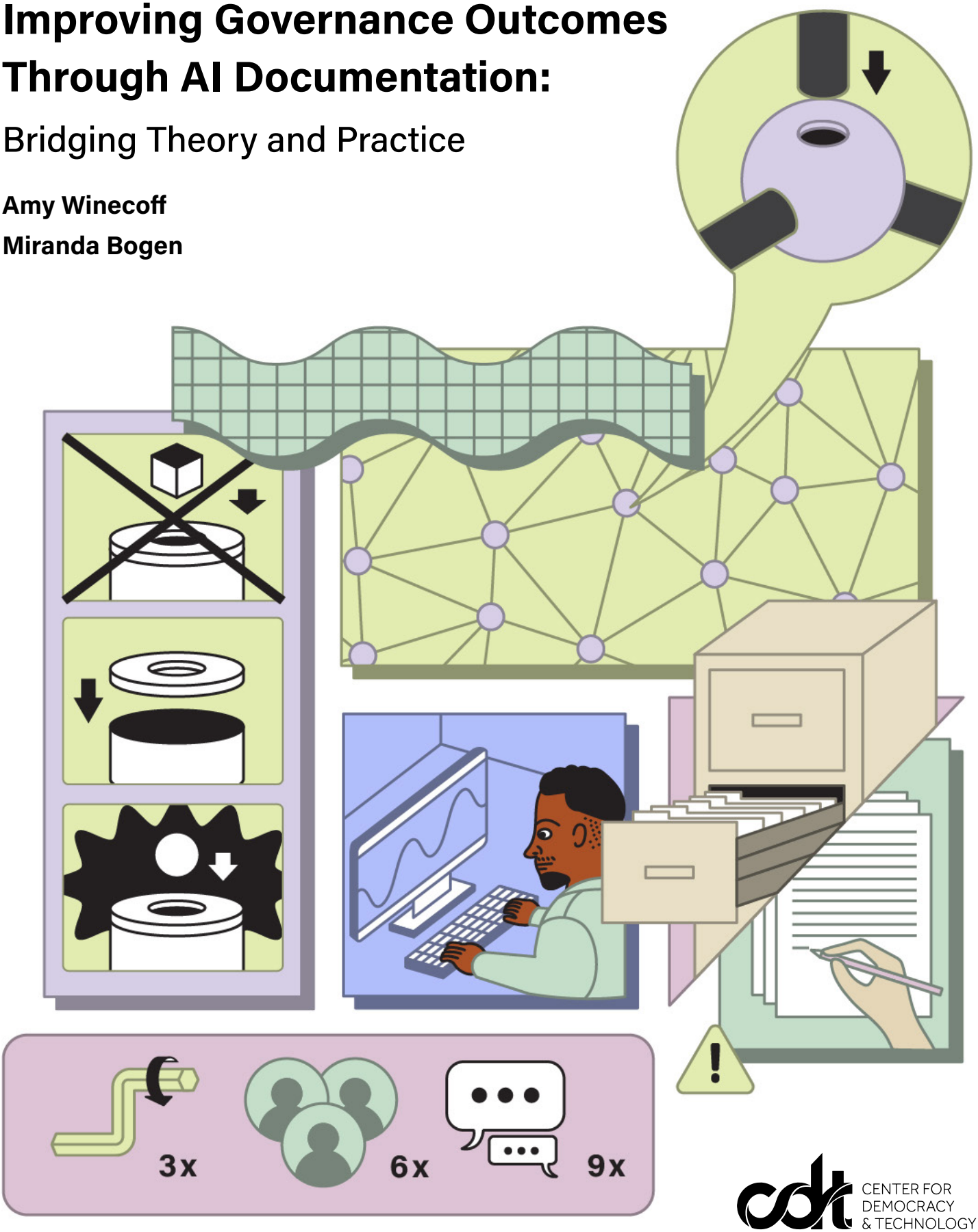
SEPTEMBER 2024

Improving Governance Outcomes Through AI Documentation:

Bridging Theory and Practice

Amy Winecoff

Miranda Bogen



CENTER FOR DEMOCRACY & TECHNOLOGY

The **Center for Democracy & Technology (CDT)** is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

CDT's AI Governance Lab develops and promotes adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems. The Lab provides public interest expertise in rapidly developing policy and technical conversations, to advance the interests of individuals whose lives and rights are impacted by AI.

SEPTEMBER 2024

Improving Governance Outcomes Through AI Documentation

Bridging Theory and Practice

Amy Winecoff
Miranda Bogen

Illustration and print layout by Gabriel Hongsdusit.

Acknowledgements

Thanks to members of the AI Governance Lab team at CDT, Emily McReynolds, Drew Courtney, Tim Hoagland, and Samir Jain their feedback and to participants of the "Improving Documentation for AI Governance" workshop, held at CDT in June 2024, for their valuable insights.

CENTER FOR
DEMOCRACY
& TECHNOLOGY

ES

Executive Summary

AI documentation is a foundational tool for governing AI systems, via both stakeholders within and outside AI organizations. It offers a range of stakeholders insight into how AI systems are developed, how they function, and what risks they may pose. For example, it might help internal model development, governance, compliance, and quality assurance teams communicate about and manage risk throughout the development and deployment lifecycle. Documentation can also help external technology developers determine what testing they should perform on models they incorporate into their products, or it could guide users on whether or not to adopt a technology. While documentation is essential for effective AI governance, its success depends on how well organizations tailor their documentation approaches to meet the diverse needs of stakeholders, including technical teams, policymakers, users, and other downstream consumers of the documentation.

This report synthesizes findings from an in-depth analysis of academic and gray literature on documentation, encompassing 37 proposed methods for documenting AI data, models, systems, and processes, along with 21 empirical studies evaluating the impact and challenges of implementing documentation. Through this synthesis, we identify key theoretical mechanisms through which AI documentation can enhance governance outcomes. These mechanisms include informing stakeholders about the intended use, limitations, and risks of AI systems; facilitating cross-functional collaboration by bridging different teams; prompting ethical reflection among developers; and reinforcing best practices

in development and governance. However, empirical evidence offers mixed support for these mechanisms, indicating that documentation practices can be more effectively designed to achieve these goals.

Our report also outlines the design trade-offs organizations must consider when developing and implementing documentation strategies. For example, customized documentation can address specific risks but may reduce comparability across documentation artifacts, whereas standardized formats promote consistency and institutionalize norms of practice but may overlook details relevant to particular systems. Organizations may also face decisions about whether to create a single, general-purpose documentation artifact or multiple tailored artifacts; while the multiple tailored formats may better serve diverse stakeholders, they are more challenging to maintain. Also, organizations must carefully determine the appropriate level of detail to include in documentation artifacts—excessive information can overwhelm users, while insufficient detail may omit critical information. We also explore the trade-offs involved in automating the documentation process and the choice of whether to develop interactive interfaces that allow stakeholders to explore the documentation more thoroughly.

The report concludes with recommendations for designing effective documentation processes. These include realistically assessing an organization's capacity for implementation, identifying the needs of key stakeholders, prioritizing essential details, and regularly evaluating progress against specific success criteria.

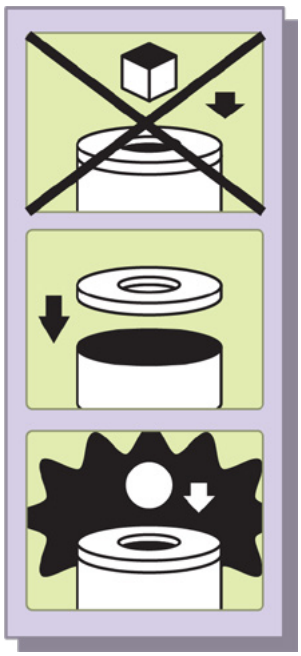
By carefully designing and implementing documentation processes that address the needs of diverse stakeholders, organizations can establish a strong foundation for robust AI system risk management. Moreover, by regularly assessing and refining their documentation practices, organizations can contribute to improved AI governance over time.

Contents

01

Introduction

Transparency in AI is widely regarded as essential for appropriately calibrating stakeholder trust and supporting accountability efforts.[1] By providing stakeholders with a clear view of an AI system's composition, operation, and development process, transparency allows for informed oversight and critical evaluation. AI documentation lays the foundation for transparency into AI systems and their construction. When AI systems are thoroughly documented, these documentation artifacts offer invaluable insights into important system features such as training data, algorithms, and risk management strategies. This information can support two broad groups of external stakeholders. Documentation can help downstream deployers and users understand how the system functions and what risks it conveys. It can also assist policymakers and researchers in holding companies accountable for the negative consequences of their AI technologies.



Yet the importance of documentation extends beyond external accountability.[i] Internally, documentation serves as a critical tool for managing AI systems throughout their lifecycle for a wide range of stakeholders. For example, documentation could help downstream technical practitioners understand the strengths, limitations, and risks of training a model on a given dataset. It might also help compliance and governance teams assess a potential system use case's compliance with company policies or legal requirements. Or

i Though the concepts of documentation and transparency are often blurred, complete documentation about an AI system does not guarantee that information will be made available to interested stakeholders, nor do transparency artifacts always reflect the full or most relevant details of an AI system even though such details may have been captured in internal documents. We distinguish between the concepts in this report in order to provide greater clarity about the presumptive goals of documentation and an exploration of what practices may more likely support those objectives.

the process of producing documentation might lead practitioners to make different decisions about how to sufficiently mitigate potential harms. As such, documentation is a foundational building block of effective AI risk management within organizations.

AI documentation lays the foundation for transparency into AI systems and their construction.

Despite its clear benefits, creating effective documentation is a complex task. No single form of documentation can meet the diverse needs of all stakeholder groups equally. Detailed technical reports on AI systems like Llama-2[2] and GPT-4,[3] which can exceed 50 pages and contain dense technical language, are valuable for AI researchers but may not be accessible to non-technical stakeholders. Similarly, technical documentation mandated by regulations such as the EU AI Act are intended to help policymakers evaluate AI risk management. However, such documentation will likely lack practical details for deployers who aim to effectively integrate these systems into their software.

To maximize the impact of documentation on AI governance, organizations must carefully define their goals and identify who will produce, maintain, and use the documentation. Tailoring the scope, level of detail, and format of documentation to suit the intended audiences and purposes is crucial. Effective documentation design requires balancing the needs of both the documentation process and the resulting artifacts, ensuring that they collectively support the organization's governance objectives.

This report presents findings from an in-depth research effort by the AI Governance Lab, exploring how documentation processes and artifacts can best support AI governance and risk management goals. Our research encompassed a review of academic and gray literature,[ii] identifying 37 proposed documentation frameworks and 21 papers with relevant empirical findings. We also incorporated insights from a multistakeholder convening hosted by CDT in June 2024.

ii Gray literature refers to publications that have not been peer-reviewed, but that present detailed theories or research findings.

The report is structured to provide a comprehensive overview of AI documentation's role in governance. We begin with a brief overview of the types of information organizations might document.[iii] We then identify four key mechanisms through which documentation could improve AI governance:

- Informing stakeholders about responsible use,
- Facilitating collaboration on AI risks,
- Encouraging ethical considerations, and
- Improving overall governance practices.

We also examine the practical barriers that may hinder the effectiveness of these mechanisms and the design considerations organizations must account for when implementing documentation processes.

These considerations include:

- Balancing customization with standardization,
- Deciding between single or multiple forms of documentation,
- Determining the extent of detail to include,
- Choosing the appropriate level of automation, and
- Selecting between static or interactive formats.

Finally, we offer recommendations to help organizations develop documentation strategies that align with their governance goals and stakeholder needs. Our findings aim to provide actionable insights for improving AI governance through better documentation practices.

iii The goal of this effort was not to review what organizations should document, but rather why and how they should document AI systems. As a result, we provide only a brief review of what organizations might document about data, models, systems, and methods. For a deeper synthesis, see https://partnershiponai.org/workstream/about-ml/

What organizations could document about AI systems

Documenting the various components of AI systems—data, models, systems, and methods—is essential for effective governance and informed decision-making. This documentation not only provides insight into the development process but also helps internal and external stakeholders understand the characteristics and potential risks associated with AI systems. Collectively, AI systems and component documentation support robust governance, informed decision-making, and effective risk management.

Within companies, many stakeholders may contribute to and use documentation. Practitioners involved in data collection, annotation, and curation could contribute to data documentation. Engineers developing data, models, and application infrastructure could contribute to and use documentation. Data scientists may create and use documentation related to data, models, and analyses. User experience (UX) designers, researchers, and product managers may use documentation to inform product designs and develop relevant public explanations. Responsible AI creators as well as governance and compliance professionals may use documentation to triage risk management work and rely on that documentation to identify and manage system risks. Organizational leaders look to elements of documentation to inform important decisions, such as whether to invest substantial resources in mitigating particular risks or whether to proceed with the launch of AI-powered products in light of the risks and benefits that have been identified.

We can categorize documentation stakeholders into two broad groups: documentation producers and documentation consumers. Documentation producers are those who actively contribute to the generation of documentation artifacts, while documentation consumers are those who read and use it. While this grouping simplifies some important distinctions, it serves as a helpful

heuristic for understanding the opportunities and tensions between groups with both shared and distinct goals. We consider both producers and consumers and the tension that might emerge between them in our report.

"AI documentation" could refer to either the process of producing documentation or the resulting artifacts. While documentation artifacts like datasheets[4] and model cards[5] inform stakeholders about the responsible use of AI, the process of creating these documents can also institutionalize best practices and foster a culture of risk management. Therefore, when considering how documentation can improve governance outcomes, we address both artifacts and processes.

AI systems and component documentation support robust governance, informed decision-making, and effective risk management.

