Understanding Implementation Challenges in Machine Learning Documentation

Jiyoo Chang
Partnership on AI, USA
jiyoo@partnershiponai.org

Christine Custis
Partnership on AI, USA
christine@partnershiponai.org

ABSTRACT

The lack of transparency in machine learning (ML) systems makes it difficult to identify sources of potential risks and harms. In recent years, various organizations have proposed standardized frameworks and processes for documentation for ML systems. However, it remains unclear how practitioners should implement and operationalize ML documentation in their workflows. We conducted semi-structured interviews with 24 practitioners in various organizational contexts to identify key implementation challenges and strategies for alleviating these challenges. Our findings indicated that addressing the *why*, *how*, and *what* of documentation is critical for implementing robust documentation practices.

CCS CONCEPTS

• **Social and professional topics** → Implementation management.

KEYWORDS

documentation, ML model evaluation, datasheets, model cards, standardization, implementation

ACM Reference Format:
Jiyoo Chang and Christine Custis. 2022. Understanding Implementation Challenges in Machine Learning Documentation. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), October 06–09, 2022, Arlington, VA, USA*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3551624.3555301

1 INTRODUCTION

In recent years, numerous organizations in public and private spheres have taken initiatives to establish ethical principles and guidelines for Artificial Intelligence (AI) technologies [18, 23]. Among a set of principles emerging in current AI ethics guidelines, one of the most prevalent values mentioned is transparency [18], which relates to transparency of the algorithms and data used to build AI systems and their governance [9]. Increasing transparency in machine learning (ML) algorithms can help diverse stakeholders such as policy makers, auditors, developers and consumers understand how the technology works, perform meaningful audits, and identify sources of harm.

EAAMO '22, October 06–09, 2022, Arlington, VA, USA

ACM ISBN 978-1-4503-9477-2/22/10…$15.00
https://doi.org/10.1145/3551624.3555301

ML documentation has been proposed as a way to operationalize transparency [25] and is meaningful both as an artifact and a process [26]. Documentation serves as an artifact that communicates the underworkings of ML systems such as development processes, decisions, and impacts to diverse audiences. It also serves as a process that helps technology builders identify potential harms throughout the ML life cycle. Additionally, documentation could benefit teams by reducing technical debt, preserving institutional knowledge, and improving communication, and enabling reproducibility [12, 21, 32].

Though documentation can serve all of these roles in theory, it remains unclear how and why practitioners currently implement and operationalize ML documentation into their work. In order to promote a broader adoption of documentation practice across industry, implementation challenges need to be examined in diverse organizational settings including startups and small- to mid-sized enterprises. In this study, we gather insights from organizations in a wide range of sizes and domains that are active in documentation practices across various implementation stages. Our aim is to consolidate their efforts into a set of best practices and help bridge the gap between responsible AI principles and practice by addressing the following research questions:

- **RQ1:** What are organizational challenges faced by teams developing AI/ML products for implementing documentation?
- **RQ2:** How do organizations currently address documentation challenges?

Moving from identification of challenges to solution-oriented discourse, we aim to provide operational insights that help practitioners foresee documentation challenges in their own settings and provide solution options for addressing them.

2 RELATED WORK

2.1 Current Documentation Practices

Current literature shows that data science and machine learning teams generally do not have rigorous and systematic documentation practices [12]. Many teams lack a rigorous practice of recording decisions and experimental results as they happen [10, 29]. Computational notebooks are commonly used during the exploratory phase, however, practitioners consider them as personal, exploratory, and messy, hindering them from sharing computational notebooks with others [28]. Studies have pointed out that many widely used datasets remain sparsely documented, which can make models trained on these datasets difficult to replicate or comprehend [2, 21]. Geiger et al found inconsistent practices particularly on labeling and annotation processes in dataset publications [13].

2.2 Types of Documentation Frameworks

In efforts to standardize documentation practices, numerous academic and industry researchers have suggested using templates and checklists as a structured way for documenting ML datasets, models, and systems [1, 3, 5, 12, 19, 20, 22, 24]. While these frameworks have different areas of focus, they provide a structure for practitioners to communicate critical information about datasets and models, such as intended uses, context, methodologies, development process, performance, potential biases, impact, and ethical implications.

Dataset documentation has been a focus in this area as data used to train models is increasingly being recognized as a source of biases [15]. Templates such as Datasheets for Datasets [12], Dataset Nutrition Labels [5], and Data statements [3] have been created as structured formats for documenting dataset's motivation, composition, collection process, recommended uses, characteristics of actors involved, ethical concerns. There are growing examples of datasheets accompanied in publications of datasets in various domains [2, 6, 11, 30]. Boyd presents evidence that datasheets help engineers recognize and understand ethical problems in training datasets [4].

