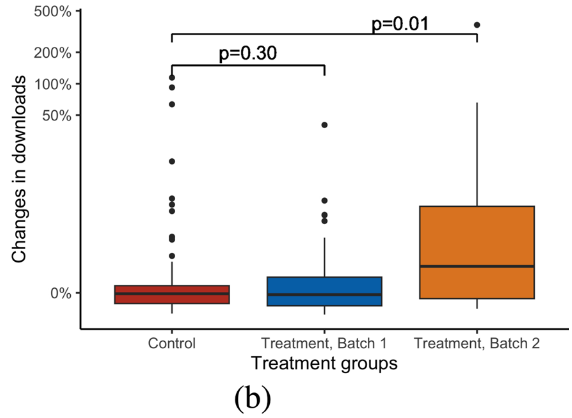
https://doi.org/10.1038/s42256-024-00857-z



Extended Data Fig. 4 | Temporal Trends of Fraction of Model Cards Containing Limitation Section. This figure illustrates the quarterly trends in the proportion of model cards that contain a limitations section, from 2020 to 2022. It highlights a noticeable decline in the occurrence of Limitation sections in model cards over time. Error bars in the plot represent the SEM, indicating the variability of the data within each quarter.

https://doi.org/10.1038/s42256-024-00857-z





Extended Data Fig. 5 | Model Card Intervention Study. (**a**) Experimental design: A schematic representation of the Model Card Intervention Study, delineating the selection of models, division into treatment (two batches) and control groups, and model card intervention process. Analysis is conducted on the document level (26 in batch 1, 16 in batch 2, and 92 in control). (**b**) Outcome: Box plots displaying the percentage change in average weekly downloads for the treatment and control groups in Batches 1 and 2. For each colour-filled box, three horizontal lines correspond to the 25th, 50th, and 75th percentiles; the upper (lower) whiskers extend from the 75th (25th) percentiles to the largest (smallest) value no further than 1.5 × interquartile range. Statistical significance (two-sided p-values) derived from a difference-in-difference analysis (using robust linear regression) is included for both batches. Overall, our analysis revealed a moderate effect of model cards on model downloads.

Extended Data Table 1 | Temporal trends in the completion rates of different model card sections

DV: Section Filled-Out Rate		Category: Audio	Category: Computer Vision	Category: Multimodal	Category: NLP	Category: RL	Category: Tabular	Category: unknown	Time (week)
Section: Environmental	Coef	-0.0153	0.0100	0.0041	0.0159	-0.0188	-0.0116	-0.0092	0.0002
	P-Value	9.75E-11	1.63E-02	4.10E-01	2.34E-25	1.67E-11	2.71E-02	3.22E-04	4.64E-24
Section: Limitations	Coef	-0.0236	0.2102	0.1288	0.1180	-0.0244	0.0329	0.0863	-0.0028
	P-Value	1.27E-04	3.22E-83	1.71E-23	1.46E-191	8.01E-04	1.61E-02	4.92E-38	0E+00
Section: Evaluation	Coef	-0.0029	0.0007	0.0051	-0.0273	-0.1263	0.5149	-0.0540	-0.0008
	P-Value	6.25E-01	9.48E-01	6.86E-01	2.09E-12	3.20E-70	0.00E+00	1.46E-16	5.51E-43
Section: Uses	Coef	-0.1409	0.0242	0.2876	-0.1179	0.6120	-0.2201	-0.0621	-0.0007
	P-Value	5.08E-109	2.98E-02	9.21E-104	3.59E-181	0.00E+00	3.40E-55	1.64E-19	1.35E-25
Section: How to Start	Coef	-0.1188	0.1180	0.0262	0.0164	-0.1820	0.8067	0.0811	-0.0031
	P-Value	1.06E-64	9.11E-22	7.33E-02	2.77E-04	3.94E-107	0E+00	1.08E-26	0E+00
Section: Citation	Coef	-0.1155	0.0372	0.0684	-0.1204	-0.1908	0.7742	0.0487	-0.0017
	P-Value	5.47E-96	1.36E-04	3.62E-09	4.51E-245	1.55E-184	0E+00	5.39E-16	1.77E-203
Section: Training	Coef	-0.0248	-0.0581	0.0912	0.2221	-0.4200	0.2974	0.1411	0.0023
	P-Value	2.81E-04	1.28E-06	1.62E-10	0.00E+00	0.00E+00	1.13E-85	1.04E-80	3.31E-244