Data

Data is a fundamental component of AI systems since it significantly influences model behavior and performance. When models are trained on datasets that do not reflect their deployment contexts, they often do not perform well. And of course, if models are trained on data reflecting existing societal biases, they risk further amplifying them. Data documentation can play a crucial role in helping practitioners identify potential issues with training datasets, thereby supporting more robust model performance and conveying important considerations to documentation consumers.[6]

Data documentation frameworks[7] typically recommend recording characteristics of each dataset and how it was created and processed. Data documentation often includes details on dataset contents, such as what each instance represents, whether labels are associated with the data, and whether instances are linked, such as in social network data. Documentation frameworks often suggest including statistical summaries of datasets, like minimum, maximum, and average feature values or histograms of their distributions. This information helps practitioners assess the appropriateness of a dataset for a specific use case.

Data documentation might also describe the data collection method, including information about the equipment used to acquire the data (e.g., sensors for environmental data or cameras for photographic images), the sources from which the dataset was derived (e.g., social media posts, online news articles), which data were included and excluded from a set, and how labels were applied (e.g., by human annotators or classification models). Understanding the data collection methods can help practitioners recognize any confounding issues or limitations that might affect how the data is best used (or not used).

Frameworks often recommend describing any preprocessing and cleaning methods applied to the data. Preprocessing might include imputation for missing values, format transformations, discretization (i.e., partitioning continuous data into buckets), tokenization (i.e., breaking larger sections of text into smaller units), and similar techniques. This information is crucial because preprocessing methods affect a dataset's compatibility with a chosen task.

Data documentation may also enumerate any known limitations or constraints on use. For example, if a dataset contains personally identifiable information, practitioners may need to employ particular mitigations to ensure their models do not later leak sensitive information. Documentation frameworks often suggest specifying the intended uses of the dataset and identifying contexts where the dataset should not be used. While it may be impossible to anticipate all potential uses, documenting some examples of in- and out-of-scope uses provides valuable insight into appropriate application contexts to potential users of that dataset who are likely to be less attuned to its potential utility and risks. Documentation of information about licenses or compliance with company policies and regulations can guide practitioners toward acceptable use.

The specific information practitioners should document may vary based on the data type and application context. For instance, speaker or author demographics may be crucial for training language models, while images' size and color range may be necessary for training computer vision models. Documenting the equipment used to collect data is necessary if it significantly affects the dataset's characteristics or risks (e.g., recording equipment for voice datasets or camera specifications for artwork[8]

images). When determining which information to include in their data documentation, organizations may benefit from consulting frameworks for documenting data that are most similar to their own. However, they will likely need to adapt the frameworks further to cater to their organizational context, products, and stakeholders.

Models

Model documentation can improve AI application development and governance. It provides downstream practitioners with insights into a model's development, performance, and potential risks, helping them make informed decisions about its use. Effective model documentation can assist developers in comparing several models (including considering non-AI approaches) and tailoring them to specific contexts. Model documentation can also identify potential harms to users or other stakeholders and highlight critical information that documentation consumers should consider when assessing whether the model is likely to pose high or unacceptable risk, and help downstream practitioners determine what mitigations to apply given a model's assessed risks.

Model documentation frameworks[9] often recommend recording essential details about the model, including its type or architecture, version, training procedures, and parameters. This information benefits those more familiar with machine learning, as different development methods come with different assumptions.[10] In some cases, noting the hardware used for development and evaluation can be important due to its potential impact on model performance.[11]

Practitioners often develop models with specific use cases in mind and choose their development methods accordingly. Documenting the intended use case helps downstream consumers assess potential mismatches between assumptions implicit in the original development aim and their own applications. And it is often important for practitioners to describe instances where the model should not be used and explain why these uses are inappropriate.

Effective model documentation should include details on how the model was tested for both performance and relevant risks,

enabling downstream stakeholders to assess the relevance and durability of those measurements in new contexts and the potential need for additional testing given a particular deployment context. The way practitioners evaluate models—through their choice of methods, metrics, and evaluation datasets—significantly influences the resulting performance estimates. Practitioners typically select metrics that they believe will best reflect the model's real-world performance and choose datasets that closely resemble the data the model will encounter in its operational environment.

For instance, if a music recommendation model is evaluated based on how accurately it mirrors users' existing preferences, the results might not be particularly relevant for a team looking to use the model to help users discover new genres. Similarly, if the model is tested on music released the previous year, those performance estimates may not accurately predict its performance on current-year music. Therefore, documenting the evaluation datasets and techniques used provides downstream stakeholders with crucial insights into the applicability of reported evaluations to their intended use. Detailed descriptions of evaluation methods also help documentation consumers interpret evaluation results, especially when they cannot directly consult the model developers, such as in the case of deployers using open-source models or off-the-shelf models accessed via API.

Some models are developed in stages, such as those pretrained on large datasets and later fine-tuned for specific tasks or adapted to prevent undesirable outputs through methods like reinforcement learning from human feedback (RLHF).[12] Documenting specifications, training processes, and evaluation methods for each development stage can help downstream consumers understand the development pipeline and its implications for their intended use case.

For pre-trained models adaptable to multiple downstream applications (i.e., "general-purpose models"), highlighting specific considerations for adaptation and use is often helpful. For example, methods to mitigate illegal, biased, or harmful outputs in pre-trained large language models may not hold when the model is fine-tuned[13] or otherwise adapted to perform specific tasks.[14] Therefore, documentation consumers might benefit from any details that documentation producers can provide about the robustness of safety guardrails to downstream adaptation or the relevance of documented pretrained model properties to various downstream contexts. Most likely, downstream users will need to conduct their own contextually relevant safety, fairness, and privacy evaluations to fully understand and mitigate potential risks in the final model implementation and should not rely upon the documented safety properties of upstream models.

Effective model documentation should include details on how the model was tested for both performance and relevant risks, enabling downstream stakeholders to assess the relevance and durability of those measurements in new contexts.

Systems, Applications, and Tasks

AI services in the real world typically consist of complex systems rather than isolated models. These systems often include multiple components, such as machine learning models, trust and safety filters, third-party software integrations, and user interfaces. The interaction among these components significantly affects the system's overall performance and risk profile — especially because the properties or behavior of individual components may change when integrated into a system.[15] Recognizing that complex AI systems are more than the sum of their parts, several research groups have proposed frameworks for documenting AI systems in their entirety.[16]

System documentation frameworks recommend organizations provide an overview of the system, including a description of its objectives, inputs and outputs, and a diagram illustrating the interactions between different components. This diagram can link to more detailed documentation for each component. Like model documentation, system documentation may describe evaluation methods and include aggregate and disaggregated results based

on relevant factors. Documentation producers may want to identify potential risks and specify contexts in which the system should not be used, especially if these considerations are not apparent from the documentation of individual components.

Documentation on development processes should ideally be produced concurrently with the development of the system.

Procedures & Methods

Organizations may also find it important to document the processes and procedures that shape the development and deployment of AI systems and their components. This documentation could include how humans annotated or evaluated training or evaluation data,[17] any privacy, legal, and ethical reviews or audits conducted,[18] and any risk mitigations that were applied. Investing time in this type of documentation may be especially beneficial because it reveals information that cannot be deduced merely from examining the system components. If such meta-data about the process is not documented throughout the development lifecycle, it may be very difficult to reconstruct later.[19] As a result, documentation on development processes should ideally be produced concurrently with the development of the system.

Organizations may also benefit from documenting machine learning methods,[20] such as techniques applied in computer vision applications[21] and strategies for enhancing model explainability.[22] Documentation of the application of the AI model or system to specific AI tasks or use cases[23] can also help practitioners understand the benefits and limitations of these applications. Information about methods and processes can provide valuable insights into what potential approaches a team might take to develop an AI system and the relative merits and drawbacks of these approaches. Such insights can help practitioners determine when particular approaches are more or less appropriate and what risks different approaches may confer.

02 How AI Documentation Can Support AI Governance

While empirical evidence directly linking documentation artifacts or processes to improved governance outcomes is limited,[iv] the theories proposed in existing documentation frameworks offer compelling hypotheses about how documentation can positively impact governance. Empirical studies involving practitioners, alongside analyses of publicly available documentation artifacts, provide an initial glimpse into the factors that contribute to accurate, high-quality, and comprehensive documentation, as well as the reasons why documentation efforts may sometimes fall short of their goals.

Our review identified four possible ways in which documentation could improve governance: informing downstream users about system development and associated risks, encouraging ethical reflection among practitioners, facilitating communication among stakeholders, and enhancing AI development and governance overall. In the following sections, we explore each of these hypothesized impacts in detail and examine the challenges organizations face in realizing these benefits.

iv Most framework authors validate their approaches by themselves documenting an actual or theoretical dataset, model, or system, which raises questions about whether practitioners with less motivation or knowledge of the documentation process would achieve similar results. In our review of 37 frameworks, only eight provided empirical evidence to guide the design, implementation, or evaluation of their proposed methods. For additional arguments about the necessity of evidence for best practices in documentation, see Winecoff and Bogen, "Best Practices in AI Documentation: The Imperative of Evidence from Practice."

Informing downstream stakeholders

The most straightforward potential impact of documentation is that it can provide essential context, such as the original motivation for developing a dataset, model, or system; the intended use cases; and details on system properties necessary for effective and responsible deployment.

Hypotheses

- **Documentation can clarify the intended uses of data, models, and systems and provide guidance on responsible deployment.** Practitioners typically develop systems with specific use cases in mind, and other development teams need to understand these original use cases to assess alignment with their goals. If aligned, teams can then modify the systems to meet their current objectives. Practitioners argue documentation can also clarify underlying assumptions, such as data collection methods and representativeness, which are critical for assessing the suitability of a dataset for specific purposes.[24] They also feel documentation could help identify potential deployment issues, such as a mismatch between the original training data and the intended deployment context.[25]

  Moreover, researchers and practitioners argue that documentation could assist downstream practitioners in making implementation decisions that support effective product development. For example, documentation could help practitioners find and compare available system components aligned with their goals,[26] including non-AI approaches.[27] It might also provide statistical summaries of data[28] or model performance evaluations such as area under the curve (AUC)[29] to help inform developers' choices of how to train models or implement appropriate guardrails.

- **Documentation can help stakeholders determine whether their planned use aligns with legal requirements and organizational policies.** Documentation could also help stakeholders ensure that planned use cases comply with legal constraints, regulatory requirements, and organizational policies. For example, certain datasets may be barred from use if the data were collected without affirmative consent or contain personally identifiable or copyrighted information. Without this awareness, practitioners might inadvertently train a model on such data, only to later discover that the model cannot be used due to policy violations.[30] Practitioners see avoiding disciplinary action for non-compliance with company policies as a key benefit of documentation.[31]

Challenges

While practitioners acknowledge that the benefits of documentation could outweigh the costs of creating and maintaining it,[32] they face significant challenges in implementing robust documentation practices in real-world organizational contexts. When organizations fail to provide clear incentives for high-quality documentation, practitioners are less likely to make documentation a priority. This lack of motivation can hinder the realization of documentation's potential benefits, particularly its ability to inform downstream users. When documentation is not prioritized, practitioners often create lower-quality artifacts that may inadequately inform or even mislead those who rely on them.

Lack of organizational incentives limits practitioners' attention to documentation. Organizations typically do not incentivize documentation production unless regulations require it for compliance or clients specifically request it.[33] Practitioners often view producing high-quality documentation as less relevant to their evaluations and promotions than development tasks that directly contribute to products.[34] Furthermore, the practitioners who benefit most from documentation are often not the same as those responsible for producing it, creating a disconnect between documentation benefits and practitioner objectives. In organizations

where governance processes mandate documentation, a lack of familiarity with AI documentation concepts among developers could potentially magnify tensions between product development and governance teams.[35]

Facing time constraints and competing demands, practitioners may deprioritize documentation, sometimes cutting corners to save time.[36] This can diminish the value of documentation artifacts, as evidenced by studies showing that practitioners sometimes reuse content from existing documentation, even when it pertains to different systems, leading to inadequate or incorrect documentation of their current system.[37] Practitioners also sometimes leave questions unanswered rather than seek out necessary information,[38] and they prioritize recording information that is valuable to them, while omitting details critical for others who are less familiar with the system.[39] Because practitioners often forget relevant information if they do not record it during development,[40] their reconstruction of key system information after the fact can be time-consuming and error prone.[41] Also, practitioners are often more motivated by compliance than normative considerations,[42] which could create blindspots as novel ethical issues outside of existing policies arise.

When documentation is not prioritized, practitioners often create lower-quality artifacts that may inadequately inform or even mislead those who rely on them.

Low-quality artifacts could misinform or mislead documentation consumers. Studies of publicly available documentation artifacts frequently highlight issues with incomplete or incorrect information, which can mislead downstream stakeholders regarding the characteristics of AI systems or components.[43] For example, many publicly available model cards lack details about out-of-scope uses, limitations, or environmental impacts. For example, less than 30% of model cards contain evaluation results, which is highly relevant information for many downstream documentation consumers.[44] Furthermore, publicly available model documentation sometimes contains incorrect license information, potentially leading those who rely on this documentation to violate license terms inadvertently.[45]

In sum, while documentation artifacts have the potential to guide AI practitioners towards informed and responsible use, current norms, priorities, and organizational dynamics often limit the extent to which they provide complete and correct information and well-thought-out guidance.