Templates for documenting models and AI systems seek to increase transparency amongst developers, users and stakeholders. Factsheets [1] are intended to incorporate information from all phases of the life cycle, from training, testing, deployment, and monitoring of models. Model Cards [22] are frameworks to provide key information about how models work and evaluation of models across demographic groups and disclosure of intended uses and other information. We are starting to see examples of public-facing model cards released with models [26 - 31]. It must be noted that templates are just one of many ways to increase transparency — other strategies such as auditing and red-teaming have also been proposed [25].

Checklists have also been presented as a shared framework for teams to assess and address biases in their data, models, and outcomes in a structured way [8]. Although checklists are not explicitly intended for documentation, checklists can prompt some teams to document their datasets and various aspects of their design and engineering practices [19].

Notably, the "Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles" (ABOUT ML) Project [26] led by the Partnership on AI seeks to consolidate disparate efforts in documentation frameworks and guidelines to works towards a standard ML documentation practice.

2.3 Implementation Challenges for Documentation

Several studies have examined challenges in applying these frameworks into practice at both organization-wide and individual practitioner levels. As might be expected, organizational culture and structure play a pivotal role in implementing responsible AI initiatives in practice [27]. Introducing new documentation standards requires changes to organizational infrastructure and workflows, and previous studies have found that employee responsibilities and incentive structures are often not kept in alignment with these changes [12]. In acknowledgement of these challenges, the authors of the "AI fairness checklist" suggest successful implementation requires checklists to be customizable by teams, and integrated into organizational goals and priorities, perhaps as metrics or key performance indicators [19].

A common challenge faced by practitioners is the perception of documentation as a burden [8] as it takes a lot of effort to complete and requires qualitative insights and knowledge from multiple stakeholders [24]. In addition, practitioners often face challenges around the loss of important details, concerns with proprietary information, and understanding the needs of diverse consumers of documentation, such as regulators and developers [14].

This paper builds on and contributes to the growing efforts in operationalizing ML documentation in two ways. First, a vast majority of current studies on documentation practices are done at large, multinational technology companies [8, 12, 17, 32], which may not fully represent a broader population of professionals in the AI/ML community. Moreover, qualitative studies on practitioners' experiences with documentation are primarily conducted in the context of a documentation tool or framework specifically designed for an organization [14, 19, 24]. We expand this area of work by interviewing practitioners at various organizational settings in terms of geographic location, size, and domain. Our findings uncover challenges unique to organization types as well as common challenges observed across interviewed organizations.

Second, much of the past research on ML documentation focuses on identifying organizational and practitioner challenges and less is known about solutions for addressing these challenges. Moving beyond the identification of challenges, this study contributes to the solution-oriented discourse and presents implementation strategies and best practices solicited from experts in ML documentation. By surveying documentation practices at less resourced and less experienced organizations in addition to Big Tech companies, we contribute to a more holistic understanding of best practices, which is critical for facilitating responsible ML development for all [16].

3 METHODS

3.1 Interviewees

We recruited interviewees using a convenience sampling strategy [31]. We reached out to responsible AI experts in documentation and posted recruitment outreach messages on social media platforms. Individuals who were interested in participating in the interview study filled out a questionnaire. Participants were selected based on the criteria that (i) they are industry practitioners in the AI/ML field, and (ii) their current work involves development or implementation of ML data, models, or systems documentation.

3.2 Materials

The semi-structured interviews followed an interview guide, which contained questions related to current documentation practices, implementation challenges, and attempted solutions. The following are samples of questions asked during the interviews:

- What is the current process for documentation? Which stakeholders are involved?
- What motivated you to adopt documentation?
- What challenges have you faced when implementing documentation?

- What are barriers to scaling documentation across the organization?
- Which organizational changes have been made to adopt documentation?
- What solutions have been implemented to overcome documentation challenges?

3.3 Interview Protocols

The interview study had two phases, starting with initial formative group interviews with six organizations from our organization's network. After gaining a broad sense of challenges and needs in documentation, we conducted 24 in-depth, semi-structured interviews on an individual basis except for one interview with two participants from the same organization.

Participants who met the selection criteria signed the informed consent forms. We conducted interviews on Zoom[1] 60 minutes. Prior to the start of the interview, we verbally asked for consent from participants to audio record the interview. Interviews were transcribed on Temi[2] and then were proofread manually to correct transcription errors and redact any personal identifiable information.