The table displays the results of a linear regression analysis conducted to assess the changes in the fill rates of different sections of model cards over time, while controlling for model categories (for example, tabular and natural language processing). The analysis, which used a two-sided test, indicates a significant trend is the swift increase in the completion rate for the Training section, rising at 0.2% per week ($p = 3.31E-244$). In contrast, most other sections are experiencing a decline in their completion rates ($p < 0.001$). An interesting exception to this pattern is the Environmental Impact section, which demonstrates a rise in its completion rate ($p = 4.64E-24$).

https://doi.org/10.1038/s42256-024-00857-z

Extended Data Table 2 | Highly Downloaded Models Have Longer Model Cards

Model Card Word Count	Top 100 Model Cards	Top 200 Model Cards	Top 500 Model Cards	Top 1000 Model Cards	Top 2500 Model Cards	Top 5000 Model Cards	All Model Cards
Means	520.78	477.76	413.88	384.23	344.35	288.35	190.78
Stderrs	33.78	28.27	21.43	12.81	9.61	5.85	3.04

Model cards in the top tier, ranked by downloads, are notably longer, suggesting a positive correlation between model card length and their usage. The total word count across all the sections is shown on the y-axis.

https://doi.org/10.1038/s42256-024-00857-z

Extended Data Table 3 | Correlation between the comprehensiveness of model card sections and model downloads

DV: Log Downloads	Section: Environmental	Section: Limitations	Section: Evaluation	Section: Uses	Section: How to Start	Section: Citation	Section: Training
Coef	0.0043	0.0092	0.0070	0.0027	0.0079	0.0106	-0.0023
P-Value	1.25E-02	1.25E-35	2.31E-22	6.45E-06	5.44E-42	4.96E-46	7.12E-09

This table shows a two-sided linear regression analysis on the log-transformed counts of model downloads against the log-transformed word counts of each section. Limitations, How to Start, and Evaluation sections are the top three contributors to increasing model downloads. Except for the Training section, all others correlated positively with model downloads ($p < 0.001$).

https://doi.org/10.1038/s42256-024-00857-z

Extended Data Table 4 | Temporal Trends in the Proportion of Discussions Related to Data in Limitation and Training Sections

Proportion of Discussions Related to Data within the Section	Year: 2022	Year: 2021	Year: 2022
Section: Limitation	30.1% (17/55)	30.1% (43/143)	30.4% (105/345)
Section: Training	38.1% (56/147)	34.7% (165/475)	30.7% (198/645)

The percentage of discussions that address data limitations within model cards has remained relatively unchanged over time: 30.1% (17 out of 55) in 2020, almost identical at 30.2% (43 out of 143) in 2021, and slightly increased to 30.4% (105 out of 345) in 2022. Conversely, in the Training section, the proportion of discussions focused on training data has been on a decline, dropping from 38.1% in 2020 to 30.7% in 2022.

nature portfolio

Corresponding author(s): James Zou

Last updated by author(s): May 12, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected through Hugging Face Hub API.
Data analysis	Data was analyzed with customized code in Python 3.10 using standard software packages along with huggingface_hub 0.13.9 and mistune 2.0.4. The analysis code is publicly available at https://github.com/Weixin-Liang/AI-model-card-analysis-HuggingFace.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The Hugging Face model cards data are public on Hugging Face at https://huggingface.co/models and can be accessed through the Hugging Face Hub API at https://huggingface.co/docs/huggingface_hub/package_reference/hf_api.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We conducted a comprehensive quantitative analysis of 32,111 AI model documentations on Hugging Face, a leading platform for distributing and deploying AI models.
Research sample	32,111 AI model documentations on Hugging Face platform.
Sampling strategy	We analyzed all non-empty model documentation hosted on Hugging Face platform as of October 1st, 2022.
Data collection	Data was collected through Hugging Face Hub API.
Timing	We analyzed model documentation hosted on Hugging Face by October 1st, 2022.
Data exclusions	Not applicable.
Non-participation	Not applicable.
Randomization	Not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*