Supporting cross-functional collaboration

Documentation has the potential to act as a bridge between different stakeholder groups involved in the AI development lifecycle, including machine learning engineers, data scientists, product managers, user experience designers, researchers, and legal teams. Since no single group oversees all stages of development, effective documentation can facilitate both indirect and direct communication among these diverse teams. Documentation artifacts can convey essential information across organizational boundaries, helping stakeholders who may not interact regularly to align their understanding of the system. Documentation can also prompt and support conversations among stakeholders directly, enabling them to work towards common goals despite their diverse backgrounds, frames of reference, or areas of expertise.[v] This bridging function of documentation could break down organizational silos[46] and promote effective collaboration across different teams.[47]

v In science and technology studies (STS), ideas, concepts, and artifacts that provide an interface for communication between different groups of people are sometimes referred to as "boundary objects." Boundary objects can serve distinct functions within the groups of people who use them, but also support coordination and understanding across groups. Documentation can serve as a boundary object since it might be used differently by developers, compliance professionals, user researchers, and other stakeholder groups, but still help each of these groups work effectively with one another.

Hypotheses

- **Documentation prompts communication between stakeholders to clarify basic details.** AI documentation could serve as an initial point of reference for downstream practitioners, offering a basic system overview that allows them to build foundational knowledge before engaging with developers for more in-depth discussions. Rather than acting as a comprehensive, self-contained repository of information, documentation artifacts could provide an initial layer of understanding, enabling downstream practitioners to identify relevant areas for further exploration and formulate more informed questions when collaborating with developers. This could foster more productive and targeted interactions between teams, enhancing their ability to work with or govern AI systems efficiently. Empirical research suggests that practitioners often approach documentation this way. For instance, UX professionals often prefer direct discussions with data scientists rather than relying solely on documentation artifacts.[48] Similarly, many AI practitioners review existing documentation only briefly before seeking further clarification through meetings or discussions.[49]

- **Documentation enables collaborative interrogation of documentation artifacts and AI systems.** In more robust forms, documentation can help stakeholder groups deliberate about the system itself and how they should document its characteristics to promote responsible use. For example, one study showed that data documentation helped teams deliberate on the socially constructed nature of gender and make collective decisions about how to annotate gender within datasets.[50] Another study found that the process of creating documentation aided practitioners in identifying and discussing trade-offs between competing objectives in system design.[51]

  By assigning some stakeholders to define requirements for documentation artifacts, others to generate content, and still others to assess the quality of the artifact, the documentation process can help stakeholders understand each others' needs and constraints in ways that could promote quality, efficiency, and responsibility in the development process.[52] Over time, documentation might also improve the technical literacy of those

who engage with documentation artifacts or processes, helping organizations create products that meet users' needs without causing harm.[53]

Challenges

- **The modern AI supply chain presents barriers to direct stakeholder communication**. Despite its potential, several challenges complicate the role of documentation in facilitating cross-functional collaboration. The modern AI supply chain, characterized by the development of general-purpose models that can be adapted to a variety of applications downstream,[54] often limits opportunities for direct communication between stakeholders. In some cases, different teams within the same organization may be responsible for developing the base model and adapting it to various specific tasks. In these instances, documentation might prompt the downstream team to directly reach out to the original development team to discuss the system's capabilities and limitations. In other cases, a team deploying a model may rely on a model developed by a different organization. In this case, it is typically more difficult if not impossible for the downstream deployment team to directly communicate with the original development team.

  In either situation, the original developers may not fully anticipate the variety of downstream uses, leading to documentation that lacks critical details needed to identify potential risks or harms. When upstream and downstream teams cannot communicate directly, it will be more challenging for them to work together to understand the model's risks within a given deployment context. This challenge is especially pronounced for less technical downstream stakeholders, who may struggle to understand the functionality and risks associated with these models.

  More interactive forms of documentation, such as those that allow users to engage with the models directly or customize the level of detail presented, could help mitigate these issues by making the information more accessible and relevant.[55] Nevertheless, more research is needed to determine how to

document the characteristics of general-purpose systems to convey both known limitations and the substantial uncertainties associated with their use. A particular challenge will be conveying this information without overwhelming documentation consumers with details that may be distracting or irrelevant, especially when communication between teams is limited.

- **Documentation that is misaligned with the needs of its intended consumers can complicate rather than facilitate stakeholder communication.** If documentation artifacts are overly technical or jargony, documentation consumers who are less technically proficient or less familiar with the system may struggle to understand which details are most relevant, or misinterpret described practices; if documentation lacks relevant details, documentation consumers may overlook risks that become apparent only later in the development process, which may require teams to substantially backtrack. In either case, mismatch in comprehension can contribute to what some refer to as "thrash," or needless disruption to normal workflows. If practitioners perceive that documentation leads to seemingly egregious friction due to misunderstandings, they may be more prone to oversimplify the documentation they produce or gloss over relevant details that they worry will lead to confusion.[56]

Prompting ethical deliberation

Researchers have suggested that documentation has the potential to encourage practitioners to consider the ethical implications of their systems, leading to more responsible development and use of data, models, and systems. By integrating ethical reflection into the documentation process, practitioners may become more aware of the potential harms their systems could cause, prompting them to make more informed decisions.

Hypotheses

- **Documentation artifacts alert documentation consumers to ethical risks**. One hypothesis is that when downstream documentation consumers consult documentation on data, models, and systems, it can prompt consideration of potential harms. This consideration could lead practitioners to make more careful decisions about if and when to use certain systems or components. Yet, empirical support for this hypothesis is inconsistent. One study of 23 practitioners examined whether reviewing data documentation artifacts could encourage practitioners to notice ethical issues, contextualize them, and formulate risk mitigation plans.[57] One subset received a scenario description, dataset, and data documentation, while the rest received only the scenario description and dataset. The researcher found that those with data documentation were more likely to notice ethical issues unprompted, though those without it often recognized concerns when the researcher pressed them — but practitioners had difficulty formulating action plans regardless of whether they had access to documentation. These findings suggest that while documentation artifacts can aid ethical deliberation, documentation is not always necessary and likely not sufficient to enable mitigation of ethical issues.

- **Producing documentation sensitizes practitioners to ethical impacts.** Documentation frameworks often ask documentation producers to consider and describe the ethical dimensions of their systems. For example, one documentation framework prompts practitioners to think about privacy implications by inquiring about sensitive information in data.[58] Another ask about known biases, ethical issues, and safety concerns,[59] and several frameworks encourage reflection on how systems might negatively impact marginalized users or populations.[60] Some researchers have suggested that documentation might be effective at promoting ethical action, even if it only engages practitioners in critical reflection relatively superficially.[61]

Yet others have taken a more explicit, deliberative approach. Drawing from value-sensitive design–a method that helps technologists identify and understand normative judgments in the development process[62]– Shen and colleagues developed the Model Authoring Toolkit to help practitioners consider diverse stakeholder values and deliberate on system design trade-offs.[63] Their qualitative and survey study of Wikipedia communities found that framing ethical reflection and documentation as a participatory process leads to more informed decisions about AI system design and deployment.

Challenges

- **Practitioners often lack awareness of the ethical impacts of AI.** The assumption that documentation artifacts and processes can promote ethical reasoning among practitioners hinges on the belief that practitioners are sufficiently aware of AI's ethical risks to users, non-users, and society that they can recognize them when prompted. However, without proper training in responsible AI practices and exposure to groups potentially harmed by AI systems, practitioners may not connect documentation with ethical risks.[64] Consequently, their consideration of harms may only partially encompass the scope of risks, especially those requiring non-technical solutions.[65] For example, practitioners often misunderstand how bias can manifest in their work, leading them to incorrectly assert in their documentation that bias concerns are not relevant.[66] While it may not be necessary for ethicists to be embedded in the work process for practitioners to make responsible choices, as some have suggested,[67] our findings point to the need for more support and training for practitioners in how to identify, manage, and effectively communicate potential risks.[68]

- **Practitioners may resist documenting ethical impacts they have identified.** Publicly available documentation artifacts, such as Hugging Face model cards and GitHub repository README files, often show little consideration of ethical concerns.[69] While this omission might occur because practitioners are unaware of potential harms, and therefore, do not document them,[70] their absence could also result from practitioners' hesitance to record

ethical concerns they did recognize. Sometimes practitioners do not readily document ethical considerations because they feel unqualified to speculate on numerous potential use cases and their possible impacts.[71] Others worry that detailing ethical concerns might give downstream stakeholders a false sense of security.[72] In our consultations with documentation stakeholders, practitioners also expressed concern that those with less AI literacy might misinterpret documentation on existing model risk mitigations or evaluations as guarantees of safety in deployment contexts. This hesitation could stem from a desire to avoid personal responsibility for their actions[73] or concerns that documenting potential ethical harms could create a legal or public relations risk if outside stakeholders gained access to documentation.[74]

- **Organizational priorities may constrain individual practitioners' ability to promote ethical development**. Even if practitioners are effectively engaged in ethical reflection by documentation, they may have little influence over their organization's development goals, deliverables, and timelines. As a result, they might be unable to make the changes to data, models, or systems that they have identified in documentation as useful or necessary to address ethical risks. One ethnographic study found that business demands, not the beliefs of data subjects or practitioners, largely determined the organization's documentation approaches.[75] For example, although practitioners who reflected on data labeling practices recognized that socially constructed identities are complex, the organization nevertheless chose to represent identity in a reductive way, such as by defining race according to discrete, mutually exclusive categories. The authors concluded that explicit and implicit power structures among internal organizational stakeholders significantly affect practitioners' documentation practices and ability to shape outcomes. Documentation approaches must be responsive to these constraints. Otherwise, documentation about ethical considerations may not promote meaningful action or may be incomplete or misleading.

Evidence on the extent to which AI documentation frameworks can truly enhance ethical decision-making is mixed. Practitioners are capable of ethical deliberation about AI, but do not always explore

ethical issues thoroughly or record the results in documentation artifacts. Even if practitioners engage in ethical deliberation when creating and using documentation, organizational environments can enable or constrain individual practitioners' normative choices.

Catalyzing best practices

Researchers have suggested that documentation can influence governance and development practices broadly. In practical settings, AI documentation doesn't function independently; documentation processes typically require practitioners to engage in other activities that can improve AI outcomes. Thus, the documentation process can catalyze behaviors that enhance the quality of AI system development and governance overall.

Hypotheses

- **Documentation can enhance scientific rigor in the AI development process.** Framework authors posit that documentation could improve the rigor with which practitioners develop AI systems.[76] By requiring practitioners to justify their development choices, documentation may lead to more careful decision-making.[77] Clear and comprehensive documentation can also support reproducibility,[78] aiding other practitioners in retraining systems consistently.[79] Since AI practitioners are often driven by a commitment to scientific rigor,[80] documentation approaches that emphasize this aspect can serve as a "value lever,"[81] further encouraging engagement with the documentation process.

- **Documentation can improve development and governance efficiency.** In large organizations, datasets, models, and systems developed for one purpose may also be useful to other teams with similar goals. Documentation can make system components more discoverable, helping practitioners avoid duplicating efforts.[82] Documentation can quickly provide information on the limitations of using a dataset, model, or system,[83] which could allow practitioners to allocate more time

to identifying and addressing any additional risks that may arise when integrating components into new systems. It can also include details relevant to company or legal policies that apply to system components, preventing practitioners from wasting time on products and features that their organizations ultimately won't approve for deployment.[84] Practitioners also point out that by documenting both unsuccessful and successful approaches, organizations can prevent repeated mistakes and identify practices that would be beneficial for the organization to disseminate broadly.[85]

Documentation might also facilitate proactive risk mitigation, which is often more efficient than reactive approaches.[86] When issues arise in already-deployed systems, organizations typically need to fix the problem without disrupting service quality or availability. The result might be a roll-back to a previous version of the system that is less performant or quick patches that may not fully resolve the issue. Documentation that helps identify potential problems before deployment can lead to more comprehensive solutions than addressing issues post hoc once they are discovered in production systems.[87]

- **Documentation can preserve institutional knowledge**. In any organization, particularly complex ones, consistently institutionalizing values and best practices can be challenging. Documentation could serve as a means of conveying technical information about systems, and the organization's policies and values that influence development and use. For instance, documentation can aid in onboarding new employees by providing information about the systems they'll work on and communicating the organization's approach to development and governance.[88]

  Furthermore, documentation is necessary to facilitate both internal and external audits.[89] Audits may focus on the components of AI systems, the overall application, or the processes used in development.[90] Documentation could help organizations demonstrate that their claims—such as providing

equitable system performance across different demographic groups—are backed by evidence. In these cases, documentation could guide auditors in selecting appropriate methods based on the auditor's objectives and the available data. When audits examine the robustness of an organization's governance, documentation becomes even more critical.[91] Without comprehensive documentation, organizations may struggle to demonstrate the integrity and consistency of their processes.

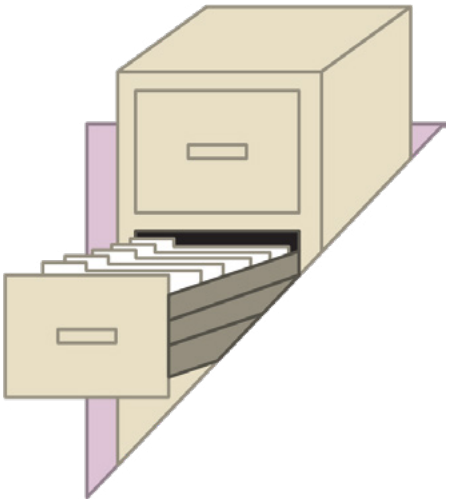
Challenges

- **Documentation requires robust integration with other processes.** The documentation process must be interconnected with other organizational processes to serve as an effective forcing function for governance and development best practices. In cases where documentation is ad-hoc and relies on tools that are not well-integrated with practitioners' workflows,[92] connecting documentation processes and artifacts to other aspects of the development and governance process becomes even more challenging. Without a solid documentation infrastructure that aligns with other organizational functions, documentation may fail to effectively enhance development and risk management.