After completing the interview, participants were compensated US$75 via electronic payment for their time. Within a couple days after the interview, participants received redacted transcripts and were given an option to remove or clarify any parts of the interview.

3.4 Analysis

We coded and analyzed the interview transcripts using thematic analysis. First, the main researcher read the interview transcripts and extracted 419 quotes that corresponded to the three interview topics: current documentation practices, implementation challenges, and solutions implemented. In the next round of coding, the quotes were codified with specific themes and then grouped into three high-level themes. The main researcher received feedback on the initial themes from a group of 9 colleagues and refined them into final three themes. The themes were organized in a way that could help practitioners foresee implementation challenges and apply potential solutions in their own settings.

4 FINDINGS

Participants had a wide range of roles. 14 participants were Data Scientists or ML engineers, 4 Executives, 3 Product Managers, 2 Ethics or Policy roles within companies, and 1 Technical writer. 14 participants were based in the US while the rest of the participants were based in Ireland, Belgium, Japan, Sweden, Netherlands, India, Switzerland, Germany, Russia, and Egypt. As shown in Table 1, we categorized Participant IDs by organization types.

4.1 Overview of Findings

We found that implementation challenges for documentation emerged around the questions of why, how, and what of documentation:

- Incentives: why are we documenting?
- Tools and Workflows: how are we documenting?
- Content: what are we documenting?

Table 1: Participants' organization types and IDs

Organization Type	Participant IDs	Count
AI Platform Companies	P2, P7, P9, P3, P10, P11, P13, P20, P21, P22, P23, P24	12
Big Tech Corporations	B1, B4, B5, B6, B8, B16, B17, B18	8
Consulting firms	C12, C14, C15, C19	4

For each question, we summarize high-level findings on major challenges (RQ1) and recommendations (RQ2) that were commonly shared by interviewees. As we interviewed organizations at various implementation stages, some of the lessons learned by organizations further along in the process might be helpful for organizations in earlier stages.

4.2 The Why of Documentation: Lack of Incentives

One of the most commonly expressed implementation challenges was understanding the value of documentation both at organizational and individual levels, a critical step for implementing documentation across the organization. More broadly, we noticed that a lack of awareness and understanding in responsible AI hindered practitioners from understanding why documentation was important and necessary.

4.2.1 Challenges on Incentives. Lack of organizational incentives. We found that some organizations had difficulty understanding the value proposition or return on investment (ROI) of documentation as one participant expressed, "People cannot find a reason to pay for something that they don't see direct value and outcome from" (C15). In the absence of specific regulation or business requirements that "force" organizations to document, organizations were not motivated to prioritize documentation amongst other competing priorities. A founder at a startup said that they felt the pressure to prioritize building products over documentation and described documentation as a "luxury" given many other priorities. In addition, organizations were generally hesitant to be transparent through documentation because how their models work is closely guarded proprietary information, as noted in previous studies.

Lack of Individual incentives. Having practitioners understand the value and necessity of documentation at an individual level was also a commonly referenced challenge. A participant leading documentation efforts at a large organization explained, "[Practitioners] receive very little benefit from the [model] cards themselves, which I think is probably true of all documentation. If it's in my head, why do I have to take the time to write it?" (B8).

At organizations where documentation was not established as a standardized practice, practitioners viewed documentation as "a nice-to-have but always an afterthought" (P15). Some engineers and developers expressed that they would rather spend time on technical projects than on documentation, because documentation

[1] https://zoom.us/for
[2] https://www.temi.com/

was viewed as non-technical work and did not affect their performance evaluation or promotion. "People don't like documenting just because they could be using their time more productively building software or building new projects" (P13). This participant went on to explain that more broadly, ML is a relatively new field that has focused a lot on innovation rather than rigor of how it is built.

Gap in understanding of responsible AI. We found that practitioners generally did not see much value in answering the types of questions posed by documentation tools. A policy manager who was implementing documentation processes at scale explains, "Teams generally are having a very difficult time understanding why the questions posed in a model card are relevant to what they're doing. In general, there's a very large disconnect between practitioners who do AI every single day and the potential harms that AI might cause. At this point, ethics of AI and building AI responsibly is still not in the vernacular of your typical AI practice" (B1).

The rationale behind this might be that few members of the AI workforce have expertise in AI ethics, i.e. those who have a strong understanding of both ML and sociotechnical systems, as further described by a participant, "Responsible AI is a socio-technical concept. It's not just like, use this library and implement these algorithms and suddenly your model is now fair and bias free. It's more so to think about the context of what your model is going to be deployed and where these harms originate and other things you can do...How do we educate people to think about the social side and how to train, upscale people and think about the sociological impacts of their algorithms?" (B4). Another participant suggested integrating ethics into ML education as a way to meet this need.