- **Poor-quality documentation can have diffuse negative impacts on governance and development.** While high-quality documentation can improve governance and development, poor-quality documentation can have negative impacts. Documentation should prevent redundant efforts, maximize time spent managing unique risks, and support proactive system development. However, these benefits depend on the quality of the documentation. Inadequate or incorrect documentation can lead to wasted time, overlooked risks, and unexpected issues that may require correction later in development. This could result in harms remaining unaddressed or inefficiencies that undermine the overall governance process.

Overall, empirical evidence largely supports the idea that documentation can have benefits beyond the scope of the documentation artifacts and processes, despite some caveats.

03 Design Considerations that Impact Documentation Outcomes



How organizations choose to document AI systems reflects both their normative values and practical constraints. Design and implementation choices are also influenced by established and emerging external requirements, such as regulations that companies must adhere to. As organizations (and regulators) aim to maximize the benefits of documentation across various types of AI systems and components, they must often navigate trade-offs in their approach. Our analysis identified common design tensions and considerations organizations must navigate. These include the degree of customization versus standardization in documentation artifacts, the extent to which documentation artifacts should be tailored to specific audiences, the level of detail documentation artifacts should contain, the amount of automation in the documentation process, and the level of interactivity that documentation artifacts should support. To an extent, these design considerations are interrelated: the level of detail documentation artifacts contain can also be a form of audience tailoring, for example. Yet each of these design tensions has partially distinct motivations and implications, so we address them separately.

Standardization vs. customization

Organizations often create datasets, models, and systems for various purposes, employing a wide range of techniques and components. This diversity leads to a variety of risks, which customized documentation is well-suited to address. At the same time, standardized documentation helps establish consistent practices and institutionalize norms. The choice between customization and standardization is not a strict binary but a spectrum. For example, organizations can create standardized templates that are adapted to different contexts, such as healthcare models versus those for software engineers, ensuring consistency while meeting specific needs. However, organizations must thoughtfully decide when to standardize and when to customize.

- **Customization can allow documentation to address organizations' or AI systems' unique capabilities and risks**. For instance, annotating the origin of data collected in different healthcare settings can be crucial, as different settings serve distinct patient populations and employ varied healthcare practitioners. This information can help practitioners identify gaps in the dataset or regions where the dataset may be less applicable.[93] Conversely, annotating the origin of code snippets from software engineers working in different contexts may be less important since code typically functions similarly across different environments.

  Echoing calls to ground approaches to AI harms within the context of deployment,[94] several frameworks recommend collectively documenting the system components that pertain to a given use case rather than separately documenting datasets, models, or systems independent of their intended use case.[95] For example, some researchers have proposed specific documentation methods for affective computing,[96] since these methods pose unique risks related to psychological

manipulation and surveillance and so may merit a customized approach. Moreover, because affective computing applications are likely to be classified as high-risk by the EU AI Act,[97] applications using affective computing are also more likely to be subject to specific documentation requirements.

Customization can also be valuable at the model level. In one study, practitioners argued that the information within a model card should be rearranged to ensure that documentation consumers clearly understand the specific model's purpose and limitations.[98] Another study found that practitioners added technical details to standardized documentation, even when instructed not to, suggesting that they believe standardized formats need flexibility to include model-specific details they find relevant.[99]

At the organizational level, customization might also be necessary. For example, a healthcare startup may require a different documentation approach than a large, well-established financial institution. In some cases, it may be appropriate to customize documentation at multiple levels to respond to both the organization's nature and the type of system it is developing.[100]

- **Standardization can facilitate system discoverability and comparisons.** While customization has many benefits, it also presents challenges. One significant advantage of standardized documentation is that it enables easier comparison of datasets, models, and systems across different projects and organizations. Standard formats facilitate the development of tools that allow practitioners to search for system components that meet particular specifications, improving discoverability and efficiency within an organization's ecosystem. Helping teams discover and use approved AI components that have already undergone relevant risk management processes rather than creating new ones can help ensure risks are not re-introduced and overlooked. Practitioners cite enhanced discoverability as one of the most significant benefits of documentation.[101]

Standardized documentation is also valuable for comparing candidate datasets, models, and systems, and for comparing AI and non-AI approaches.[102] When documentation artifacts vary significantly in the type of information they contain or how this information is presented, these comparisons become more challenging.[103]

- **Standardization can help institutionalize norms of practice.** A more diffuse but critical benefit of standardization is its role in helping the AI community converge on norms of practice and communication.[104] Practitioners sometimes struggle with how to complete documentation or what level of detail to include,[105] leading them to rely heavily on existing examples, which may not always be appropriate in their context. For instance, in one study, practitioners frequently copied documentation from previous projects into their current projects, which could result in practitioners including irrelevant information or omitting relevant details.[106] Standardization can mitigate these issues by providing clear expectations for what information should be included and what users should expect to find in documentation artifacts, thus encouraging broader adoption and more consistent practices.[107]

- **Standardization can facilitate structured risk management.** If documentation producers are instructed to select certain details from structured categories, or certain fields are constrained to particular structured formats, these fields can be used to trigger certain governance actions, such as scheduling a review or requiring that a certain mitigation be applied prior to proceeding. On the other hand, such process automation can mean that concepts and risks are oversimplified, and that processes or mitigations are recommended in cases where they may not be appropriate, while leading other relevant risks to be overlooked.

Tailored vs. general-purpose

When designing documentation strategies, organizations must consider the diverse needs of multiple stakeholder groups. Different stakeholders often require distinct forms of documentation to be effective. To meet these varied needs, organizations might choose to produce tailored documentation for each stakeholder group. However, this approach presents challenges in terms of creating, managing, and maintaining multiple documentation formats. Alternatively, organizations might opt for a single, general-purpose document that all stakeholders can use, though this could compromise the effectiveness of the documentation for specific audiences.

Stakeholder-specific documentation can address the unique needs of different groups of practitioners. While a single-format artifact can reduce the complexity of documentation management, it risks failing to meet the specialized needs of different groups. For instance, documentation designed for data scientists might not be accessible or useful to UX professionals or other non-technical stakeholders.[108] Similarly, documentation aimed at public transparency might lack the depth needed by internal decision-makers handling risk management.

Stakeholder-specific documentation, meanwhile, can address the unique needs of different groups of practitioners. These groups, including users, policymakers, data scientists, engineers, lawyers, and others,[109] often have distinct goals and use different terminologies. These differences can influence the type of documentation produced and the most helpful forms for each group. Documentation tailored to its specific audience is more likely to provide stakeholders with the necessary information for accomplishing their goals. Without a clear target audience, documentation producers might only address what is relevant to their own team,[110] potentially overlooking critical information needed by other teams[111] or presenting it in an unusable format.

Stakeholder-specific documentation can be more accessible to non-technical stakeholders. Tailored documentation may be

especially important when organizations must make technical information accessible to non-technical stakeholders, both within and outside the organization. Within organizations, tailored documentation can ensure that less technical stakeholders, such as compliance or governance professionals, understand key technical considerations at an appropriate level of detail for performing their work. Outside organizations, tailored documentation can ensure that end users understand enough about how systems work to make informed decisions about system use.

While many researchers propose frameworks that aim to make documentation comprehensible to both technical and non-technical practitioners,[112] empirical studies indicate that non-technical audiences often struggle to understand even simplified technical details.[113] For example, a study of non-technical practitioners using documentation to assess model risks found that a significant proportion failed to correctly identify the model's basic purpose.[114] Even when documentation presents technical details in simpler terms, non-technical practitioners may still struggle to grasp technical information, such as accuracy metrics.[115] This suggests that even when technical information is presented in simpler terms, non-technical stakeholders may still face challenges in interpreting technical information necessary for risk assessment when it is not specifically tailored to their needs.

Stakeholder-specific documentation can better leverage practitioners' areas of expertise. Tailoring documentation allows stakeholders to focus on the most relevant information for their roles, thereby enhancing the organization's ability to identify and manage risks. For instance, in developing a documentation framework for healthcare data, machine learning experts prioritized the dataset's composition since that is important for model training and performance. Conversely, healthcare experts concentrated on details, such as how medical diagnoses were assigned (e.g., by physicians or lab tests), data collection sites, and the calibration of diagnostic equipment, as these aspects are crucial for interpreting the dataset's relevance and limitations.[116] By emphasizing information pertinent to their expertise, stakeholders can collectively conduct more thorough analyses and potentially uncover issues that others might miss.[117]

Stakeholder-specific documentation may be more actionable. One common issue with general-purpose documentation is that it may not clearly guide stakeholders on how to use the information provided.[118] This issue is particularly significant when consumers fail to connect documentation with the ethical impacts of their work,[119] indicating a need for prescriptive guidance to manage AI risks effectively. The actions practitioners take in response to documentation often depend on their specific roles for two reasons. First, when documentation is aligned with the purpose of their role, practitioners are better able to understand the actions necessary to manage risk. For example, information about how well a system meets user needs may be more actionable for a UX professional than a machine learning engineer. Second, practitioners typically have the authority to act in specific ways within organizations. For instance, compliance teams cannot implement changes to the production codebase, and data scientists are not responsible for conducting legal reviews. Tailored documentation that aligns with each group's roles and responsibilities can guide them toward actions they are empowered to take.

Single-format documentation can help reduce confusion that might arise from multiple forms of documentation. The creation and maintenance of multiple documentation formats come with significant challenges. Single-format documentation can help reduce confusion and avoid the fragmentation that occurs when information is scattered across different formats, such as README files, wikis, and slide decks.[120] Even with centralized repositories that link different versions, managing multiple formats can still lead to confusion about the existence and authority of these artifacts and make it difficult for practitioners to access the information they need and develop a cohesive understanding.

Single-format documentation is often easier to produce and maintain. If practitioners are required to create multiple forms of documentation to meet the needs of different stakeholders, organizations must ensure sufficient time is allocated for this process. Without adequate time, practitioners may rush, leading to low-quality documentation. In cases where time constraints are a

concern, it may be more effective to focus on producing a single, high-quality document that is regularly updated. At a minimum, all documentation should clearly indicate its last update to help users gauge its relevance and accuracy. Single-format documentation may be particularly efficient to create when it incorporates some level of standardization.

Interactive vs. static

Producing multiple forms of documentation is one way to support a variety of systems, stakeholders, and goals; however, these multiple forms of documentation create challenges for organizations to manage. Interactive documentation offers an alternative solution by enabling stakeholders to access the specific information they need while maintaining overall usability and comprehension.

Interactive documentation can accommodate multiple stakeholders with a single artifact. Since stakeholders often require different levels of detail depending on their expertise and needs, interactive documentation allows users to selectively access the information they need and reduce or eliminate the need to create multiple documentation versions. For instance, interactive system diagrams might allow practitioners to click on specific components to access technical details,[121] while expandable sections can provide additional context as necessary.[122]

Interactive documentation can enhance comprehension for non-technical stakeholders. Empirical research indicates that such documentation can significantly improve understanding, particularly among those without technical backgrounds.[123] Interactive features, like Hugging Face's model inference API[124] or OpenAI's developer playground,[125] allow practitioners to engage with models and systems directly. This hands-on experience helps foster intuitive understanding of how system inputs and outputs are connected,[126] which is crucial for designing products that utilize the system's capabilities and for assessing potential risks. Interactive access is especially valuable when exploring pre-trained, general-purpose models, whose functionality can vary significantly in different implementations.[127]

Static documentation presents consistent information in a consistent format. Static documentation provides uniform information to all stakeholders, fostering a common understanding that could enhance communication, collaboration, and decision-making. Also, without good information architecture, interactive documentation can overwhelm documentation consumers.[128] Moreover, unlike interactive documentation, which requires an additional engineering effort, static documentation formats are likely more manageable for organizations to support.

Comprehensive vs. concise

Organizations must determine what level of detail to include in documentation artifacts. Whereas exhaustive documentation might ensure practitioners have access to all the necessary information, concise documentation may improve ease of production and use.

Concise documentation can help stakeholders focus on the most relevant details for the intended audience. For documentation aimed at downstream consumers, including internal risk management professionals, it is essential to provide enough detail to support decision-making without overwhelming them. Information overload can lead to selective attention, where decision-makers may focus on certain details while neglecting others, potentially degrading the quality of their decisions,[129] especially under time pressure.[130] For example, faced with extensive and complex documentation on AI systems, consumers might miss critical information or become frustrated and abandon the effort entirely. As one participant in a research study noted, "I would lose patience after 30-40 seconds if I have to put a lot of effort into finding what I'm looking for."[131] Concise documentation that highlights the most important information for a given stakeholder group can help internal decision-makers manage risk more effectively by ensuring they focus on critical details.

Concise documentation may be easier to produce and maintain, increasing accuracy and encouraging more frequent use. Exhaustive documentation may be burdensome to create and keep updated, particularly if the process is manual. For instance, in the initial case studies of an extensive data documentation framework,

practitioners reported that completing the documentation took two to three hours, excluding the time needed to gather the required information.[132] While a few hours might be manageable for a single project, this time commitment becomes overwhelming when scaled across an entire organization, potentially involving hundreds or thousands of datasets, models, or systems. In such scenarios, spreading effort thinly across all systems might detract from the focus on high-risk or high-impact components, undermining critical risk management efforts.

Exhaustive documentation provides a more comprehensive overview, accommodating a wider range of applications. While concise documentation has its advantages, it may omit details that, although not relevant to most uses, are crucial for specific applications. Exhaustive documentation provides a more comprehensive overview that can accommodate a broader range of applications. For example, consider a dataset containing patient information and healthcare outcomes over a year. If a model uses this dataset to predict patients' health outcomes over time, the model's accuracy depends on the validity and reliability of the data's timestamps. Thus, documenting how these timestamps were applied is essential for practitioners intending to use the dataset for this purpose. On the other hand, if the dataset is used for clustering patients into broad categories, temporal information becomes less critical. Therefore, the project's purpose and the specific stakeholders involved should shape the level of detail included in the documentation.

Manual vs. automated

Documentation encompasses two main types of information: data that can be extracted automatically from the system and information that requires human input. Information that could potentially be derived directly from source code or generated through scripts interacting with system components includes elements such as data quantity, distribution, statistical properties, model types and parameters, system flow, and software library

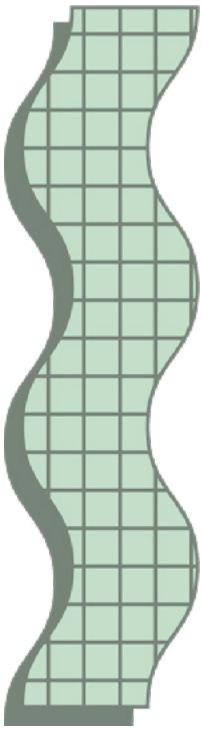
usage. In contrast, information related to human decision-making processes and organizational context requires manual input. This includes the motivations for developing specific datasets or models, criteria for selecting data sources, reasoning behind chosen methodologies, outcomes of compliance reviews, contact information for responsible parties, decommissioning procedures, and ethical considerations addressed during development.

Manual documentation may encourage deeper ethical reflection. One of the potential purposes of both producing and using documentation within an organization is to prompt developers to consider the ethical implications of their systems. However, the degree of automation in the documentation process can significantly influence how thoroughly practitioners engage with these considerations. Fully automated documentation requires minimal direct involvement, which may not encourage the level of critical reflection necessary to address complex ethical issues. As a result, some experts advocate for completing documentation manually, even when automation is feasible, to ensure that practitioners engage deeply with the material.[133]

Automated documentation can enable more frequent updates and minimize human error. Manually creating documentation is time-consuming and prone to errors, particularly if it is not well integrated with practitioners' existing tools and workflows.[134] Practitioners have noted that automation can reduce the time required to complete documentation and increase the likelihood that it is kept up-to-date.[135] Automation also reduces the burden on practitioners to recall critical information after the fact, which can be particularly valuable in complex, fast-paced environments.[136]

Organizations should also be aware that practitioners facing multiple competing priorities may circumvent manual processes by implementing their own forms of unaccountable automation.[137] Well-designed automation of some portions of documentation could reduce this risk.

04 Recommendations



Because organizations have unique requirements and operate in dramatically different contexts, there is no one-size-fits-all approach to documentation processes or a universal set of information that organizations should capture for all datasets, models, or systems. However, existing theories and evidence suggest several key considerations organizations should take into account when designing and implementing their documentation strategies.

- **Identify the primary objectives for documentation and design approaches to align with these goals.** Interventions are usually most successful when tailored to specific aims. Organizations should intentionally identify, prioritize, and articulate their documentation goals, understanding that achieving all goals simultaneously and equally may not be feasible. Organizations should not default to using existing templates without reflecting on whether those templates are sufficient to support their identified documentation goals. Instead, organizations should determine their documentation processes, what information they document, and the format they use based on these objectives. Organizations should also establish key performance indicators for their documentation goals and assess progress regularly, allowing for continuous improvement. Policymakers should similarly consider what documentation goals they aim to facilitate and craft requirements to align with those goals.

- **Evaluate operational resources and constraints and incentivize the production of robust documentation within that context.** Organizations should realistically evaluate their available resources for supporting documentation, including time, personnel, and expertise for creating and maintaining documentation. Attempting to implement comprehensive documentation without the necessary organizational capacity and incentives can lead to unexpected and harmful failures. For instance, if practitioners feel rushed and view documentation as a secondary task, they might resort to inappropriate practices such as copying and pasting from previous projects, leaving sections incomplete, inputting incorrect information instead of finding the correct data, or creating informal, unaccountable methods for producing documentation. It is generally better for documentation to be less exhaustive but more accurate and reliable than to attempt to cover more ground but do so inadequately. If organizations can assess their resources and strengths and determine the most effective areas for documentation, they can establish basic approaches to support their goals and build on them over time. Practitioners should understand what is expected of them in terms of documentation, and opportunities to cut corners should be minimized.

There is no one-size-fits-all approach to documentation processes or a universal set of information that organizations should capture for all datasets, models, or systems.

- **Identify critical stakeholders and their documentation needs.** When designing and implementing AI documentation approaches, organizations should start by identifying the stakeholders who will most frequently and meaningfully interact with documentation processes and artifacts. Organizations should consider the needs of stakeholders who will consume documentation artifacts and the expertise and capacity of those who will produce the documentation. Suppose, for example, that organizations design their approach to produce documentation artifacts that are maximally informative to downstream consumers but that cannot feasibly be produced by upstream practitioners. In that case, their approach is unlikely to be successful when implemented, and so would require adding staff to documentation producer teams who can effectively meet the needs of documentation consumers. Organizations should also consider documentation stakeholders' technical and

non-technical knowledge to ensure that their documentation approaches balance the nature and level of detail with accessibility.

- **Institutionalize responsible AI and risk management practices beyond documentation.** Ethical deliberation is a skill that practitioners can develop over time with intentional practice,[138] but organizations need to implement additional interventions beyond documentation to help cultivate this skill.[139] Even when practitioners engage in documentation processes, they may overlook critical risks or fail to recommend appropriate mitigations without a deeper understanding of how AI can harm diverse individuals and communities.[140] Therefore, organizations should strive to embed responsible AI and risk management efforts throughout their AI development practices. When responsible AI is integrated into the organization's culture and broader practices, documentation processes that prompt practitioners to reflect on risks are more likely to be informed by a comprehensive understanding necessary for effective risk management.

  Conversely, relying solely on documentation to encourage critical reflection is unlikely to be sufficient and may even lead to negative outcomes. Documentation of potential ethical harms will likely be more effective when it focuses on concrete system properties, such as data containing personally identifiable information or existing societal biases, rather than on speculative concerns, which can be subjective and heavily influenced by the expertise and perspectives of the documentation producers. Alternatively, if certain ethical issues are of particular concern for organizations – for example, fairness issues in AI applications where inequitable outcomes would have legal implications – organizations may want to specifically advise producers on how to identify and communicate about these issues. This may have the added benefit of improving standardization and institutionalizing norms and values.

- **Establish a plan for storage, organization, and maintenance of documentation artifacts.** A robust documentation infrastructure can increase the likelihood that stakeholders can easily access the information they need and minimize confusion when multiple forms of documentation exist for different stakeholders or levels of detail. Good infrastructure can help practitioners develop tools for searching and comparing documentation artifacts, ensuring that development and risk management efforts are efficient and effective.

Organizations should strive to embed responsible AI and risk management efforts throughout their AI development practices.

Lastly, documentation tools are most likely effective when incorporated within practitioners' existing workflows. For example, rather than integrating directly with practitioners' software engineering interfaces, some documentation tools require practitioners to copy information from their code into external templates or forms. In these cases, practitioners are more likely to make errors and engage minimally with the ethical hazards of their work.[141] Sound documentation infrastructure that works with practitioners' existing workflows is more likely to realize success than isolated, ad-hoc tools.

- **Use standardization as the default approach, and intentionally deviate from it when necessary.** Standardization offers many advantages. It allows practitioners to compare documentation artifacts, facilitates search and discoverability of information, establishes consistent expectations for documentation producers and consumers, and can help institutionalize norms of practice. For these reasons, some level of standardization in documentation processes and artifacts is likely to be helpful for most organizations for most AI systems.

  To balance standardization and customization, organizations might develop customizable documentation templates or interchangeable modules.[142] Templates could, for example, allow for customization of subsections while maintaining consistent content and structure within those subsections. For instance, companies may include a subsection on the

labeling process for AI systems based on supervised learning but exclude it for unsupervised AI systems since they do not require labels. Where companies choose to embrace a greater degree of customization, they should ensure that the benefits of this customization outweigh the benefits of maintaining standardization.

- **Leverage automation for information extraction and incorporate manual processes for scoped critical synthesis.** Some helpful information for documentation, such as dataset composition, statistical properties, model training methods, parameters, and software library or third-party API usage, can be extracted from AI systems using automated procedures, contributing to consistent, high-fidelity documentation of these system elements. Yet the aspects of systems that are easiest to document automatically are not always the most critical. For example, by manually documenting the motivation for a dataset's creation, practitioners can enable downstream stakeholders to better assess whether this motivation aligns with their own prospective uses. This information could be more critical for decision making than additional plots of the distribution of the dataset's features. As a result, organizations should think carefully about what automatically extracted information and manually generated details to include. Organizations should bear in mind that errors in automated documentation can have a greater negative impact than manual errors, as they can persist through each update. Thus, organizations should carefully review any automated pipelines for producing documentation.

- **Establish a quality assurance or review process for documentation.** Quality assurance and review ensures that all critical information is accurately captured and clearly conveyed, safeguarding against potential oversights, errors, and misrepresentations by those creating the documentation. When documentation producers and consumers have different expertise, documentation review processes can ensure that information is accessible and actionable. The review might also help organizations identify opportunities for improving their documentation approach.

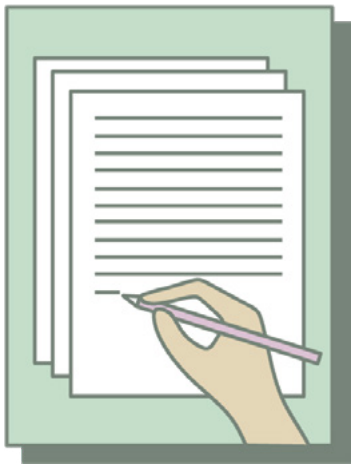
- **Leverage empirical evidence and feedback for continuous improvement.** Organizations should evaluate potential documentation approaches via pilot projects and periodically collect empirical data on the efficacy of their documentation practices in meeting the needs of relevant stakeholders. Organizations should plan to assess the usability of documentation frameworks and their effectiveness in achieving the organization's goals,[143] and adapt their documentation strategies by conducting regular assessments as their needs or technology evolves. Given the limited empirical data on the effectiveness of documentation approaches,[144] organizations can also contribute to the broader AI community by publicly sharing their findings on what works best in different contexts. Organizations should recognize that empirical evidence and key performance indicators can be qualitative and quantitative. Whereas quantitative evidence can help organizations assess general trends at scale, qualitative evidence can help them understand unique issues in greater depth.

- **Consider interactive documentation interfaces for general-purpose models in particular.** Documenting the characteristics of general-purpose models and their potential risks is particularly challenging due to their wide range of possible applications. When downstream practitioners cannot directly communicate with upstream model developers, interpreting the general documentation in the context of specific applications can be difficult. Several studies have shown that interfaces allowing practitioners to observe system outputs in response to given inputs can enhance their understanding of model capabilities and risks.[145] Consequently, interactive interfaces can serve as a valuable complement to traditional, static documentation.

- **Focus initially on core information needs for documentation and evaluate potential gaps over time.** When designing and implementing new documentation strategies, organizations should start by identifying the core pieces of information that will meet the needs of key stakeholders. (In other words, they should start by designing a concise version of documentation.) They should plan to consult with stakeholders who use this documentation to determine whether any additional information is necessary to perform their essential duties and whether any of the initial information provided in documentation artifacts is unnecessary, and incorporate learnings from empirical research to augment documentation as needed. Organizations can use this agile approach to manage the information in documentation artifacts, ensuring that the time spent producing and using documentation is maximally efficient and effective at serving stakeholder requirements.

05 Conclusion

This report underscores the critical role of documentation in enhancing AI governance, emphasizing that effective documentation practices are essential for managing AI risks and fostering responsible system development. To support robust governance, organizations must tailor their documentation processes and artifacts to meet the specific needs and constraints of their stakeholders. Establishing clear success criteria—such as accuracy, comprehensiveness, and usability—and regularly assessing progress against these goals is crucial for maintaining the effectiveness of documentation strategies.



Although this report primarily focuses on documentation that can enhance internal governance, documentation processes and artifacts can also form the foundation for transparency efforts directed at external stakeholders. By cultivating strong internal documentation practices and continually evaluating their success, organizations can build the necessary infrastructure to create detailed, actionable records of system development and risk management. These records are necessary for communicating effectively with external audiences, such as regulators and the public.

However, documentation designed for internal use may not seamlessly translate to external contexts. The differences in expertise, needs, and objectives between internal and external stakeholders require careful consideration. Organizations must adapt their documentation processes and artifacts to bridge these gaps, ensuring that they are both accessible and informative to all relevant parties. By doing so, organizations can enhance their governance efforts and contribute to a more transparent and accountable AI ecosystem.

06

Appendix

Methods

Publication Sampling

We employed a purposeful sampling approach[146] to identify proposed documentation frameworks and relevant empirical studies. Given our specific interest in how AI documentation can support governance, our use of a purposeful sampling approach allowed us to focus on proposed frameworks and empirical studies that are directly relevant to our research questions. It also afforded us the flexibility to iteratively refine our sample of papers based on emerging findings.