4.2.2 Recommendations for Incentives. Communicate the value proposition of documentation. Participants shared that it was crucial to communicate to practitioners the value that documentation generates at different levels of the organization. The value proposition of documentation can be shared via top-down or bottom-up efforts. Participants shared a number of benefits of documentation that they have experienced in their work:

(1) Documentation makes ML projects more scalable. Documentation can serve as an effective tool for knowledge transfer, making it easier to onboard new members and interact with other groups in the organization.

(2) Documentation strengthens team coherence, bringing team members on the same page and allowing teams to keep track of other projects and expertise in the organization, as explained by an engineering manager: "Our team has an interest to document things well, just so that other people are also aware of what we are doing, if they can help and maybe unblock or brainstorm something together" (B16).

(3) Documentation helps build institutional knowledge and guide the decision-making process. Documentation from previous projects can instruct decisions for current projects. One engineer said, "I realized that revisiting what we've done before has become like an extremely, extremely important thing in our day to day. Being able to go back and see what we've done saves us a lot of time" (P22).

Additionally, a product manager described how documentation helps prevent organizations from making the same mistakes. "Effectively you have a working document of tenants and principles from lessons learned and that keeps the company overall from making too many bad decisions over again" (B18).

(4) Documentation enables easier backtracking of errors and reduces technical debt in the long run, as described by a policy manager, "If we can't track how AI is built now, it will take years to disentangle down the line" (B1).

(5) Documentation makes models more robust by helping teams find gaps and potential biases. "If you rigorously evaluate tests or machine learning models before you put them in production, you can be more confident that they won't go haywire or be something unexpected when you deploy them...So you can understand the robustness across different sets of features" (B4).

(6) Documentation helps build trust with users and consumers by enhancing understanding of models. An engineer described how documentation enhances trust with their customers: "They know how [the AI model] works so that they can trust it and they can be more efficient" (P3). A product manager shared the long-term benefit of being transparent with users, "Transparency builds trust and ultimately trust will win over in the long haul" (B18).

One way to present the value of documentation and increase its awareness to an organization was sharing the use case of documentation. For example, a product manager who started using model cards with his team shared the implementation journey at an organization-wide meeting, which generated interest among other teams: "People from other teams heard about it and set up meetings to have us talk about it and how we are using it and how we are putting it together" (P21). It is important to note that this organization had buy-in from leadership to prioritize transparency and ethics, which may have contributed to the enthusiasm of adopting documentation amongst teams.

Education and training in responsible AI. Initiatives from the leadership are critical, as one participant said, "When it's pushed from the leadership, people really make an effort to make a change" (B5). One of the ways that the leadership teams can contribute to the adoption of documentation and culture of documentation is by upskilling the organization on responsible AI topics through education and training initiatives. For example, an organization shared how they brought up awareness through a series of talks. "More recently, we've focused on just bringing up awareness through a series of conversations with the broader team in general. We recently have done AI talks within the data scientists that have honed on things like bias in AI versus model maintenance" (P22).

Beyond targeting data scientists and engineers, a few participants mentioned the importance of creating educational materials for product managers or senior leadership and increasing awareness of these topics across the company. A participant gave an example of an organization-wide educational program on AI ethics: "An e-learning program to all employees and such programming includes what AI ethics is, and why that matters and what kind of incidents actually happen in the market" (B6).

These kinds of top-down educational efforts are potential ways to help bridge the gap in understanding of responsible AI amongst practitioners and help explain why practices such as documentation are important and necessary.

4.3 The How of Documentation: Tools and Workflows

An important aspect of implementing documentation was equipping teams with the right tools and workflows. Practitioners sometimes faced technical barriers and lack of clarity when adopting existing tools and frameworks for documentation. The process of documentation was often viewed as "tedious" due to the time-consuming information gathering process.

4.3.1 Challenges for Documentation Tools and Workflows. *Technical barriers with tools.* One common form of documentation that practitioners discussed was the use of fairness and explainability tools to help profile datasets and models. Many of these tools required programming skills which created barriers to non-technical users. In addition, with a broad number of existing tools, choosing the right tool was overwhelming for practitioners. For example, some participants had difficulty determining which fairness metrics to use to help document performance on different groups of population and needed more guidance on criteria for fairness attributes. Practitioners expressed that a repository of available tools for documentation would be helpful.