We identified an initial seed sample of publications from a systematic review focused on AI documentation methods relevant to EU regulation.[147] We cross-referenced the initial sample against the references used by the Partnership on AI in developing their ABOUT ML framework,[148] which seeks to provide technology organizations with guidance on what to document about their AI systems. We chose these initial sources because they allowed us to focus on publications that establish best practices for industry documentation and address policy initiatives aimed at institutionalizing these practices. We reviewed the cited references within each of these initial works for additional proposed documentation frameworks and empirical studies related to documentation. We also consulted with academic researchers and industry practitioners with experience producing and using documentation to identify additional frameworks and studies. We excluded papers that centered on tools for implementing documentation (e.g., automated documentation code libraries) or that were not specifically focused on AI system documentation, such as fairness checklists.

This process yielded a sample of 37 proposed frameworks, eight of which included empirical evidence related to their implementation. Because of the relative dearth of empirical findings within publications proposing approaches, we further expanded our sample by searching for references within these 37 approaches that empirically evaluated documentation methods. We also conducted searches in the archives of the ACM Conference on Human Factors in Computing Systems[149] and the ACM Conference on Computer Supported Cooperative Work[150] using terms such as "datasheet," "model card," "AI documentation," "model documentation," "data card," and "data documentation." This allowed us to identify an additional 13 publications with empirical findings. Recognizing the importance of industry perspectives, we included non-peer-reviewed papers, such as technical whitepapers, acknowledging that they often provide insights not found in academic literature. While our sampling approach was not exhaustive, we reached theoretical saturation after analyzing the selected publications, as no new themes emerged. Therefore, we concluded our sampling at this point.

Table 1. List of Publications

	Author	Year	Framework Type	Evaluation Type
1	Adkins et al.,	2022	Method, process, or task	Feasibility analysis
2	Arnold et al.,	2019	System	Feasibility analysis
3	Baracaldo et al.,	2022	System	Feasibility analysis
4	Bender & Friedman	2018	Data	Feasibility analysis
5	Bhat et al.,	2023	Evaluation only	Practitioner lab study
6	Blasch et al.,	2020	System	Feasibility analysis
7	Boyd	2021	Evaluation only	Practitioner lab study
8	Brajovic et al.,	2023	System	Feasibility analysis
9	Chang & Custis	2022	Evaluation only	Practitioner real-world study
10	Chmielinski et al.,	2020	Data	Feasibility analysis
11	Chmielinski et al.,	2024	Method, process, or task	None
12	Crisan et al.,	2022	Model	Feasibility analysis, Practitioner lab study
13	Díaz et al.,	2023	Data	Feasibility analysis
14	Gebru et al.,	2021	Data	Feasibility analysis
15	Geiger et al.,	2020	Evaluation only	Artifact study
16	Gilbert et al.,	2023	System	Feasibility analysis
17	Heger et al.,	2022	Evaluation only	Practitioner real-world study

	Author	Year	Framework Type	Evaluation Type
18	Hind et al.,	2019	Evaluation only	Practitioner real-world study
19	Holland et al.,	2018	Data	Feasibility analysis
20	Hupont & Gomez	2022	Method, process, or task	Feasibility analysis
21	Liang et al.,	2024	Evaluation only	Artifact study
22	Liao et al.,	2023	Evaluation only	Practitioner lab study
23	Marone & Van Durme	2023	Data	Feasibility analysis
24	McMillan-Major et al.	2021	Data, Model	Feasibility analysis
25	McMillan-Major, Bender, & Friedman	2024	Data	Practitioner lab study
26	Miceli et al.	2021	Evaluation only	Practitioner real-world study
27	Mitchell et al.,	2019	Model	Feasibility analysis
28	Mohammad	2022	Method, process, or task	Feasibility analysis
29	Moore, Llao, & Subramonyam	2023	Evaluation only	Practitioner lab study
30	Nunes et al.,	2022	Evaluation only	Practitioner lab study
31	Papakyriakopoulos et al.,	2023	Data	Feasibility analysis
32	Pepe et al.,	2024	Evaluation only	Artifact study
33	Procope et al.,	2022	System	Feasibility analysis

	Author	Year	Framework Type	Evaluation Type
34	Pushkarna et al.,	2022	Data	Practitioner real-world study
35	Raji & Yang	2020	System	None
36	Reid & Williams	2023	Evaluation only	Practitioner real-world study Artifact study
37	Richards et al.,	2020	System	Feasibility analysis
38	Roman et al.	2023	Data	Feasibility analysis
39	Rostamzadeh et al.,	2022	Data	Practitioner lab study
40	Shen et al.,	2022	Model	Practitioner lab study
41	Shimorina & Belz	2021	Method, process, or task	None
42	Soh	2021	Data	None
43	Sokol & Flach	2020	Method, process, or task	Feasibility analysis
44	Srinivasan et al.,	2021	Data	Feasibility analysis
45	Stoyanovich & Howe	2019	Data, Model	Feasibility analysis
46	Subramaniam et al.,	2023	Data	Practitioner lab study
47	Sun et al.,	2019	Data	Feasibility analysis
48	Tagliabue et al.,	2021	Method, process, or task	Feasibility analysis
49	Yang et al.,	2018	Method, process, or task	Feasibility analysis
50	Zheng et al.,	2022	Data	Practitioner lab study

Data Analysis

Our data analysis followed an abductive approach,[151] iterating between inductively derived codes, as in grounded theory,[152] and codes based on the theoretical motivations of our research. The first author conducted the initial coding by applying descriptive codes aligned with our research goals, such as identifying how AI documentation enhances internal governance and the challenges of using documentation as a governance tool. After discussing the results with the second author, we refined these codes into higher-order categories during the axial coding phase. For instance, codes like "identify intended purpose" and "specify out-of-scope uses" were grouped under the category "inform downstream documentation consumers." In the final thematic coding phase, we grouped axial codes into broader theoretical mechanisms through which AI documentation could support robust governance, the challenges in achieving these impacts, and the design trade-offs faced by organizations.

Publication Sample

For each framework, we classified whether it primarily focused on data, models, systems, or methods/tasks/processes. When a framework did not fit neatly into one of these categories, we assigned it to the most appropriate category or categories based on its primary focus.

We also categorized the type of evaluation each framework or independent empirical research study employed. Frameworks employing a "feasibility analysis" are those in which the framework's authors or another group applied the framework to create documentation for a hypothetical or actual dataset, model, system, or method. This type of analysis demonstrates that the framework could theoretically be used for its intended purpose but does not

involve empirical evaluation with practitioners in research or real-world settings. If a study developed a documentation artifact for the purpose of an empirical study, we classified this as part of the empirical study rather than as a feasibility analysis, as empirical studies offer a more rigorous evaluation.

We classify publications as employing a "practitioner lab study" when the evaluation involved practitioners within a controlled research setting. We classify "practitioner real-world studies" as those examining practitioner methods and practices within their real-world work environments. Both lab and real-world studies have unique strengths, and neither is inherently more rigorous or useful than the other. We define "artifact studies" as studies of publicly available documentation artifacts such as Github repository documentation or Hugging Face model cards.

In some instances, framework authors mentioned consulting relevant stakeholders during the design or refinement of their framework. However, if these consultations are only briefly mentioned or not elaborated upon, we do not classify the work as including a practitioner study.

07 Endnotes

1 Liao and Vaughan, "AI Transparency in the Age of LLMs."

2 Touvron et al., "Llama 2."

3 OpenAI, "GPT-4V(Ision) System Card."

4 Gebru et al., "Datasheets for Datasets."

5 Mitchell et al., "Model Cards for Model Reporting."

6 Gebru et al., "Datasheets for Datasets."

7 Bender and Friedman, "Data Statements for Natural Language Processing"; Chmielinski et al., "The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence"; Díaz et al., "SoUnD Framework"; Gebru et al., "Datasheets for Datasets"; Marone and Van Durme, "Data Portraits"; McMillan-Major et al., "Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation," 2021; Pushkarna, Zaldivar, and Kjartansson, "Data Cards"; Roman et al., "Open Datasheets"; Rostamzadeh et al., "Healthsheet"; Soh, "Building Legal Datasets"; Srinivasan et al., "Artsheets for Art Datasets"; Stoyanovich and Howe, "Nutritional Labels for Data and Models"; Subramaniam et al., "Comprehensive and Comprehensible Data Catalogs"; Sun et al., "MithraLabel"; Zheng et al., "Network Report"; Papakyriakopoulos et al., "Augmented Datasheets for Speech Datasets and Ethical Decision-Making."

8 Reid and Williams, "Right the Docs"; Srinivasan et al., "Artsheets for Art Datasets."

9 McMillan-Major et al., "Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation," 2021; Stoyanovich and Howe, "Nutritional Labels for Data and Models"; Crisan et al., "Interactive Model Cards," 2022; Mitchell et al., "Model Cards for Model Reporting"; Shen et al., "The Model Card Authoring Toolkit"; Gilbert et al., "Reward Reports for Reinforcement Learning."

10 Mitchell et al., "Model Cards for Model Reporting."

11 Zhang et al., "Dissecting the Runtime Performance of the Training, Fine-Tuning, and Inference of Large Language Models."

12 Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback."

13 Qi et al., "Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!"

14 Shaikh et al., "On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning."

15 Andrew Smart et al., "What Is Sociotechnical AI Safety? What Do We Want It To Be? A FAccT Community Workshop."

16 Arnold et al., "FactSheets"; Baracaldo et al., "Towards an Accountable and Reproducible Federated Learning"; Brajovic et al., "Model Reporting for Certifiable AI"; Green et al., "System Cards, a New Resource for Understanding How AI Systems Work"; Procope et al., "System-Level Transparency of Machine Learning"; Yang et al., "A Nutritional Label for Rankings"; Blasch, Sung, and Nguyen, "Multisource AI Scorecard Table for System Evaluation."

17 Shimorina and Belz, "The Human Evaluation Datasheet 1.0."

18 Raji et al., "Closing the AI Accountability Gap."

19 Belz et al., "Missing Information, Unresponsive Authors, Experimental Flaws"; Reid and Williams, "Right the Docs."

20 Adkins et al., "Prescriptive and Descriptive Approaches to Machine-Learning Transparency"; Baracaldo et al., "Towards an Accountable and Reproducible Federated Learning"; Sokol and Flach, "Explainability Fact Sheets"; Tagliabue et al., "DAG Card Is the New Model Card."

21 Adkins et al., "Prescriptive and Descriptive Approaches to Machine-Learning Transparency."

22 Sokol and Flach, "Explainability Fact Sheets."

23 Hupont and Gomez, "Documenting Use Cases in the Affective Computing Domain Using Unified Modeling Language"; Mohammad, "Ethics Sheets for AI Tasks."

24 Zheng et al., "Network Report"; Rostamzadeh et al., "Healthsheet"; Reid and Williams, "Right the Docs."

25 Boyd, "Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data."

26 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata"; Reid and Williams, "Right the Docs"; Shen et al., "The Model Card Authoring Toolkit"; Zheng et al., "Network Report."

27 Shen et al., "The Model Card Authoring Toolkit."

28 Holland et al., "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards."

29 Mitchell et al., "Model Cards for Model Reporting."

30 Roman et al., "Open Datasheets."

31 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

32 Heger et al.

33 Miceli et al., "Documenting Computer Vision Datasets"; Chang and Custis, "Understanding Implementation Challenges in Machine Learning Documentation."

34 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata"; Chang and Custis, "Understanding Implementation Challenges in Machine Learning Documentation."

35 "Improving Documentation for AI Governance;" Ahlawat, Winecoff, and Mayer, "Minimum Viable Ethics."

36 Liao et al., "Designerly Understanding"; Miceli et al., "Documenting Computer Vision Datasets"; Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata"; Chang and Custis, "Understanding Implementation Challenges in Machine Learning Documentation."

37 Pushkarna, Zaldivar, and Kjartansson, "Data Cards"; Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

38 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

39 Hind et al., "Experiences with Improving the Transparency of AI Models and Services."

40 Hind et al.

41 Reid and Williams, "Right the Docs."

42 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

43 Geiger et al., "Garbage in, Garbage Out?"; Pepe et al., "How Do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study"; Liang et al., "What's Documented in AI?"

44 Liang et al., "What's Documented in AI?"

45 Pepe et al., "How Do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study."

46 Chmielinski et al., "The CLeAR Documentation Framework for AI Transparency Recommendations for Practitioners and Context for Policymakers."

47 Pushkarna, Zaldivar, and Kjartansson, "Data Cards"; Gilbert et al., "Reward Reports for Reinforcement Learning"; Raji and Yang, "ABOUT ML"; Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata"; Crisan et al., "Interactive Model Cards," 2022; Srinivasan et al.,

"Artsheets for Art Datasets"; Mohammad, "Ethics Sheets for AI Tasks"; Richards et al., "A Methodology for Creating AI FactSheets"; Shen et al., "The Model Card Authoring Toolkit."

48 Liao et al., "Designerly Understanding."

49 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

50 Pushkarna, Zaldivar, and Kjartansson, "Data Cards."

51 Shen et al., "The Model Card Authoring Toolkit."

52 Chang and Custis, "Understanding Implementation Challenges in Machine Learning Documentation."

53 Liao et al., "Designerly Understanding."

54 Vaswani et al., "Attention Is All You Need."

55 Crisan et al., "Interactive Model Cards," 2022; Liao et al., "Designerly Understanding"; Moore, Liao, and Subramonyam, "fAIlureNotes."

56 "Improving Documentation for AI Governance."

57 Boyd, "Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data."

58 Gebru et al., "Datasheets for Datasets."

59 Arnold et al., "FactSheets."

60 Brajovic et al., "Model Reporting for Certifiable AI"; McMillan-Major et al., "Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation," 2021; Shen et al., "The Model Card Authoring Toolkit."

61 Bender and Friedman, "Data Statements for Natural Language Processing."

62 Friedman, Kahn Jr., and Borning, "Value Sensitive Design and Information Systems."

63 Shen et al., "The Model Card Authoring Toolkit."

64 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

65 Madaio et al., "Learning about Responsible AI On-The-Job."

66 Hind et al., "Experiences with Improving the Transparency of AI Models and Services."

67 McLennan et al., "Embedded Ethics."

68 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata"; Madaio et al., "Learning about Responsible AI On-The-Job."

69 Pepe et al., "How Do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study"; Liang et al., "What's Documented in AI?"

70 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata.

71 Heger et al.; Nunes et al., "Using Model Cards for Ethical Reflection"; Hind et al., "Experiences with Improving the Transparency of AI Models and Services"; Crisan et al., "Interactive Model Cards," 2022.

72 Crisan et al., "Interactive Model Cards," 2022.

73 Nunes et al., "Using Model Cards for Ethical Reflection."

74 "Improving Documentation for AI Governance."

75 Miceli et al., "Documenting Computer Vision Datasets."

76 Framework authors posit that documentation could improve the rigor with which practitioners develop AI systems.

77 Chmielinski et al., "The CLeAR Documentation Framework for AI Transparency Recommendations for Practitioners and Context for Policymakers"; Bender and Friedman, "Data Statements for Natural Language Processing"; Holland et al., "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards."

78 Hind et al., "Experiences with Improving the Transparency of AI Models and Services"; Adkins et al., "Prescriptive and Descriptive Approaches to Machine-Learning Transparency"; Baracaldo et al., "Towards an Accountable and Reproducible Federated Learning."

79 Adkins et al., "Prescriptive and Descriptive Approaches to Machine-Learning Transparency."

80 Winecoff and Watkins, "Artificial Concepts of Artificial Intelligence."

81 Shilton, "Values Levers."

82 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

83 Rostamzadeh et al., "Healthsheet."

84 Roman et al., "Open Datasheets."

85 Miceli et al., "Documenting Computer Vision Datasets"; Chang and Custis,

"Understanding Implementation Challenges in Machine Learning Documentation."

86 Mitchell et al., "Model Cards for Model Reporting."

87 Holland et al., "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards."

88 Adkins et al., "Prescriptive and Descriptive Approaches to Machine-Learning Transparency"; Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata"; Chang and Custis, "Understanding Implementation Challenges in Machine Learning Documentation."

89 Brajovic et al., "Model Reporting for Certifiable AI"; Crisan et al., "Interactive Model Cards," 2022; Hupont and Gomez, "Documenting Use Cases in the Affective Computing Domain Using Unified Modeling Language"; Miceli et al., "Documenting Computer Vision Datasets."

90 Mökander et al., "Auditing Large Language Models."

91 Clavell, "Checklist for AI Auditing."

92 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata"; Bhat et al., "Aspirations and Practice of Model Documentation."

93 Rostamzadeh et al., "Healthsheet."

94 Narayanan and Kapoor, "AI Safety Is Not a Model Property"; Hutchinson et al., "Evaluation Gaps in Machine Learning Practice"; Nicholas, "Grounding AI Policy: Towards Researcher Access to AI Usage Data."

95 Rostamzadeh et al., "Healthsheet"; Srinivasan et al., "Artsheets for Art Datasets"; Chmielinski et al., "The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence"; Mohammad, "Ethics Sheets for AI Tasks"; Hupont and Gomez, "Documenting Use Cases in the Affective Computing Domain Using Unified Modeling Language."

96 Hupont and Gomez, "Documenting Use Cases in the Affective Computing Domain Using Unified Modeling Language."

97 Hupont and Gomez.

98 Crisan et al., "Interactive Model Cards," 2022.

99 Bhat et al., "Aspirations and Practice of Model Documentation."

100 Chmielinski et al., "The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence"; Gebru et al., "Datasheets for Datasets"; Holland et al., "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards"; Roman et al., "Open Datasheets"; Stoyanovich and Howe, "Nutritional

Labels for Data and Models"; Richards et al., "A Methodology for Creating AI FactSheets."

101 Reid and Williams, "Right the Docs"; Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

102 Shen et al., "The Model Card Authoring Toolkit"; Mitchell et al., "Model Cards for Model Reporting"; Stoyanovich and Howe, "Nutritional Labels for Data and Models"; Zheng et al., "Network Report"; Gilbert et al., "Reward Reports for Reinforcement Learning"; Hind et al., "Experiences with Improving the Transparency of AI Models and Services"; Reid and Williams, "Right the Docs"; Chmielinski et al., "The CLeAR Documentation Framework for AI Transparency Recommendations for Practitioners and Context for Policymakers."

103 Reid and Williams, "Right the Docs."

104 Arnold et al., "FactSheets"; McMillan-Major et al., "Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation," 2021; McMillan-Major, Bender, and Friedman, "Data Statements."

105 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata"; Chang and Custis, "Understanding Implementation Challenges in Machine Learning Documentation"; Pushkarna, Zaldivar, and Kjartansson, "Data Cards"; Miceli et al., "Documenting Computer Vision Datasets."

106 Pushkarna, Zaldivar, and Kjartansson, "Data Cards."

107 McMillan-Major et al., "Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation," 2021; Roman et al., "Open Datasheets"; Pushkarna, Zaldivar, and Kjartansson, "Data Cards."

108 Liao et al., "Designerly Understanding."

109 Micheli et al., "The Landscape of Data and AI Documentation Approaches in the European Policy Context."

110 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

111 Hind et al., "Experiences with Improving the Transparency of AI Models and Services."

112 McMillan-Major, Bender, and Friedman, "Data Statements"; McMillan-Major et al., "Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation," 2021; Holland et al., "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards"; Mohammad, "Ethics Sheets for AI Tasks"; Roman et al., "Open Datasheets"; Richards et al., "A Methodology for Creating AI FactSheets"; Tagliabue et al., "DAG Card Is the New Model Card"; Sokol and Flach, "Explainability Fact Sheets"; Pushkarna, Zaldivar, and Kjartansson, "Data Cards."

113 Crisan et al., "Interactive Model Cards," 2022; Liao et al., "Designerly Understanding"; Shen et al., "The Model Card Authoring Toolkit"; Miceli et al., "Documenting Computer Vision Datasets."

114 Crisan et al., "Interactive Model Cards," 2022.

115 Shen et al., "The Model Card Authoring Toolkit."

116 Rostamzadeh et al., "Healthsheet."

117 Rostamzadeh et al.; Pushkarna, Zaldivar, and Kjartansson, "Data Cards."

118 Adkins et al., "Prescriptive and Descriptive Approaches to Machine-Learning Transparency"; Crisan et al., "Interactive Model Cards," 2022; Stoyanovich and Howe, "Nutritional Labels for Data and Models."

119 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

120 Heger et al.

121 Procope et al., "System-Level Transparency of Machine Learning."

122 Crisan et al., "Interactive Model Cards," 2022.

123 Crisan et al.; Liao et al., "Designerly Understanding"; Moore, Liao, and Subramonyam, "fAIlureNotes."

124 https://huggingface.co/docs/api-inference/

125 https://platform.openai.com/playground

126 Liao et al., "Designerly Understanding."

127 Crisan et al., "Interactive Model Cards," 2022; Liao et al., "Designerly Understanding"; Moore, Liao, and Subramonyam, "fAIlureNotes."

128 Crisan et al., "Interactive Model Cards," 2022.

129 Phillips-Wren and Adya, "Decision Making under Stress."

130 Hahn, Lawson, and Lee, "The Effects of Time Pressure and Information Load on Decision Quality."

131 Crisan et al., "Interactive Model Cards," 2022.

132 Bender and Friedman, "Data Statements for Natural Language Processing."

133 Gebru et al., "Datasheets for Datasets"; McMillan-Major, Bender, and Friedman, "Data Statements."

134 Bhat et al., "Aspirations and Practice of Model Documentation."

135 Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

136 Hind et al., "Experiences with Improving the Transparency of AI Models and Services."

137 "Improving Documentation for AI Governance."

138 Vallor, Raicu, and Green, "Overview of Ethics in Tech Practice."

139 E.g., interventions based on value sensitive design such as Ballard, Chappell, and Kennedy, "Judgment Call the Game."

140 Madaio et al., "Learning about Responsible AI On-The-Job"; Chang and Custis, "Understanding Implementation Challenges in Machine Learning Documentation."

141 Bhat et al., "Aspirations and Practice of Model Documentation."

142 Holland et al., "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards"; Richards et al., "A Methodology for Creating AI FactSheets"; Pushkarna, Zaldivar, and Kjartansson, "Data Cards"; Bhat et al., "Aspirations and Practice of Model Documentation"; Hind et al., "Experiences with Improving the Transparency of AI Models and Services"; McMillan-Major et al., "Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation," 2021; Heger et al., "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata."

143 Berman, Goyal, and Madaio, "A Scoping Study of Evaluation Practices for Responsible AI Tools."

144 Winecoff and Bogen, "Best Practices in AI Documentation: The Imperative of Evidence from Practice."

145 Crisan et al., "Interactive Model Cards," 2022; Liao et al., "Designerly Understanding."

146 Palinkas et al., "Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research."

147 Micheli et al., "The Landscape of Data and AI Documentation Approaches in the European Policy Context."

148 https://partnershiponai.org/workstream/about-ml/

149 https://dl.acm.org/conference/chi

150 https://cscw.acm.org/2024/

151 Timmermans and Tavory, "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis"; Tavory and Timmermans, *Abductive Analysis: Theorizing Qualitative Research*.

152 Strauss and Corbin, *Basics of Qualitative Research Techniques*.

08 References

Adkins, David, Bilal Alsallakh, Adeel Cheema, Narine Kokhlikyan, Emily McReynolds, Pushkar Mishra, Chavez Procope, Jeremy Sawruk, Erin Wang, and Polina Zvyagina. "Prescriptive and Descriptive Approaches to Machine-Learning Transparency." In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–9, 2022. https://doi.org/10.1145/3491101.3519724.

Ahlawat, Archana, Winecoff, Amy, and Jonathan Mayer. "Minimum Viable Ethics: From Institutionalizing Industry AI Governance to Product Impact." arXiv. September 11, 2024. https://arxiv.org/abs/2409.06926

Andrew Smart, Shazeda Ahmed, Jacob Metcalf, Atoosa Kasirzadeh, Luca Belli, Shalaleh Rismani, Roel Dobbe, et al. "What Is Sociotechnical AI Safety? What Do We Want It To Be? A FAccT Community Workshop." June 4, 2024.

Arnold, Matthew, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, et al. "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity." arXiv, February 7, 2019. http://arxiv.org/abs/1808.07261.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, et al. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." arXiv, April 12, 2022. http://arxiv.org/abs/2204.05862.

Ballard, Stephanie, Karen M. Chappell, and Kristen Kennedy. "Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology." In *Proceedings of the 2019 on Designing Interactive Systems Conference*, 421–33. San Diego CA USA: ACM, 2019. https://doi.org/10.1145/3322276.3323697.

Baracaldo, Nathalie, Ali Anwar, Mark Purcell, Ambrish Rawat, Mathieu Sinn, Bashar Altakrouri, Dian Balta, et al. "Towards an Accountable and Reproducible Federated Learning: A FactSheets Approach." arXiv, February 24, 2022. http://arxiv.org/abs/2202.12443.

Belz, Anya, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, et al. "Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP." arXiv, August 7, 2023. http://arxiv.org/abs/2305.01633.

Bender, Emily M., and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics 6* (December 2018): 587–604. https://doi.org/10.1162/tacl_a_00041.

Berman, Glen, Nitesh Goyal, and Michael Madaio. "A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations." In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–24. Honolulu HI USA: ACM, 2024. https://doi.org/10.1145/3613904.3642398.

Bhat, Avinash, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin L. C. Guo. "Aspirations and Practice of Model Documentation: Moving the Needle with Nudging and Traceability." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17, 2023. https://doi.org/10.1145/3544548.3581518.

Blasch, Erik, James Sung, and Tao Nguyen. "Multisource AI Scorecard Table for System Evaluation." In *AAAI FSS20: Artificial Intelligence in Government and Public Sector*, 2020.

Boyd, Karen L. "Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data." *Proceedings of the ACM on Human-Computer Interaction 5*, no. CSCW2 (October 13, 2021): 1–27. https://doi.org/10.1145/3479582.

Brajovic, Danilo, Niclas Renner, Vincent Philipp Goebels, Philipp Wagner, Benjamin Fresz, Martin Biller, Mara Klaeb, Janika Kutz, Jens Neuhuettler, and Marco F. Huber. "Model Reporting for Certifiable AI: A Proposal from Merging EU Regulation into AI Development." arXiv, July 21, 2023. http://arxiv.org/abs/2307.11525.

Chang, Jiyoo, and Christine Custis. "Understanding Implementation Challenges in Machine Learning Documentation." In *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8. Arlington VA USA: ACM, 2022. https://doi.org/10.1145/3551624.3555301.

Chmielinski, Kasia, Sarah Newman, Chris N Kranzinger, Michael Hind, Jennifer Wortman Vaughan, Margaret Mitchell, Julia Stoyanovich, et al. "The CLeAR Documentation Framework for AI Transparency Recommendations for Practitioners and Context for Policymakers." The Shorenstein Center on Media, Politics and Public Policy, May 2024. https://shorensteincenter.org/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers/.

Chmielinski, Kasia S, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. "The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence," 2020. https://arxiv.org/abs/2201.03954.

Clavell, Gemma Galdon. "Checklist for AI Auditing." *European Data Protection Board*, June 2024. https://www.edpb.europa.eu/system/files/2024-06/ai-auditing_checklist-for-ai-auditing-scores_edpb-spe-programme_en.pdf.