Additionally, some practitioners had trouble integrating the documentation tools with current workflows, as one participant put it, "[Documentation tool] has its own separate API, it's hard to get it to play well with the traditional Python workflow" (B4). Another participant said, "When you have this big collection of tools that don't necessarily talk well to one another, either you write the integration layer yourself or you end up doing things by hand" (P2).

In another case, a data scientist found it difficult to export outputs from profiling tools which are used to obtain metadata and said that they got errors every single time. They went on to suggest a rationale for this, "I don't know whether it's more about tools or people's awareness. If people are more aware that we need to do this, then there'll be better tools that come out" (C7).

Tedious information gathering process. Perception of documentation as a burden is not new as mentioned earlier in Related Work. Participants in this study commonly described documentation as 'tedious' and time-consuming. Some pointed out that they had to do extra work to gather scattered information. For instance, a data scientist described having to rerun calculations when summarizing data, "If I want to find something about a particular column, I have to go back, write a code and then copy and paste it or write in my own words to put it in my documentation. I wish I didn't have to write code for every single question that I have about data before I write it" (C7).

Information was also scattered around the organization. For example, a participant had to "go through different people to find out just one thing about one single feature" (C7). Practitioners who worked with clients had to access information from clients, which were sometimes not available as one participant mentioned, "Some information is not clear from the client side so we need to ask about it again. Maybe the client doesn't know the information, so it will be unknown for us" (P24).

In order to make the information gathering process less tedious, many engineers expressed that an automated tool that captures information through the ML lifecycle or an interactive tool that gives prompts and questions for users to answer could help lessen the burden of documentation.

High turnover of project owners. We found that tracking down information from data or model owners who changed teams or left the company was a major challenge, especially at larger organizations. Without robust documentation in place, knowledge about projects got lost when those project owners left, as one participant described, "That person who left knows all about the model they developed, but nobody else knows" (C15).

Another participant said that even in a span of a few months, a model had changed many times and the model owner left the company, which made it difficult to update the model documentation. This challenge was especially prevalent for organizations that were retrospectively documenting their older models and systems.

4.3.2 Recommendations for Documentation Tools and Workflows. *Frequently used tools and frameworks.* Documentation tools that were most frequently mentioned by participants were Github[3], Confluence[4], Google Docs[5], and Overleaf[6]. These tools commonly provide a centralized place for stakeholders to collaborate on documentation and share documents.

Several practitioners mentioned having referred to The Assessment List for Trustworthy Artificial Intelligence, published by the European Commission's AI High-Level Expert Group [7]. This guide aims to help practitioners assess whether their AI systems adheres to seven requirements of Trustworthy AI, which includes Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Non-discrimination and Fairness, Societal and Environmental Wellbeing, and Accountability. Practitioners mentioned using this guide to inform the structure of the documentation content.

Forcing functions. In order to operationalize documentation, participants shared examples of forcing functions they implemented in their workflows:

- Requiring documentation for the review process before launching products
- Making documentation a project deliverable
- Assigning documentation as a task in agile workflows
- Having documentation at the center how meetings are run

A participant who works at a large company that has documentation as part of its culture elaborated that having documentation as the center of meetings provided incentives for creating good documentation: "Having regular check-ins is a good forcing function, like you want to make sure you've written everything very carefully in case there's follow-up, or having a big meeting with people about something especially with cross-functional stakeholders who don't see your work all the time. That's a good way to make sure you do a good job of documentation" (B18). In another example, an engineer who worked in an agile structure had documentation assigned as a task. He explained how having documentation scoped in the project plan incentivized him: "Being able to finish your tickets and getting your work done itself is a reward" (P22).

[3]https://github.com/
[4]https://www.atlassian.com/software/confluence
[5]https://docs.google.com
[6]https://www.overleaf.com

Rough notes along the way. Many participants found it effective to "dump notes" throughout the development process or keep a running report to write something in everyday. For example, a consultant said, "Whatever parameters or important metrics are there I collected, stored it, and I just dumped it into a file and then used that file at the end of the lifecycle" (C12). The rough notes would be polished and formalized at the end of a sprint or a stage of the ML lifecycle. A participant working in an agile workflow said after the assigned task on the ticket was completed, the team extracted information contained in the tickets and integrated it into their documentation. This system helped the team not lose any important information.

4.4 The What of Documentation: Content

Practitioners had various challenges determining what to put in their documentation and how much detail and complexity to include. Engineers found it difficult to translate technical information in ways that could be widely understood. In order to overcome some of these challenges, some teams had multiple stakeholders contribute to the documentation and iteratively developed the documentation content using feedback from the documentation audience.

4.4.1 Challenges for Documentation Content. Issues with off-the-shelf templates. About 37% (9 out of 24) of the participants used publicly available templates such as model cards and factsheets. These templates provided a useful starting point on what to document and were customized to organizational needs. For example, one participant mentioned how a documentation template they tried to use contained too many questions, some of which were not easy to answer, so they abridged the template to their needs. Notably, multiple participants discussed how questions about anticipating possible downstream impact, use, or misuse of their datasets and models were particularly difficult to answer. One participant explained that "we don't know what we don't know" and that incomplete knowledge of systems could lead to blindspots when answering these questions.

Finding the right level of detail. Many participants found it challenging to find the right amount of details to include in the documentation without being overly long or burdensome. Given that building ML models involves exploratory and iterative work, documenting every decision made along the way and explaining how the model was created could be overwhelming: "There's a tension between, I don't want to put too much, because then it just turns into a wall of documentation and no one's going to read it, but if you don't put enough, then maybe you're not really exploring the limitations of the model in various domains. (P21).

Some participants explained that this tension of having too much or too little details came from not knowing who the target audience is or what their audience might need. In terms of prioritizing content of documentation, nascent teams especially faced challenges as they lacked structure in their documentation, "Since we don't have a template at this point of time and the team is also very new, we don't have any defined structure; how we want to record this entire ML algorithm or speak less about algorithms, or do we need to speak more about the methods..." (C14).

Translating technical information for non-technical audiences. Engineers had particular challenges when translating technical information to non-technical audiences. We found that simplifying complex algorithms to an audience who might not have even a basic understanding in ML was challenging. Creating non-technical documentation often requires a diverse set of skills such as quantitative analysis, qualitative insights, and data visualization, which not all engineers or developers are trained or equipped to properly do so. For example, an ML engineer said that in their career, they have only written technical documentation, such as providing docstrings on functions, rather than non-technical documentation with qualitative insights. Another related challenge experienced by engineers was switching between coding and writing as one engineer described, "If I'm in a coding mode, I lose my language ability. I cannot write the documentation and code in the same hour" (C15). Another engineer said, "I'm focused on the programming part, so sometimes it's hard for me to switch to English and write a complete sentence" (P10). Although it would help to have time blocked out for documentation, in many cases, time for documentation was not allocated by managers or from clients.

4.4.2 Recommendations for Documentation Content. Collaborative documentation efforts. As mentioned above, engineers are usually tasked with documentation work, even non-technical documentation, because they work most closely with datasets and models and know their technical details. Given the challenges they reasonably face in this task, it is worthwhile to consider how non-technical stakeholders can lessen the documentation burden from engineers and contribute their expertise to create documentation targeted to a wide range of audiences.

- *Product or Project Managers:* give an overall structure of documentation, provide qualitative insights
- *Business Analysts or Sales Specialists:* provide a high-level overview of models or services, ensure that technical information is comprehensible to non-technical audiences, create data visualization
- *User Experience Designer or Content Designers:* word documentation questions in a way that is understandable by documentation creators, make documentation more usable and readable to consumers
- *Technical writers:* polish writing, translate technical information to diverse audiences

We also found that many teams iteratively created and revised their documentation in feedback loops, whether internally with team members or externally with clients. A participant shared how their team set up an interdependent feedback system, which required everyone to complete their part on time:

"Actually the team was set up in a very smart way. So they had business analysts writing the requirements that they had developers do them and then testers. And so each of them were kind of accountable to each other. So you couldn't slack because the other person has to check on you, the one who wrote this requirement or the one who tested this requirement and vice versa. So that was a very good feedback loop for everyone involved. And you had to write good documentation as a business analyst because the next person will not get it otherwise." (C15)

Table 2: Summary of Main Findings

Themes	Major Challenges (**RQ1**)	Recommendations (**RQ2**)
1. The Why: Incentives	Understanding the value of documentation	Communicate the value proposition of documentation
	Lack of knowledge in responsible AI	Education and training in responsible AI
2. The How: Tools & Workflows	Technical barriers with tools	Frequently used tools and frameworks
		Forcing functions
	Tedious information gathering process	Time allocation
		Rough notes along the way
	High turnover of project owners	
3. The What: Content	Issues with using off-the-shelf templates	Collaborative documentation efforts
	Finding the right level of detail	Iterative feedback from documentation audience
	Translating technical to non-technical	

Iterative feedback from documentation audience. - A few participants mentioned that getting frequent feedback from their documentation readers helped them make sure that their documentation is not too lengthy or missing any important information. A consultant who develops AI solutions for clients said:

"We also make sure to get a lot of feedback from our clients. So if it's something that we do send to our client, we will many times actually iterate on whatever we sent based on their feedback. So in a way, you don't want to get feedback that the documentation is not good, or you want to make sure you are covering all bases that you're making a client happy." (P22)

This iterative process helped practitioners find the right balance in content and complexity by centering on the needs of the documentation users. Practitioners who worked heavily with clients, such as consults, especially used this iterative process to meet their clients' wants and needs and explain how the AI solution was developed and how to implement the models in a way that made sense to their clients.

4.5 Challenges by Organization Type

We observed that certain types of organizations experienced some of the challenges mentioned above more significantly in addition to unique challenges given their business model, resources, and experience.

AI consultancy firms and platform companies. These organizations had unique challenges in the context of communicating how the AI solution was developed and how to implement the models to their clients. Notably, documentation depended heavily on clients' wants, needs, and working style. One participant said that while some clients were very organized and requested thorough documentation, other clients explicitly told them not to prioritize documentation. Some consultants voiced that they felt uncertain about telling their clients what is considered "ethical". The diverse problems and requirements from clients make it difficult to standardize documentation. In addition, consultants had to often make sure that clients actually read and utilize the documentation to implement models and reproduce results. One consultant said, "It seems like more people than not are the kind of people who call tech support before reading docs" (P21). Therefore, creatin documentation that is highly usable and readable to their clients was a priority to these types of organizations.

Startups and smaller organizations. A major challenge for smaller organizations was the lack of resources to develop and maintain documentation. A participant working at a startup who felt the pressure to prioritize building products over documentation described documentation as a "luxury" given many other competing priorities. Another participant said that their products are not mature enough to have thorough documentation. A CTO of a startup said, "I can't justify hiring a risk manager at this point in time. If I go to my board and say, I'm going to be hiring a risk manager, as we're doing all these things, or I hired another data scientist, the word would be to add another data scientist or add more features to the product." (P13). Finding the motivation to invest the time and effort into documentation was especially challenging for startups.

Big Tech companies. Big tech companies had unique challenges of implementing documentation at scale given a large number of models in production. One of the challenges was selecting the models that should have documentation amongst many models with varying levels of maturity and impact. Creating documentation for older models was especially challenging, because it was hard to track information from model owners who had moved teams or left the organization. A barrier to standardizing a documentation template across an organization is that teams used different tech stacks. In addition, certain terms were used and understood differently by teams, especially between teams that build products and teams that review the products.

5 CONCLUSION

Based on the interview study with 24 AI/ML practitioners, we identified key implementation challenges in organizational and individual incentives, documentation tools and workflows, and documentation content. As potential solutions to address some of these challenges, we highlighted aligning the organization on why documentation was important, equipping teams with effective documentation tools and workflows, and collaboratively creating documentation with expertise and feedback from multiple stakeholders. Our findings indicate that answering the why, how, and what of documentation will be an important step for organizations to implement robust and sustainable documentation practices.

One of the limitations of this study is that our findings are based on qualitative interviews from practitioners with a wide range of roles from organizations of different sizes, maturity, domains, and geographic locations. Although this provided perspectives from diverse organizational contexts beyond what had mostly been studied in previous work, the challenges and recommendations may not apply broadly to all practitioners or organizations. This may be due to the heavy use of qualitative data and a small sample size in each organization type. In future work, we seek to apply and test some of the recommendations with practitioners and help establish best practices for documentation. In addition, we hope to address some of the practitioners' needs mentioned such as a repository of documentation frameworks or automated documentation tools in the future.

We hope that this work helps practitioners foresee documentation challenges in their own settings and provide solution options for addressing them. The impact we believe this work has and will continue to have is helping to create an organizational infrastructure for ethics in ML and helping to increase responsible technology development and deployment via transparency and accountability.

ACKNOWLEDGMENTS

We want to thank all the interviewees that took the time to discuss their work and share valuable insights.

REFERENCES

[1] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. 2019. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv:1808.07261 [cs] (Feb. 2019). http://arxiv.org/abs/1808.07261 arXiv: 1808.07261.

[2] Jack Bandy and Nicholas Vincent. 2021. Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus. arXiv:2105.05241 [cs] (May 2021). http://arxiv.org/abs/2105.05241 arXiv: 2105.05241.

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

[4] Karen L. Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (Oct. 2021), 1–27. https://doi.org/10.1145/3479582

[5] Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2020. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. (2020), 7.

[6] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC : Question Answering in Context. arXiv:1808.07036 [cs] (Aug. 2018). http://arxiv.org/abs/1808.07036 arXiv: 1808.07036.

[7] European Commission. 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. (2020).

[8] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, Tracks & Data: an Algorithmic Bias Effort in Practice. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk, 1–8. https://doi.org/10.1145/3290607.3299057

[9] Virginia Dignum. 2017. Responsible Autonomy. arXiv:1706.02513 [cs] (June 2017). http://arxiv.org/abs/1706.02513 arXiv: 1706.02513.

[10] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show Your Work: Improved Reporting of Experimental Results. arXiv:1909.03004 [cs, stat] (Sept. 2019). http://arxiv.org/abs/1909.03004 arXiv: 1909.03004.

[11] Christian Garbin, Pranav Rajpurkar, Jeremy Irvin, Matthew P. Lungren, and Oge Marques. 2021. Structured dataset documentation: a datasheet for CheXpert. arXiv:2105.03020 [cs, eess] (May 2021). http://arxiv.org/abs/2105.03020 arXiv: 2105.03020.

[12] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. arXiv:1803.09010 [cs] (March 2020). http://arxiv.org/abs/1803.09010 arXiv: 1803.09010.

[13] Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Jan. 2020), 325–336. https://doi.org/10.1145/3351095.3372862 arXiv: 1912.08320.

[14] Michael Hind, Stephanie Houde, Jacquelyn Martino, Aleksandra Mojsilovic, David Piorkowski, John Richards, and Kush R. Varshney. 2019. Experiences with Improving the Transparency of AI Models and Services. arXiv:1911.08293 [cs] (Nov. 2019). http://arxiv.org/abs/1911.08293 arXiv: 1911.08293.

[15] Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem". Patterns 2, 4 (April 2021), 100241. https://doi.org/10.1016/j.patter.2021.100241

[16] Aspen Hopkins and Serena Booth. 2021. Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (July 2021), 134–145. https://doi.org/10.1145/3461702.3462527 arXiv: 2110.02932.

[17] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. arXiv:2010.13561 [cs] (Jan. 2021). http://arxiv.org/abs/2010.13561 arXiv: 2010.13561.

[18] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 9 (Sept. 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[19] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376445

[20] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. (2021), 19.

[21] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, Virtual Event Canada, 161–172. https://doi.org/10.1145/3442188.3445880

[22] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency (Jan. 2019), 220–229. https://doi.org/10.1145/3287560.3287596 arXiv: 1810.03993.

[23] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence 1, 11 (Nov. 2019), 501–507. https://doi.org/10.1038/s42256-019-0114-4

[24] David Piorkowski, Daniel González, John Richards, and Stephanie Houde. 2020. Towards evaluating and eliciting high-quality documentation for intelligent systems. arXiv:2011.08774 [cs] (Nov. 2020). http://arxiv.org/abs/2011.08774 arXiv: 2011.08774.

[25] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. arXiv:2001.00973 [cs] (Jan. 2020). http://arxiv.org/abs/2001.00973 arXiv: 2001.00973.

[26] Inioluwa Deborah Raji and Jingying Yang. 2020. ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles. arXiv:1912.06166 [cs, stat] (Jan. 2020). http://arxiv.org/abs/1912.06166 arXiv: 1912.06166.

[27] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices. arXiv:2006.12358 [cs] (March 2021). https://doi.org/10.1145/3449081 arXiv: 2006.12358.

[28] Adam Rule, Aurélien Tabard, and James D. Hollan. 2018. Exploration and Explanation in Computational Notebooks. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173606

[29] D Sculley, Jasper Snoek, Ali Rahimi, and Alex Wiltschko. 2018. ON PACE, PROGRESS, AND EMPIRICAL RIGOR. (2018), 4.

[30] Ismaïla Seck, Khouloud Dahmane, Pierre Duthon, and Gaëlle Loosli. 2018. Baselines and a datasheet for the Cerema AWP dataset. arXiv:1806.04016 [cs, stat] (June 2018). http://arxiv.org/abs/1806.04016 arXiv: 1806.04016.

[31] Samuel J. Stratton. 2021. Population Research: Convenience Sampling Strategies. Prehospital and Disaster Medicine 36, 4 (Aug. 2021), 373–374. https://doi.org/10.1017/S1049023X21000649

[32] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. arXiv:2001.06684 [cs, stat] (April 2020). http://arxiv.org/abs/2001.06684 arXiv: 2001.06684.