Crisan, Anamaria, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. "Interactive Model Cards: A Human-Centered Approach to Model Documentation." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 427–39, 2022. https://doi.org/10.1145/3531146.3533108.

Díaz, Mark, Sunipa Dev, Emily Reif, Emily Denton, and Vinodkumar Prabhakaran. "SoUnD Framework: Analyzing (So)Cial Representation in (Un)Structured (D)Ata." arXiv, December 1, 2023. http://arxiv.org/abs/2311.17259.

Friedman, Batya, Peter H. Kahn Jr., and Alan Borning. "Value Sensitive Design and Information Systems." In *Human-Computer Interaction and Management Information Systems: Foundations Advances in Management Information Systems*, 5:348-372. Armonk, NY: M.E. Sharpe, 2006. http://ndl.ethernet.edu.et/bitstream/123456789/18272/1/87..Neelke%20Doorn.pdf#page=67.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for Datasets." arXiv, December 1, 2021. http://arxiv.org/abs/1803.09010.

Geiger, R. Stuart, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. "Garbage in, Garbage out?: Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?" In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–36. Barcelona Spain: ACM, 2020. https://doi.org/10.1145/3351095.3372862.

Gilbert, Thomas Krendl, Nathan Lambert, Sarah Dean, Tom Zick, and Aaron Snoswell. "Reward Reports for Reinforcement Learning." arXiv, March 19, 2023. http://arxiv.org/abs/2204.10817.

Green, Nekesha, Chavez Procope, Adeel Cheema, and Adekunle Adediji. "System Cards, a New Resource for Understanding How AI Systems Work," 2021. https://ai.facebook.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/.

Hahn, Minhi, Robert Lawson, and Young Gyu Lee. "The Effects of Time Pressure and Information Load on Decision Quality." *Psychology & Marketing* 9, no. 5 (1992): 365–78.

Heger, Amy K., Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata." arXiv, August 24, 2022. http://arxiv.org/abs/2206.02923.

Hind, Michael, Stephanie Houde, Jacquelyn Martino, Aleksandra Mojsilovic, David Piorkowski, John Richards, and Kush R. Varshney. "Experiences with Improving the Transparency of AI Models and Services." arXiv, November 11, 2019. http://arxiv.org/abs/1911.08293.

Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards," May 2018. https://arxiv.org/abs/1805.03677.

Hupont, Isabelle, and Emilia Gomez. "Documenting Use Cases in the Affective Computing Domain Using Unified Modeling Language." arXiv, September 19, 2022. http://arxiv.org/abs/2209.09666.

Hutchinson, Ben, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. "Evaluation Gaps in Machine Learning Practice." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1859–76. Seoul Republic of Korea: ACM, 2022. https://doi.org/10.1145/3531146.3533233.

"Improving Documentation for AI Governance." Workshop, Center for Democracy & Technology (Virtual), June 20, 2024.

Liang, Weixin, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. "What's Documented in AI? Systematic Analysis of 32K AI Model Cards." arXiv, February 7, 2024. http://arxiv.org/abs/2402.05160.

Liao, Q. Vera, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. "Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience." arXiv, February 20, 2023. http://arxiv.org/abs/2302.10395.

Liao, Q. Vera, and Jennifer Wortman Vaughan. "AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap." arXiv, August 7, 2023. http://arxiv.org/abs/2306.01941.

Madaio, Michael, Shivani Kapania, Rida Qadri, Ding Wang, Andrew Zaldivar, Remi Denton, and Lauren Wilcox. "Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations." In The *2024 ACM Conference on Fairness, Accountability, and Transparency*, 1544–58. Rio de Janeiro Brazil: ACM, 2024. https://doi.org/10.1145/3630106.3658988.

Marone, Marc, and Benjamin Van Durme. "Data Portraits: Recording Foundation Model Training Data." arXiv, December 14, 2023. http://arxiv.org/abs/2303.03919.

McLennan, Stuart, Amelia Fiske, Daniel Tigard, Ruth Müller, Sami Haddadin, and Alena Buyx. "Embedded Ethics: A Proposal for Integrating Ethics into the Development of Medical AI." *BMC Medical Ethics* 23, no. 1 (December 2022): 6. https://doi.org/10.1186/s12910-022-00746-3.

McMillan-Major, Angelina, Emily M. Bender, and Batya Friedman. "Data Statements: From Technical Concept to Community Practice." *ACM Journal on Responsible Computing* 1, no. 1 (March 31, 2024): 1–17. https://doi.org/10.1145/3594737.

McMillan-Major, Angelina, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. "Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards." In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics* (GEM 2021), 121–35, 2021. https://doi.org/10.18653/v1/2021.gem-1.11.

Miceli, Milagros, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. "Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices." In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 161–72. Virtual Event Canada: ACM, 2021. https://doi.org/10.1145/3442188.3445880.

Micheli, Marina, Isabelle Hupont, Blagoj Delipetrev, and Josep Soler-Garrido. "The Landscape of Data and AI Documentation Approaches in the European Policy Context." Ethics and Information Technology 25, no. 4 (October 28, 2023): 56. https://doi.org/10.1007/s10676-023-09725-7.

Mitchell, Margaret, 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*, 220–29, 2019. https://doi.org/10.1145/3287560.3287596.

Mohammad, Saif M. "Ethics Sheets for AI Tasks." arXiv, March 19, 2022. http://arxiv.org/abs/2107.01183.

Mökander, Jakob, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. "Auditing Large Language Models: A Three-Layered Approach." *AI and Ethics*, May 30, 2023. https://doi.org/10.1007/s43681-023-00289-2.

Moore, Steven, Q. Vera Liao, and Hariharan Subramonyam. "fAIlureNotes: Supporting Designers in Understanding the Limits of AI Models for Computer Vision Tasks." arXiv, February 22, 2023. https://doi.org/10.48550/arXiv.2302.11703.

Narayanan, Arvind, and Sayash Kapoor. "AI Safety Is Not a Model Property." *AI Snake Oil* (blog), March 12, 2024. https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property.

Nicholas, Gabriel. "Grounding AI Policy: Towards Researcher Access to AI Usage Data," August 2024. https://cdt.org/wp-content/uploads/2024/08/2024-08-12-CDT-Research-Grounding-AI-Policy-report-final.pdf.

Nunes, José Luiz, Gabriel D. J. Barbosa, Clarisse Sieckenius De Souza, Helio Lopes, and Simone D. J. Barbosa. "Using Model Cards for Ethical Reflection: A Qualitative Exploration." In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems*, 1–11. Diamantina Brazil: ACM, 2022. https://doi.org/10.1145/3554364.3559117.

OpenAI. "GPT-4V(Ision) System Card," September 25, 2023. https://openai.com/index/gpt-4v-system-card/.

Palinkas, Lawrence A, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. "Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research." *Administration and Policy in Mental Health and Mental Health Services Research* 42 (2015): 533–44.

Papakyriakopoulos, Orestis, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. "Augmented Datasheets for Speech Datasets and Ethical Decision-Making." In 2023 *ACM Conference on Fairness, Accountability, and Transparency*, 881–904. Chicago IL USA: ACM, 2023. https://doi.org/10.1145/3593013.3594049.

Pepe, Federica, Vittoria Nardone, Antonio Mastropaolo, Gerardo Canfora, and Gabriele Bavota. "How Do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study," In Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension, 370–381. Lisbon, Portugal. 2024. https://dl.acm.org/doi/abs/10.1145/3643916.3644412.

Phillips-Wren, Gloria, and Monica Adya. "Decision Making under Stress: The Role of Information Overload, Time Pressure, Complexity, and Uncertainty." *Journal of Decision Systems* 29, no. sup1 (August 18, 2020): 213–25. https://doi.org/10.1080/12460125.2020.1768680.

Procope, Chavez, Adeel Cheema, David Adkins, Bilal Alsallakh, Emily McReynolds, Grace Pehl, Erin Wang, and Zvyagina, Polina. "System-Level Transparency of Machine Learning," February 22,

2022. https://ai.meta.com/research/publications/system-level-transparency-of-machine-learning/.

Pushkarna, Mahima, Andrew Zaldivar, and Oddur Kjartansson. "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI." arXiv, April 3, 2022. http://arxiv.org/abs/2204.01075.

Qi, Xiangyu, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. "Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" arXiv, October 5, 2023. http://arxiv.org/abs/2310.03693.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. Barcelona Spain: ACM, 2020. https://doi.org/10.1145/3351095.3372873.

Raji, Inioluwa Deborah, and Jingying Yang. "ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles." arXiv, January 7, 2020. http://arxiv.org/abs/1912.06166.

Reid, Kathy, and Elizabeth T. Williams. "Right the Docs: Characterising Voice Dataset Documentation Practices Used in Machine Learning." arXiv, March 19, 2023. http://arxiv.org/abs/2303.10721.

Richards, John, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. "A Methodology for Creating AI FactSheets." arXiv, June 27, 2020. https://doi.org/10.48550/arXiv.2006.13796.

Roman, Anthony Cintron, Jennifer Wortman Vaughan, Valerie See, Steph Ballard, Nicolas Schifano, Jehu Torres, Caleb Robinson, and Juan M. Lavista Ferres. "Open Datasheets: Machine-Readable Documentation for Open Datasets and Responsible AI Assessments." arXiv, December 11, 2023. http://arxiv.org/abs/2312.06153.

Rostamzadeh, Negar, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. "Healthsheet: Development of a Transparency Artifact for Health Datasets." arXiv, February 25, 2022. http://arxiv.org/abs/2202.13028.

Shaikh, Omar, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. "On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning." arXiv, June 4, 2023. http://arxiv.org/abs/2212.08061.

Shen, Hong, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. "The Model Card Authoring Toolkit: Toward Community-Centered, Deliberation-Driven AI Design." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 440–51. Seoul Republic of Korea: ACM, 2022. https://doi.org/10.1145/3531146.3533110.

Shilton, Katie. "Values Levers: Building Ethics into Design." *Science, Technology, & Human Values* 38,

no. 3 (May 2013): 374–97. https://doi.org/10.1177/0162243912436985.

Shimorina, Anastasia, and Anya Belz. "The Human Evaluation Datasheet 1.0: A Template for Recording Details of Human Evaluation Experiments in NLP." arXiv, March 17, 2021. http://arxiv.org/abs/2103.09710.

Soh, Jerrold. "Building Legal Datasets." arXiv, November 3, 2021. http://arxiv.org/abs/2111.02034.

Sokol, Kacper, and Peter Flach. "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67. Barcelona Spain: ACM, 2020. https://doi.org/10.1145/3351095.3372870.

Srinivasan, Ramya, Emily Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, and Beth Coleman. "Artsheets for Art Datasets," 2021.

Stoyanovich, Julia, and Bill Howe. "Nutritional Labels for Data and Models." *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2019.

Strauss, Anselm, and Juliet Corbin. *Basics of Qualitative Research Techniques*. Sage Publications, 1998.

Subramaniam, Pranav, Yintong Ma, Chi Li, Ipsita Mohanty, and Raul Castro Fernandez. "Comprehensive and Comprehensible Data Catalogs: The What, Who, Where, When, Why, and How of Metadata Management." arXiv, February 1, 2023. http://arxiv.org/abs/2103.07532.

Sun, Chenkai, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. "MithraLabel: Flexible Dataset Nutritional Labels for Responsible Data Science." In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2893–96. Beijing China: ACM, 2019. https://doi.org/10.1145/3357384.3357853.

Tagliabue, Jacopo, Ville Tuulos, Ciro Greco, and Valay Dave. "DAG Card Is the New Model Card." arXiv, November 20, 2021. http://arxiv.org/abs/2110.13601.

Tavory, Iddo, and Stefan Timmermans. *Abductive Analysis: Theorizing Qualitative Research*. University of Chicago Press, 2014.

Timmermans, Stefan, and Iddo Tavory. "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis." *Sociological Theory* 30, no. 3 (2012): 167–86.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." *Meta AI* (blog), July 18, 2023. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

Vallor, Shannon, Irina Raicu, and Brian Green. "Overview of Ethics in Tech Practice." *Markkula Center for Applied Ethics* (blog), May 2018. https://www.scu.edu/ethics-in-technology-practice/overview-of-ethics-in-tech-practice/.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz

Kaiser, and Illia Polosukhin. "Attention Is All You Need," 2017. https://doi.org/10.48550/arXiv.1706.03762.

Winecoff, Amy A., and Elizabeth Anne Watkins. "Artificial Concepts of Artificial Intelligence: Institutional Compliance and Resistance in AI Startups." In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 788–99. Oxford United Kingdom: ACM, 2022. https://doi.org/10.1145/3514094.3534138.

Winecoff, Amy, and Miranda Bogen. "Best Practices in AI Documentation: The Imperative of Evidence from Practice," July 25, 2024. https://cdt.org/insights/best-practices-in-ai-documentation-the-imperative-of-evidence-from-practice/.

Yang, Ke, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, Hv Jagadish, and Gerome Miklau. "A Nutritional Label for Rankings." In *Proceedings of the 2018 International Conference on Management of Data*, 1773–76. Houston TX USA: ACM, 2018. https://doi.org/10.1145/3183713.3193568.

Zhang, Longteng, Xiang Liu, Zeyu Li, Xinglin Pan, Peijie Dong, Ruibo Fan, Rui Guo, et al. "Dissecting the Runtime Performance of the Training, Fine-Tuning, and Inference of Large Language Models." arXiv, December 1, 2023. http://arxiv.org/abs/2311.03687.

Zheng, Xinyi, Ryan A. Rossi, Nesreen Ahmed, and Dominik Moritz. "Network Report: A Structured Description for Network Datasets." arXiv, June 7, 2022. http://arxiv.org/abs/2206.03635.

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY