Practitioners' Perceptions of Data Quality Requirements in Machine Learning-Enabled Systems: An Exploratory Survey

Kevin Pontes
Federal University of Ceará
Itapajé, Brazil
kevinpontes@alu.ufc.br

Arthur Cordeiro
Federal University of Ceará
Itapajé, Brazil
arthur.cordeiro32@alu.ufc.br

Jessica da Silva
Federal University of Ceará
Quixadá, Brazil
jessicasilva2696@gmail.com

Jhefferson Souza
Federal University of Ceará
Itapajé, Brazil
jhefferson.souza@alu.ufc.br

Emanuel Ávila
Federal University of Ceará
Fortaleza, Brazil
emanuel.pires@alu.ufc.br

Marcelo Martins
Federal University of Ceará
Quixadá, Brazil
mmartins@ufc.br

Hugo Villamizar
fortiss GmbH
Munich, Germany
guarinvillamizar@fortiss.org

Anderson Uchôa*
Federal University of Ceará
Itapajé, Brazil
andersonuchoa@ufc.br

Abstract

Data quality affects the performance, reliability, and trustworthiness of machine learning (ML)-enabled systems, but empirical evidence on how practitioners treat it as a requirement remains limited. This paper reports an exploratory survey with 56 ML practitioners from seven countries, investigating how they perceive, prioritize, evaluate, and manage data quality requirements. We analyzed closed-ended and Likert-scale questions using descriptive statistics and bootstrapped confidence intervals, and open-ended responses through qualitative coding. Our results show that participants associate data quality mainly with intrinsic characteristics, especially consistency, completeness, and reliability. They also prioritize accuracy, reliability, consistency, credibility, and completeness, while governance- and lifecycle-oriented characteristics, such as compliance, traceability, and currentness, receive lower priority. Data quality is mostly addressed during data collection and preparation and evaluated indirectly via model performance metrics, whereas continuous monitoring, centralized documentation, and systematic discussion are less frequent. Participants also report challenges with incomplete data, inconsistent sources, outdated data, limited validation tools, and insufficient organizational support. These findings suggest that data quality is recognized as critical, but is often managed through fragmented, performance-oriented, and team-dependent practices. We discuss implications for Requirements Engineering, including the need to explicitly elicit, specify, validate, and monitor data quality requirements across the ML lifecycle.

Keywords

ML-enabled systems; Data Quality Requirements; Requirements Engineering; Practitioners' Perceptions

1 Introduction

Machine Learning (ML)-enabled systems are increasingly deployed in domains where reliability, robustness, and trustworthiness are critical. Unlike traditional software, their behavior is strongly shaped by data, making data quality a central concern for system correctness [35, 36]. Poor data quality can degrade model performance, introduce bias, and lead to unreliable or unsafe decisions, ultimately undermining stakeholder trust [25].

Despite its importance, data quality in ML systems remains difficult to define, operationalize, and manage. It is commonly described through attributes such as *accuracy*, *consistency*, *completeness*, and *currentness* [12], but these attributes are context-dependent and often tied to system-level goals. Moreover, data quality issues are not confined to preprocessing; they may arise across the ML lifecycle, from data collection and integration to deployment and monitoring [15, 17]. Thus, data quality should be understood as a continuous engineering concern rather than a one-time preparation task.

Prior work has shown that data-related problems are pervasive in ML systems and often more difficult to address than model-related issues [2]. At the same time, studies report that data work is frequently undervalued, leading teams to handle quality issues reactively and inconsistently [31]. In parallel, Requirements Engineering (RE) research argues that data-related concerns should be treated as non-functional requirements (NFRs), enabling systematic specification, validation, and traceability [11, 37].

This creates a gap between the recognized importance of data quality and the way it is commonly handled in practice. Decisions about data quality are often implicit, local, and driven by immediate needs, such as improving model performance, rather than by explicit quality goals. This gap is relevant because data quality decisions depend on domain risks, system objectives, and project constraints. For instance, emphasizing performance metrics may overlook traceability or compliance, while focusing on completeness may conflict with currentness in dynamic environments. Yet, we still lack empirical evidence on how practitioners navigate these decisions in practice, particularly how they *perceive*, *prioritize*, and *manage* data quality as a requirement.

To address this gap, we conducted an exploratory survey with 56 ML practitioners from seven countries, covering diverse roles and experience levels. Our analysis examines: (i) how practitioners perceive and prioritize data quality characteristics; (ii) how data quality is incorporated and evaluated across the ML lifecycle; and (iii) the technical and organizational challenges faced in managing data quality. Our results reveal a systematic gap between perception and practice in how data quality is handled in ML-enabled systems. We summarize our main findings as follows:

- **Data quality is primarily associated with intrinsic characteristics**, in which practitioners prioritize attributes such as *accuracy* (92.9%), *reliability*, *consistency*, and *credibility*

*Also with North Carolina State University, Raleigh, United States.

(83.9% each), while governance- and lifecycle-oriented attributes, such as *currentness* (44.6%), *compliance* (53.6%), and *traceability* (55.4%), receive less emphasis;
- **Data quality is addressed early but weakly sustained across the lifecycle**, in which practitioners mainly focus on collection/preparation and evaluate quality through model performance metrics, while continuous monitoring and centralized documentation are less common;
- **Evaluation practices are predominantly performance-driven**, in which practitioners rely on model metrics, manual inspection, and validation datasets to assess data quality, with limited use of explicit data quality metrics or requirement-driven evaluation approaches; and
- **Data quality challenges are persistent**, with practitioners reporting issues such as incomplete data (70%), cross-source inconsistencies (62%), outdated data (50%), and limited validation tools and organizational support.

2 Background and Related Work

ML is an approach to learning algorithms (called models) from data [41]. Unlike deterministic systems, where developers explicitly define decision logic, ML derives decision-making procedures from data, often resulting in difficult-to-understand models. This makes evaluation and accountability more challenging when these systems behave unexpectedly [5, 16]. When ML models are integrated into applications, which we refer to as ML-enabled systems in this paper, humans may be exposed to automated decisions, sometimes without being aware that ML is used. In turn, data quality refers to the degree to which data satisfy the requirements of their intended use. Established standards such as ISO/IEC 25012 [12] conceptualize data quality as a multidimensional concept, including characteristics such as accuracy, completeness, consistency, currentness, and traceability. In addition, ISO/IEC 5259 [13] discusses data quality in analytics and AI/ML contexts, reinforcing the relevance of data quality as an engineering concern for ML-enabled systems. Table 1 lists the data quality characteristics used in this study [12].

Table 1: Data quality characteristics investigated in this study

Characteristic	Description
Accuracy	Refers to the extent to which data correctly represent the real-world values, events, or entities they are intended to describe.
Completeness	Refers to the extent to which all necessary data are available and missing values do not compromise the intended use of the dataset.
Consistency	Refers to the extent to which data remain coherent and non-contradictory across records, datasets, formats, or data sources.
Credibility	Refers to the extent to which data can be trusted based on their source, collection process, and perceived reliability.
Currentness	Refers to the extent to which data are up to date and available within a time frame suitable for the intended ML task.
Accessibility	Refers to the extent to which data can be accessed and used by authorized stakeholders when needed.
Compliance	Refers to the extent to which data handling follows applicable rules, standards, regulations, or organizational policies.
Reliability	Refers to the extent to which data can be depended upon to support stable and trustworthy ML outcomes.
Efficiency	Refers to the extent to which data can be processed, stored, and used without unnecessary computational cost or delay.
Traceability	Refers to the extent to which the origin, transformations, versions, and use of data can be tracked throughout the lifecycle.
Understandability	Refers to the extent to which data, their meaning, structure, and context can be understood by practitioners and stakeholders.
Availability	Refers to the extent to which data are available for use when required by development, training, testing, or operation activities.
Recoverability	Refers to the extent to which data can be restored after loss, corruption, or failure.

Data quality is relevant across the ML development lifecycle. During data collection, teams need to ensure that data are relevant, representative, and collected from reliable sources. During preparation, practitioners commonly perform cleaning, integration, normalization, transformation, balancing, and outlier handling. During model training and validation, data quality affects model performance metrics, such as accuracy, precision, recall, and F1-score. After deployment, continuous monitoring may be necessary to detect distribution shifts, inconsistencies, or performance degradation over time [17, 36]. Therefore, data quality should not be treated as a one-time concern restricted to preprocessing, but as a continuous concern from conception to operation. Despite its importance, managing data quality requirements in ML-enabled systems remains challenging. Practitioners may face several issues such as inconsistent data from multiple sources, incomplete datasets, outdated records, lack of standardization, and difficulty defining objective metrics for data quality evaluation [8]. They may also face process-related and organizational challenges, such as insufficient documentation, limited tool support, lack of continuous validation practices, and limited collaboration with other teams [14, 24].

Engineering Challenges and Data Issues in ML Systems. Prior studies identify data as a major source of complexity in ML-enabled systems. Sculley et al. [32] showed that ML systems accumulate technical debt due to data dependencies and undisciplined data management. Kalinowski et al. [15], in a large-scale international survey, found that data-related problems appear across all phases of the ML lifecycle, from data collection to model monitoring. Similarly, Amershi et al. [2] report that data quality issues are a significant class of engineering problems in industrial ML systems, often requiring more effort than model-related issues. From an RE perspective, Horkoff [11] and Vogelsang and Borg [37] further argue that data-related concerns are frequently underspecified and rarely treated as explicit NFRs in ML projects.

Data Quality Issues and Data Smells. Another line of research examines low-quality data through *data smells*, a concept analogous to code smells in software engineering. Foidl et al. [8] introduced a catalog of 36 data smells grouped into *Believability*, *Understandability*, and *Consistency Smells*, which may increase technical debt and data-induced failures in AI-based systems. Similarly, Shome et al. [34] analyzed 25 public datasets and identified 14 recurring data smells, showing that these problems also appear in benchmark datasets used to train ML models. While these studies help operationalize observable data quality problems, they focus mainly on identifying and classifying quality issues, rather than on how practitioners reason about data quality as a requirement.

Data Quality as Requirements and Development Practices. The studies most directly related to our work examine data quality from requirements, governance, and development-process perspectives. Gudivada et al. [10] offered an early characterization of data quality issues in big data and ML, arguing that traditional data quality approaches are insufficient for the scale and heterogeneity of modern ML pipelines. Sambasivan et al. [31] documented the phenomenon of *data cascades*, showing that data work is often undervalued compared with model development, leading teams to address quality issues reactively. Priestley et al. [28] further argue that classical data quality frameworks, such as the *fitness-for-use* view, are insufficient for contemporary ML contexts and fail to capture emerging requirements.

Our contribution. While the importance of data quality for ML-enabled systems is already well established, we provide exploratory evidence on how practitioners translate this concern into requirements-related decisions. We examine which data quality characteristics they prioritize, how they evaluate, and where gaps emerge between recognizing data quality as critical and managing it through explicit requirements practices. This complements prior work on data quality issues, data smells, and ML engineering challenges by focusing on practitioners' perceptions of data quality as a requirement across the ML lifecycle.

3 Study Design

To define our goal and research questions (RQs), we use the Goal-Question-Metric (GQM) template [3]. Our goal is: **analyze** data quality requirements in ML-enabled systems; **for the purpose of** understanding how practitioners perceive, prioritize, evaluate, and manage data quality throughout the development of ML-enabled systems; **with respect to** perceived importance, development practices, and challenges; **from the viewpoint of** practitioners involved in the development of ML-enabled systems; **in the context of** an international survey spanning seven countries.

RQ_1: *How do practitioners perceive and prioritize data quality characteristics in ML-enabled systems?* – RQ_1 investigates how practitioners understand data quality as a requirement, and which characteristics they consider more relevant. This helps reveal how data quality is conceptualized by practitioners and whether certain quality concerns are systematically prioritized over others.

RQ_2: *How do practitioners evaluate and incorporate data quality throughout the ML development lifecycle?* – RQ_2 examines the practices adopted to evaluate, document, communicate, and monitor data quality within ML projects. In particular, it investigates at which stages of the lifecycle data quality is addressed, how its impact on model behavior is measured, and how quality-related concerns are communicated across teams. By answering RQ_2, we can characterize whether data quality is treated mainly as an upfront concern during data collection and preparation, or whether it is continuously managed across the development lifecycle.

RQ_3: *What challenges do practitioners face when ensuring data quality in ML projects?* – RQ_3 identifies the obstacles that hinder the effective treatment of data quality requirements in ML projects. These challenges may include issues related to data sources, incomplete or outdated datasets, lack of standardization, difficulty in defining objective metrics, limited tool support, insufficient documentation, weak version control practices, and limited cross-team collaboration. By answering RQ_3, we can shed light on the practical barriers that prevent data quality from being treated as a formal requirement in ML-enabled systems.

3.1 Survey Design

To answer our RQs, we chose to use the survey method. To support the design and execution of the survey, we followed the guidelines proposed by Linåker et al. [20]. We describe each step as follows.

Step 1: Identifying the target population and sampling. Based on our study goal, our target population consists of practitioners with experience in the development of ML-enabled systems, including developers, data scientists, data engineers, ML engineers, and professionals who have worked directly or indirectly with ML projects. Since we focus on perceptions and practices related to data quality requirements, we targeted participants with experience in data processing, data preparation, requirements engineering, or ML model development. We recruited participants through specialized groups related to ML, posts on the researchers' LinkedIn profiles, and direct contacts previously established by the authors. This recruitment strategy allowed us to reach professionals with different backgrounds, roles, and levels of experience in ML projects.

Step 2: Designing the survey instrument. We designed and administered the survey using *Google Forms*. The survey comprised 22 questions, including closed-ended, multiple-choice, Likert-scale, and open-ended items, organized into four sections (see Table 2). These sections progressively collected information on participants' profiles, experience with data quality, perception of data quality as a requirement, and implementation of such requirements in practice. At the beginning of the survey, we presented the study purpose, the researchers involved, and the Free and Informed Consent Form (FICF).[1] Participants were informed that participation was voluntary and anonymous, and that they could withdraw at any time without any penalty. The survey did not request sensitive personal data. The instrument collected both quantitative and qualitative evidence through structured responses and open-ended questions that allowed participants to elaborate on their experiences and perceptions. We describe the four sections as follows.

The first section comprises seven questions (Q1–Q7) to characterize participants' demographic and professional profiles, including age, location, gender, education level, role in ML-enabled systems, seniority, and number of ML projects in which they had been involved. The second section comprises one matrix question (Q8) designed to assess participants' familiarity with data processing activities commonly performed in ML projects, such as data cleaning, normalization, outlier detection, data integration, data transformation, data validation, automation of data processing, monitoring of data quality, use of libraries or tools, and dataset splitting. Participants reported their skill levels using a 5-point Likert scale [19], ranging from *Very high* to *Very low*, with an additional *Not applicable* option. This section helped us understand whether participants' perceptions were grounded in practical experience with data quality-related activities.

The third section comprises eight questions (Q9–Q16) designed to investigate how participants conceptualize and prioritize data quality in ML-enabled systems. We used ISO/IEC 25012 [12] to select the survey characteristics because it provides a well-established data quality vocabulary. Although ISO/IEC 5259 [13] addresses analytics and AI/ML contexts, our instrument focuses on general data quality characteristics applied to ML-enabled systems, not the full set of ML-specific requirements. We also limited the set to keep the survey manageable and reduce ambiguity. For example, we did not include precision separately because, in ML, the term is commonly used as a model performance metric; instead, we used accuracy to refer to data correctness. Other characteristics less aligned with our research questions, such as portability and confidentiality, were also left out. First, we asked participants to report the first five words

[1]The study complies with ethical and privacy standards established by the Brazilian General Data Protection Law (Law No. 13,709/2018).

Table 2: Survey Questions

ID	Question	Type
	Free and Informed Consent Form (FICF)	
	I – General Participant Characterization	
Q1	How old are you?	C
Q2	In which country or state do you currently reside?	O
Q3	What's your gender?	C
Q4	What is your level of education?	C
Q5	Which role best describes your current activities in the development of machine learning-enabled systems?	C
Q6	How senior is your role in the team?	C
Q7	How many machine learning-based projects have you been involved in? Please provide your best estimate.	O
	II – Characterization of Data Quality Experience in ML-enabled Systems	
Q8	How would you rate your skill level in relation to the following activities that involve the data processing stage in a machine learning-enabled system?	L
	III – Perception of Data Quality as a Requirement for ML-enabled Systems	
Q9	What are the first five words that come to mind when you think of data quality as a requirement in machine learning projects?	O
Q10	What experience do you have with Requirements Engineering activities in your projects? Have you ever been involved in defining, analyzing or documenting requirements? If so, can you share a little about what that process was like?	O
Q11	Based on your experience, how **important** are the following data quality characteristics for machine learning-enabled systems?	L
Q12	For those characteristics that you considered important, please describe the reason for your rating. (Optional)	O
Q13	Regarding data quality characteristics, how would you characterize the degree of **priority** of each characteristic for machine learning-based system projects?	L
Q14	For those characteristics that you considered to be a priority, please describe the reason for your classification. (Optional)	O
Q15	Based on your experience, how do you balance different data quality characteristics, such as accuracy, understandability and computational efficiency, when developing machine learning-enabled systems?	O
Q16	How can data version control impact data quality characteristics in machine learning projects?	C
	IV – Implementation of Data Quality Requirements in the Development Process	
Q17	Considering the machine learning projects you have participated in, how is data quality incorporated into the development process?	M
Q18	How do you measure the impact of data quality on the performance of the machine learning model?	M
Q19	How often are data quality requirements formally discussed during the development of machine learning projects?	C
Q20	How do you document and communicate data quality issues within your team and with other stakeholders?	M
Q21	What are the main challenges in guaranteeing data reliability in machine learning projects?	M
Q22	How often do you receive support from other teams (e.g., data team, data scientists) to deal with data quality issues?	C

C: Closed; M: Multiple; O: Open; L: Likert

they associate with data quality as a requirement and to describe their experience with RE activities. Before the Likert-scale questions, we provided definitions of the data quality characteristics listed in Table 1. Participants then assessed these characteristics along two dimensions: *importance*, using a 5-point Likert scale [19] from *Not important* to *Very important*; and *priority*, using a 5-point Likert scale from *Not a priority* to *Essential*. We also asked them to justify their ratings, describe how they balance trade-offs among data quality attributes, and indicate how data version control may impact data quality in ML projects. This section addresses RQ_1.

The fourth section comprises six questions (Q17–Q22) designed to investigate how data quality requirements are incorporated into the development process of ML-enabled systems. We asked participants how data quality is considered during development, including assessment during data collection and preparation, continuous monitoring across the model lifecycle, and validation through test sets before training. We also asked how they measure the impact of data quality on model performance, considering A/B testing, performance metrics (e.g., precision and recall), and manual review of results. In addition, we investigated how often data quality requirements are formally discussed, how data quality issues are documented and communicated within teams and with stakeholders, and what challenges practitioners face in ensuring data reliability. Finally, we asked how often they receive support from other teams, such as data teams or data scientists. This section addresses RQ_2 and RQ_3, capturing both the practices adopted throughout the ML lifecycle and the challenges faced by practitioners.

Step 3: Validating and administering the survey. Before administering the survey, we conducted a pilot test [21] with two specialists to evaluate the clarity, consistency, and relevance of the questions: one data engineer with a master's degree in Informatics and one Software Engineering graduate pursuing a PhD in Computer Science. Their feedback led to refinements in the instrument, including wording adjustments, removal of redundant questions, and inclusion of new items to better capture data quality requirements in ML-enabled systems. The pilot responses were used only to refine the instrument and were not included in the main study. After the pilot, we made the survey publicly available online in Portuguese and English using *Google Forms*. The survey targeted professionals with current or previous experience in the development of ML-enabled systems. We disseminated it through ML-related groups, the researchers' LinkedIn profiles, and direct contacts. The survey remained available from February 13 to March 1, 2025. During this period, we collected 56 valid responses.

Step 4: Conducting data analysis. To address our RQs, we combined quantitative and qualitative analyses according to the type of data collected. For closed-ended and Likert-scale questions, we complemented descriptive statistics with bootstrapped confidence intervals. Since the distribution of professionals involved in ML-enabled systems is unknown, bootstrapping is suitable for estimating uncertainty without assuming a specific population distribution [22, 39]. This strategy has also been used in survey-based SE research [1, 38]. For each RQ, we considered the number of valid responses as the empirical sample size n. We generated $S = 1000$ bootstrap resamples with replacement, each with size n, and computed the corresponding statistic for each resample [7]. The resulting bootstrap distribution was used to estimate confidence intervals. Responses marked as *Not applicable* were excluded. We adopted 1,000 resamples, which has been reported as sufficient for meaningful estimates in empirical SE contexts [18].

For $\mathbf{RQ}_1$, we analyzed Q9–Q16 to examine how participants conceptualize, evaluate, prioritize, and balance data quality characteristics. We grouped semantically similar terms from Q9, used Q10 to contextualize participants' RE experience, and analyzed Q11–Q14 through response distributions, medians, high-level ratings, and Spearman's correlation between *importance* and *priority* [40]. We qualitatively analyzed Q15 using Grounded Theory procedures [4, 6] and examined Q16 through frequency distributions. For $\mathbf{RQ}_2$, we analyzed Q17–Q20 using frequencies and percentages to characterize how data quality is incorporated, measured, discussed, documented, and communicated during ML development. For $\mathbf{RQ}_3$, we analyzed Q21–Q22 to identify data quality challenges and cross-team support. Finally, we grouped participants by role, seniority, and ML experience to explore differences across profiles.

4 Results and Discussion

Overview of the participants' characterization. Participants were primarily between 25–34 years old (60.7%), followed by 35–44 (16.1%), 18–24 (7.1%), and 45–54 (7.1%). Most of the participants were male (80.4%), with 19.6% identifying as female. Regarding geographic distribution, the sample is dominated by participants from Brazil (42, 75%), followed by the United States (5, 8.9%), Germany (4, 7.1%), and France (2, 3.6%), with limited representation of one participant from other countries (Ireland, Colombia, and China). In terms of experience, participants were mostly junior (37.5%), followed by mid-level (28.6%), senior (26.8%), and interns (7.1%). On average, participants reported experience in 5.48 ML projects. Participants also showed a high level of academic qualification: Doctorate (21.4%), Doctoral students (32.1%), Master's degree holders (8.9%), Master's students (12.5%), Specialization (8.9%), Undergraduate (14.3%), and High School (1.8%). Regarding professional roles, most participants were Data Scientists (46.4%) or Developers (30.4%), with smaller representation from Data Engineers (7.1%), ML Engineers (3.6%), Managers (3.6%), Researchers, and DevOps Engineers (1.8%, each).

Figure 1 shows participants' self-assessed skills in ten data processing activities (Q8). We observed higher proficiency in core preprocessing tasks, especially data cleaning, train/test split, normalization/standardization, and the use of data processing libraries. In contrast, lower confidence is reported for production-oriented activities, especially pipeline automation and production monitoring. This suggests that participants were more familiar with data preparation than with production-oriented data lifecycle activities.

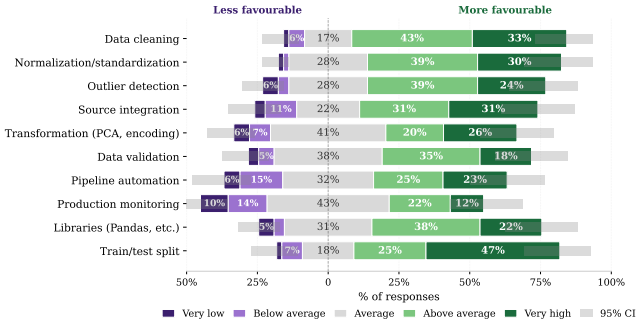

Figure 1: Self-assessed data processing skills

4.1 Perception of Data Quality (RQ_1)

To answer $\mathbf{RQ_1}$, we considered five complementary aspects: (i) *conceptualization*, i.e., the concepts and terms practitioners associate with data quality (Q9); (ii) *experience with RE*, i.e., practitioners' previous experience in analyzing and documenting system requirements (Q10); (iii) *importance and prioritization*, i.e., how practitioners evaluate and rank data quality characteristics (Q11–Q14); (iv) *alignment between importance and priority*, i.e., the relationship between perceived relevance and practical prioritization decisions; and (v) *versioning perception*, i.e., how practitioners perceive the impact of data version control on data quality (Q16).

The most frequent words associated with data quality. Participants were asked to provide the first five words that came to mind when thinking of data quality in ML projects (Q9). This question aimed to capture the immediate associations and key concepts practitioners link to data quality. Figure 2 shows the frequency distribution of the reported terms. The x-axis represents the position of the words in the order provided by the participants (from 1 to 5), corresponding to the sequence in which participants listed them.

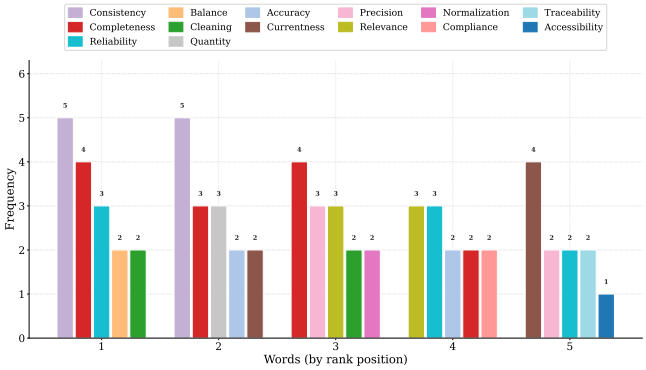

Figure 2: Top-5 words by rank frequency.

At *ranking* = 1, *Consistency* appears as the most frequently mentioned term (five mentions), followed by *Completeness* and *Reliability*, both with four mentions. This indicates that participants primarily associate data quality with the internal coherence and dependability of data. At *ranking* = 2, the distribution becomes more diverse, with *Consistency* again appearing frequently (five mentions), followed by *Completeness* and *Quantity* (three mentions each), and *Accuracy* and *Currentness* appearing with lower frequencies. This suggests that participants also relate data quality to correctness, volume, and temporal aspects.

At *ranking* = 3, *Completeness* emerges as the most frequent term (four mentions), followed by *Precision* and *Relevance* (three mentions each). These terms reflect both intrinsic data properties and aspects related to usefulness and interpretation. At *ranking* = 4, the distribution remains diverse, with *relevance* and *reliability* appearing more frequently (three mentions each), followed by several other characteristics with lower frequencies. This suggests that additional contextual and application-dependent aspects of data quality become more prominent. Finally, at *ranking* = 5, *Currentness* appears as the most frequent term, although overall mentions remain sparse and distributed across multiple characteristics.

> **Finding 1:** Participants primarily associate data quality with intrinsic properties such as *Consistency*, *Completeness*, and *Reliability*, while other characteristics appear more frequently in later ranking positions.

Experience with Requirements Engineering activities. Q10 asked participants about their experience defining, analyzing, or documenting requirements. Among those engaged with RE, responses indicate diverse involvement, including full lifecycle participation, stakeholder coordination, and the use of specific techniques and artifacts. For brevity, the full categorization is available in our replication package [27]. Participants described active involvement in RE. P31 reported having *"participated in all phases"*, while P53 explained that, as a tech lead, the work involves *"coordinating with stakeholders (POs, CPO, etc) to understand business needs"*. Others emphasized artifacts and iterative practices: P9 mentioned *"formal*

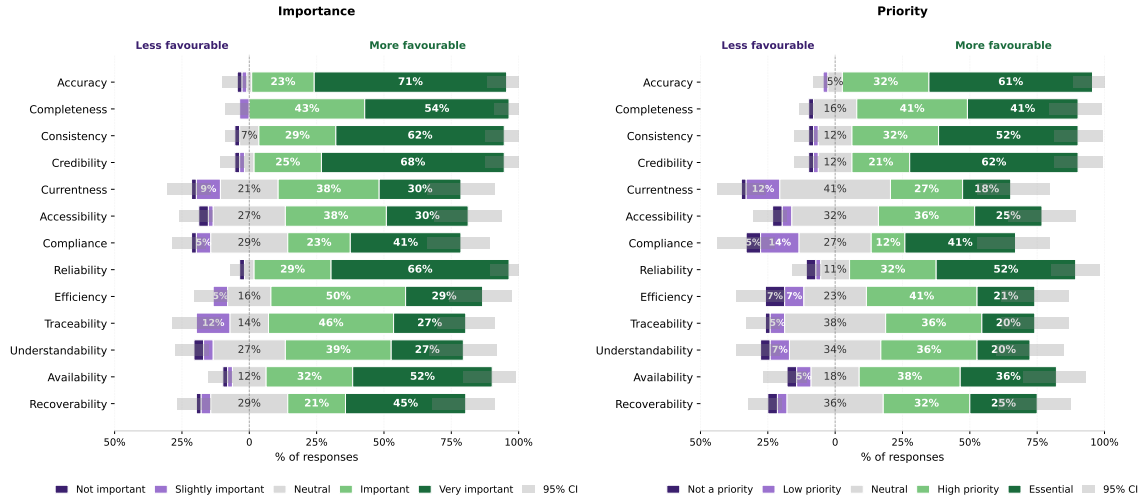

Figure 3: Distribution of perceived importance and priority of data quality characteristics.

documentation of requirements in detailed documents", including functional and non-functional specifications and use cases, while P30 noted that *"we always hold a meeting with the client where they identify the scenarios and cases they want to achieve"*. Participants also reflected on RE's value and challenges. P27 described it as *"the crucial step in defining a good scope, avoiding friction between the team and the client"*, whereas P37 called it *"a costly process, but if done well it's extremely useful"*. Overall, engaged practitioners portray RE as an adaptive process of aligning stakeholder needs, technical feasibility, and business context.

Analysis of the perceived importance and priority of data quality characteristics. We asked participants to evaluate 13 data quality characteristics in terms of their perceived *importance* and *priority*. Figure 3 overviews the distribution of responses.

Regarding *importance* (Q11), responses were concentrated in the *important* and *very important* categories. *Completeness* received the highest combined positive evaluation (96.4%), followed by *Accuracy* and *Reliability* (94.6% each), *Credibility* (92.9%), and *Consistency* (91.1%). In the *very important* category, *Accuracy* stood out (71.4%), followed by *Credibility* (67.9%), *Reliability* (66.1%), and *Consistency* (62.5%). These results suggest that participants attribute high importance to characteristics related to correctness, dependability, and internal coherence. In contrast, *Compliance* (64.3%), *Understandability* (66.1%), and *Recoverability* (66.1%) received lower positive evaluations, suggesting a more context-dependent perception.

To complement these results, we asked participants to justify their importance ratings (Q12). Participants framed *Completeness*, *Accuracy*, *Reliability*, *Credibility*, and *Consistency* as foundational for trustworthy ML outcomes, often referring to the *"garbage in, garbage out"* principle. For instance, P8 stated that *"if the data is a mess, the model will only produce sophisticated garbage"*, while P18 argued that inconsistent, incomplete, or inaccurate data can compromise *"the entire result from your ML pipeline"*. At the same time, participants highlighted that importance is not absolute, but depends on project context and lifecycle stage. P12 distinguished characteristics that are *"non-negotiable from the initial moment"* from those that can be addressed later, such as documentation and accessibility. Similarly, P17 noted that importance depends on *"the scope of the project and the learning objective"*.

> **Finding 2:** Participants perceive *Completeness*, *Accuracy*, *Reliability*, *Credibility*, and *Consistency* as foundational for trustworthy ML outcomes, while other characteristics appear more context-dependent.

Regarding *priority* (Q13), *Accuracy* shows the highest combined proportion of *high priority* and *essential* ratings (92.9%), followed by *Reliability*, *Consistency*, and *Credibility* (83.9% each), and *Completeness* (82.1%). Conversely, *Currentness* receives the lowest combined priority score (44.6%), followed by *Compliance* (53.6%), *Understandability* and *Traceability* (55.4% each), suggesting that governance- and context-related characteristics are less frequently prioritized in practice. To provide additional insight, participants described the reasoning behind their classifications (Q14). The responses show that highly prioritized characteristics are associated with the reliability of results, representativeness of real-world phenomena, and production readiness. For instance, P31 stated that poor-quality data makes *"all results [...] dubious and unreliable"*, while P27 argued that consistency and completeness indicate how accurately data represent real-world scenarios. P32 also emphasized that decisions based on data require information to be *"accurate and consistent"*.

Participants also described prioritization as constrained by project needs, domain risks, and available resources. For instance, P20 noted that several characteristics may be critical, but prioritizing them *"will depend too much on the project in question"*. P21 explained that acceptable precision depends on error tolerance, compliance becomes critical in sensitive domains such as healthcare or banking, and efficiency depends on real-time requirements.

> **Finding 3:** When prioritizing data quality, participants mainly emphasize correctness- and coherence-related characteristics, while governance- and lifecycle-oriented characteristics are less frequently treated as immediate priorities.

Analysis of correlation between perceived importance and priority. To examine the relationship between perceived importance (Q11) and priority (Q13), we computed correlations for each data quality characteristic. We defined $\mathbf{H_1}$**:** there is a strong correlation between perceived importance and priority for a given data quality characteristic. The null hypothesis is $\mathbf{H_0}$**:** there is no strong correlation between perceived importance and priority for that characteristic. We applied the Shapiro-Wilk test [33] to assess data normality and observed that the data are not normally distributed. Thus, we used Spearman's rank correlation coefficient [40] to measure the association between importance and priority. We considered a confidence level of 95% ($p < 0.05$). We categorized the correlation values following Salkind [30]: very strong ($0.80 \leq |\rho| \leq 1.00$), strong ($0.60 \leq |\rho| < 0.80$), moderate ($0.40 \leq |\rho| < 0.60$), weak ($0.20 \leq |\rho| < 0.40$), and weak or no relationship ($0.00 \leq |\rho| < 0.20$). Table 3 shows the correlation results.

Table 3: Spearman's ρ between importance and priority

Data quality characteristic	ρ	p-value	Strength
Compliance	0.832	< 0.001	Very strong
Currentness	0.714	< 0.001	Strong
Traceability	0.693	< 0.001	Strong
Accuracy	0.637	< 0.001	Strong
Availability	0.633	< 0.001	Strong
Consistency	0.589	< 0.001	Moderate
Completeness	0.585	< 0.001	Moderate
Understandability	0.584	< 0.001	Moderate
Credibility	0.573	< 0.001	Moderate
Recoverability	0.571	< 0.001	Moderate
Reliability	0.544	< 0.001	Moderate
Accessibility	0.542	< 0.001	Moderate
Efficiency	0.526	< 0.001	Moderate

We can observe that all data quality characteristics exhibit statistically significant correlations ($p < 0.001$), ranging from moderate to very strong. In particular, *Compliance* shows a very strong correlation ($\rho = 0.832$), followed by *Currentness* ($\rho = 0.714$), *Traceability* ($\rho = 0.693$), *Accuracy* ($\rho = 0.637$), and *Availability* ($\rho = 0.633$), indicating that participants tend to prioritize these characteristics when they perceive them as important. However, the majority of characteristics exhibit only moderate correlations, such as *Completeness* ($\rho = 0.585$), *Consistency* ($\rho = 0.589$), *Credibility* ($\rho = 0.573$), and *Reliability* ($\rho = 0.544$). This suggests that, although these characteristics are widely perceived as important, their prioritization is not as strongly aligned and may depend on contextual constraints, such as project requirements, cost, or implementation effort.

Interestingly, even characteristics with strong correlations (e.g., *Compliance*) showed gaps between importance and priority levels, indicating that correlation alone does not capture differences in absolute ratings. Based on these results, we confirm H_1 for five characteristics (*Compliance*, *Currentness*, *Traceability*, *Accuracy*, and *Availability*), which presented strong or very strong correlations. For the remaining eight characteristics, we reject H_1 and retain H_0, since they presented moderate correlations rather than strong ones. In summary, perceived importance is generally associated with prioritization, but this association varies across characteristics and does not eliminate the influence of practical constraints.

Finding 4: Importance and priority are significantly correlated, but this does not imply full alignment in practice, since prioritization may still depend on constraints and trade-offs.

Balancing data quality characteristics. We also asked participants how they balance different data quality characteristics, such as accuracy, understandability, and computational efficiency, in ML-enabled systems (Q15). We report the most prominent categories here; the full categorization is available in our replication package [27]. Practitioners describe this balance as a *context-driven decision process*, shaped by domain, system constraints, and expected risks. Decisions may *"depend on the problem to be solved"* (P25), on how *"the application needs to respond to the user"* (P3), or on *"business requirements"* (P51). In critical domains, accuracy and reliability tend to prevail, while in real-time or resource-constrained systems efficiency becomes more central (P45). Some participants also express an *explicit hierarchy*, usually placing accuracy before understandability and computational efficiency (P2, P14).

A second theme is the recognition of *trade-offs and inherent tensions* as a structural condition of ML development. Participants acknowledge that improving one characteristic may require sacrificing another, such as *"settle for a slightly less precise but much faster model"* (P56) or accepting incomplete but more reliable data (P16). To handle these tensions, they report *balancing strategies and reconciliation techniques*, including model selection (P55), feature engineering and dimensionality reduction (P6), and iterative refinement (P20). However, some responses indicate an *absence of systematic consideration*, where decisions are driven by available data rather than explicit trade-off reasoning, as illustrated by P44: *"we simply work on the data we have available"*. Overall, balancing data quality appears to be *pragmatic, iterative, and rarely formalized*, despite being widely recognized as essential.

On data version control. We asked participants how data version control impacts data quality characteristics (Q16). Most participants (91.1%) selected the option stating that version control *ensures data consistency and traceability over time, allowing dataset changes to be documented and verified.* A small share selected that it *eliminates the need to document changes* (5.4%), while 1.8% selected that it *increases data availability without requiring consistency checks between versions.* Overall, this suggests strong agreement that versioning supports data integrity and traceability in ML projects.

Finding 5: Balancing data quality characteristics is mainly *context-dependent* and *trade-off-driven.* Practitioners adapt priorities to domain risks, system goals, and resource constraints, often using accuracy as an anchor, while relying on practical strategies that are rarely formalized.

4.2 Data Quality in ML Development ($\mathbf{RQ_2}$)

To answer $\mathbf{RQ_2}$, we analyzed four aspects: (i) *adoption of practices*, i.e., how data quality is incorporated into the development process (Q17); (ii) *impact measurement*, i.e., how practitioners evaluate the effect of data quality on model performance (Q18); (iii) *formal discussion*, i.e., how often data quality requirements are explicitly discussed during development (Q19); and (iv) *communication practices*, i.e., how data quality issues are documented and communicated within teams and with stakeholders (Q20). Figure 4 summarizes practitioners' responses. Since Q17, Q18, and Q20 allowed multiple selections, we report the number of mentions for each option,

since participants could select more than one alternative. For Q19, a single-choice question, we report the proportion of participants.

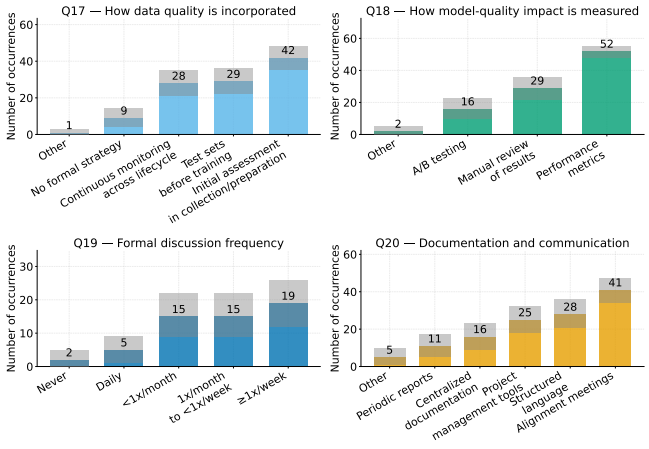

Figure 4: Synthesis of responses (Q17-Q20)

The results for Q17–Q20 indicate that data quality is primarily treated as an upfront concern in ML development, rather than as a continuously managed process. In Q17, *initial assessment in collection/preparation* was the most frequently mentioned practice (42 mentions), followed by *test sets before training* (29 mentions) and *continuous monitoring across the lifecycle* (28 mentions). Only 9 mentions referred to *no formal strategy*. The *other* response, *"I did not directly participate in any"* (P24), indicates limited direct experience rather than an additional practice. Overall, data quality is usually considered, although not always continuously managed.

For Q18, the impact of data quality was mainly assessed through *model performance metrics* (52 mentions), followed by *manual review of results* (29 mentions) and *A/B testing* (16 mentions). The *other* responses reinforce this performance-oriented pattern: P15 emphasized monitoring metrics and confusion matrices, focusing on *"Recall because it is a health-related project"*, while P45 combined performance analysis with *"data exploration analysis before training models"*. Overall, practitioners seem to evaluate data quality mostly through its effects on model behavior, with exploratory analysis acting as a complementary data-centered practice.

Regarding Q19, formal discussions about data quality requirements varied across projects. The most frequent response was *at least once a week* (19 participants, 33.9%), followed by *once a month to less than once a week* and *less than once a month* (15 participants, 26.8% each). Only 5 participants reported *daily* discussions (8.9%), and only 2 participants indicated that such discussions never occur (3.6%). Finally, in Q20, communication practices relied mainly on *alignment meetings* (41 mentions), followed by *structured language* (28 mentions), *project management tools* (25 mentions), and *centralized documentation* (16 mentions). The *other* responses complement this pattern by mentioning informal channels, such as *"Slack"* (P15), embedded development spaces, such as *"In the same MR"* (P21), and structured environments, such as *"Project management tools (Jira, Trello, Asana, and Confluence"* (P42). However, they also reveal cases in which communication was limited or absent, as participants reported *"No experience"* (P48) and *"Only solo projects"* (P49). This suggests that communication about data quality still appears to depend largely on team interaction, while persistent documentation mechanisms are less frequently adopted.

> **Finding 6:** Data quality in ML projects is mainly addressed early and evaluated via model performance metrics, but remains fragmented, with limited monitoring, documentation, and systematic discussion across the lifecycle.

To better understand how data quality practices vary across participants, we combined responses to (Q17–Q20) with profile information collected at the beginning of the survey. We compared responses across three profile dimensions: career stage, role type, and experience with ML projects. We defined the following groups: (a) *early-career practitioners*, including trainees and junior participants (25 respondents); (b) *mid-career practitioners*, including full-level participants (16 respondents); (c) *senior-level practitioners*, including senior participants (15 respondents); (d) *data-oriented practitioners*, including data scientists, data engineers, and ML engineers (32 respondents); (e) *product-oriented practitioners*, including software developers, managers, and researchers (24 respondents); (f) practitioners with *up to five ML projects* (36 respondents); and (g) practitioners with *more than five ML projects* (20 respondents). After defining these groups, we compared their responses to Q17, Q18, Q19, and Q20. Figure 5 shows responses by profile.

Data-oriented roles drive continuous practices. Our results indicate that data-oriented participants are more engaged in activities that extend beyond the initial stages of development. For instance, *continuous monitoring across the lifecycle* was more frequent among data-oriented participants (18%) than product-oriented ones (7%). Similarly, *performance metrics* were more often reported by the former group (32%) than the latter (21%). This suggests that responsibilities related to data handling tend to persist throughout the lifecycle and may depend strongly on specialized data roles.

Data quality is still mostly front-loaded. *Initial assessment during collection/preparation* was the most recurrent practice across profiles, especially among data-oriented participants and those with up to five ML projects (23% each). In contrast, *continuous monitoring across the lifecycle* appeared less frequently, suggesting that data quality is still treated mainly as an upfront concern rather than as a continuously managed requirement. **The assessment relies mainly on model performance metrics.** *Performance metrics* were the dominant strategy for measuring the impact of data quality on model performance, particularly among data-oriented participants (32%) and those with up to five ML projects (34%). *Manual review of results* also appeared as a complementary strategy, especially in the latter group (20%). This indicates that data quality is often operationalized indirectly through model behavior rather than through standalone data quality metrics.

Communication remains informal and weakly standardized. *Alignment meetings* were the most common mechanism for communicating data quality issues, especially among data-oriented participants and those with up to five ML projects (22% each). More structured mechanisms such as *centralized documentation* (up to 10%) and *periodic reports* (up to 6%) were less frequent, suggesting that data quality knowledge is often shared through team-dependent practices rather than persistent artifacts. **Experience may influence formalization.** Participants with up to five ML

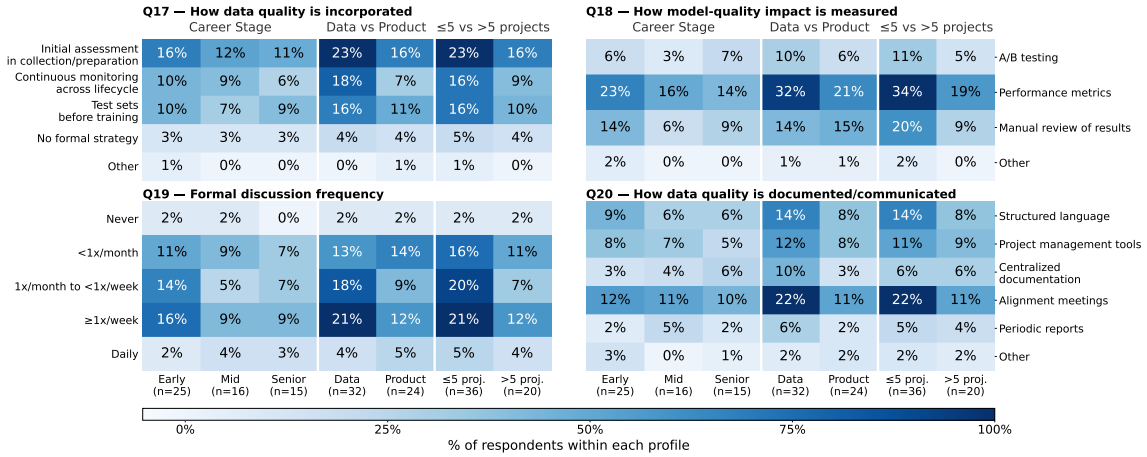

Figure 5: Distribution of responses (Q17–Q20) within each practitioner profile.

projects reported higher use of *performance metrics* (34%) and *alignment meetings* (22%) than those with more than five projects (19% and 11%, respectively). One possible interpretation is that less experienced participants rely more on explicit practices, whereas more experienced ones may rely more on tacit knowledge.

> **Finding 7:** Data quality practices vary by practitioner profile, but remain mostly front-loaded and informally communicated. Data-oriented practitioners report more monitoring and metric-based assessment, while less experienced ones rely more on explicit practices and alignment meetings.

4.3 Data Quality Challenges (RQ_3)

To answer $\mathbf{RQ_3}$, we analyzed two complementary aspects: (i) *challenges identification*, i.e., the main difficulties in ensuring data reliability in ML projects (Q21); and (ii) *organizational support*, i.e., how often practitioners receive support from other teams when dealing with data quality issues (Q22). Figure 6 summarizes practitioners' responses. Since Q21 allowed multiple selections, we report the proportion of participants mentioning each challenge. For Q22, a single-choice question, we report the proportion of participants who select each option.

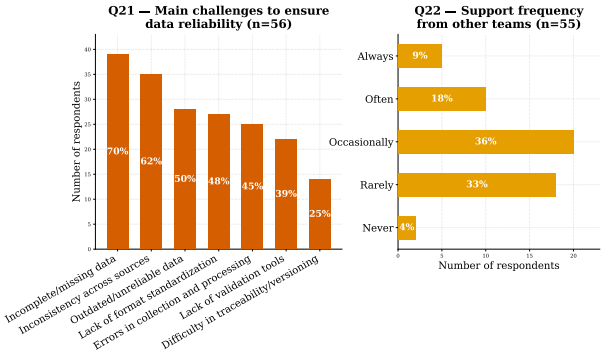

Figure 6: Summarization of responses (Q21-Q22)

Main challenges faced by practitioners. Figure 6 shows that the most frequently reported challenges were *incomplete or missing data* (70%, $n = 39$), *inconsistency across data sources* (62%, $n = 35$) and *outdated or unreliable data* (50%, $n = 28$). Practitioners also reported governance and engineering challenges, such as *lack of format standardization* (48%, $n = 27$), *errors in data collection and processing* (45%, $n = 25$), *lack of validation tools* (39%, $n = 22$), and *difficulty in traceability and versioning* (25%, $n = 14$). The only *other* response mentioned issues related to *data structure* and *data nature* (P15). These results suggest that data quality challenges span data collection, integration, validation, and maintenance, requiring coordination across activities and roles.

Organizational support. We asked participants how often they receive support from other teams (e.g., data teams or data scientists). Figure 6 also shows that most participants receive support only occasionally (36%) or rarely (33%), while fewer reported receiving support often (18%) or always (9%); one participant did not answer this question. This suggests that, although data quality challenges are recurrent, support from specialized teams is not consistently available, reinforcing the need for more integrated and collaborative data quality management.

> **Finding 8:** Participants report recurrent data quality challenges, especially incomplete data, inconsistency across sources, and outdated data, while organizational support is often only occasional or rare.

5 Implications for Requirements Engineering

Treating data quality as a requirements problem, not only a preprocessing concern. Our results show that participants strongly value intrinsic data quality characteristics, especially *Completeness*, *Accuracy*, *Reliability*, *Credibility*, and *Consistency*. However, governance- and lifecycle-oriented characteristics, such as *Currentness*, *Compliance*, *Traceability*, and *Understandability*, received lower priority. This suggests that data quality is recognized as important, but is not always translated into explicit requirements across the ML lifecycle. Requirements Engineering (RE) should therefore make these characteristics visible during elicitation, especially when domain risks, regulations, privacy concerns, or operational constraints make them critical. This reinforces prior work arguing that RE for ML-enabled systems must adapt to data-dependent development settings, where requirements are shaped

not only by software behavior but also by data properties, privacy concerns, and organizational practices [1, 23, 37].

Specifying data quality requirements beyond performance metrics. Participants often reported assessing data quality through model performance metrics. Although useful, these metrics are insufficient as requirements evidence: a model may perform well while relying on data that are outdated, weakly documented, hard to trace, or difficult to audit. Thus, data quality requirements should be specified in assessable terms, including acceptance criteria, dataset assumptions, quality thresholds, validation checks, versioning rules, and monitoring conditions. For researchers and tool builders, this points to RE support that connects data quality requirements to ML pipeline artifacts, such as dataset contracts, quality checks, version histories, and monitoring dashboards. This aligns with prior work showing that production ML systems require explicit data validation and engineering practices beyond model-level evaluation [2, 26]. Dataset documentation approaches, such as datasheets and data cards, also provide useful starting points for connecting data quality requirements to concrete artifacts [9, 29].

Managing data quality requirements as lifecycle and organizational concerns. Our results indicate that data quality practices remain mostly front-loaded: initial assessment during collection and preparation was the most frequent practice, while continuous monitoring, centralized documentation, and structured communication were less frequent. Participants also reported recurrent challenges related to incomplete data, inconsistency across sources, and outdated or unreliable data, often with limited organizational support. These results suggest that data quality requirements should not be managed only at the beginning of development. RE for ML-enabled systems should support continuous validation, change management, stakeholder communication, and clear responsibility assignment across data, development, and business roles. This concern is consistent with prior evidence that undervalued data work can create downstream failures and that data dependencies are a major source of technical debt in ML systems [31, 32].

6 Threats to Validity

We discuss threats to the study validity [40] as follows.

Construct and internal validity. The survey instrument may not fully capture the intended constructs, such as practitioners' perceptions of data quality requirements and their implementation across the ML lifecycle. To mitigate this threat, we designed the questionnaire based on prior work on data quality and requirements engineering [1, 15] and structured it using the GQM [3], aligning the study goal, RQs, and survey questions. We also provided definitions of data quality characteristics before Likert-scale questions and validated the survey through a pilot study with domain experts [21], which led to wording refinements, removal of redundancies, and improved clarity. Moreover, because our instrument was mainly based on ISO/IEC 25012, it may not capture the full set of AI/ML-specific data quality concerns discussed in ISO/IEC 5259. Regarding internal validity, participants' responses may have been influenced by their roles, experience, or organizational context. To reduce this threat, we collected demographic and professional data, conducted complementary analyses across participant profiles, and used structured questions with predefined options and examples to reduce misinterpretation. In addition, the combination of quantitative (descriptive statistics with bootstrapping) and qualitative analysis (Grounded Theory procedures) helped triangulate findings and reduce interpretation bias [4, 6].

Conclusion and external validity. The generalizability of our findings is limited by the sample composition. Although participants came from multiple countries, most responses came from Brazil, and the sample includes many early- and mid-career practitioners, several of them with strong academic backgrounds. Thus, our results should be interpreted as exploratory evidence rather than as representative of the global population of ML practitioners or mature industrial ML settings. Future studies should adopt more balanced samples across countries, industry sectors, organizational contexts, ML maturity levels, and data-oriented roles. Regarding conclusion validity, the relatively small sample size and response variability, especially in open-ended questions, may affect the reliability of the results. To mitigate this, we used bootstrapping with 1,000 resamples to estimate confidence intervals without assuming a specific population distribution [7, 18, 22]. For qualitative data, we followed Grounded Theory procedures [4], including open and axial coding, and conducted the analysis collaboratively to reduce subjectivity and reach consensus. Nevertheless, some nuance may be lost when aggregating qualitative insights into categories.

7 Conclusion and Future Work

This paper investigated how practitioners perceive, prioritize, evaluate, and manage data quality requirements in ML-enabled systems. Based on an exploratory survey with 56 practitioners from seven countries, we found that participants mainly associate data quality with intrinsic characteristics, especially consistency, completeness, reliability, accuracy, and credibility. By contrast, compliance, traceability, currentness, and understandability received lower priority, suggesting that governance- and lifecycle-oriented concerns are often handled according to project context.

The results also show that data quality is frequently recognized as important, but not always managed as an explicit requirement. Participants reported addressing data quality mostly during data collection and preparation, and often evaluating it through model performance metrics. This practice may overlook requirements-level concerns such as outdated data, weak traceability, limited documentation, auditability, and compliance evidence. Thus, data quality in ML-enabled systems still appears fragmented, performance-oriented, and dependent on local team practices. Future work should investigate how to elicit, specify, monitor, and trace data quality requirements across ML pipelines, as well as replicate this study with larger and more diverse samples and interviews to better understand data quality trade-offs in practice.

Artifact Availability

All artifacts from this study, including datasets, analysis scripts, and documentation, are available in our replication package [27].

Acknowledgments

This research was partially funded by the Brazilian funding agencies FUNCAP (Grant BP6-00241-00276.01.00/25), and CNPq (Grants 201708/2025-6, 403304/2025-3, and 404406/2023-8).

References

[1] Antonio Pedro Santos Alves, Marcos Kalinowski, Görkem Giray, Daniel Mendez, Niklas Lavesson, Kelly Azevedo, Hugo Villamizar, Tatiana Escovedo, Helio Lopes, Stefan Biffl, et al. 2023. Status quo and problems of requirements engineering for machine learning: Results from an international survey. In *International Conference on Product-Focused Software Process Improvement*. Springer, 159–174.

[2] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 291–300.

[3] Victor R Basili-Gianluigi Caldiera and H Dieter Rombach. 1994. Goal question metric paradigm. *Encyclopedia of software engineering* 1, 528-532 (1994), 6. https://www.cs.umd.edu/~mvz/handouts/gqm.pdf

[4] Juliet Corbin and Anselm Strauss. 2008. Basics of qualitative research: Techniques and procedures for developing grounded theory. *Thousand Oaks* 3 (2008), 1–400. https://archive.org/details/basicsofqualitat0000stra

[5] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* 23, 226 (2022), 1–61.

[6] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. Selecting empirical methods for software engineering research. In *Guide to Advanced Empirical Software Engineering*. Springer, 285–311.

[7] Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

[8] Harald Foidl, Michael Felderer, and Rudolf Ramler. 2022. Data smells: Categories, causes and consequences, and detection of suspicious data in ai-based systems. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*. 229–239.

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

[10] Venkat Gudivada, Amy Apon, and Junhua Ding. 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software* 10, 1 (2017), 1–20.

[11] J. Horkoff. 2019. Non-functional requirements for machine learning: challenges and new directions. In *27th IEEE International Requirements Engineering Conference (RE)*. IEEE, 386–391.

[12] International Organization for Standardization. 2008. ISO/IEC 25012:2008 — Software engineering — Software product quality requirements and evaluation (SQuaRE) — Data quality model.

[13] International Organization for Standardization and International Electrotechnical Commission. 2024. *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples*. Technical Report ISO/IEC 5259-1:2024. ISO/IEC, Geneva, Switzerland. https://www.iso.org/standard/81088.html

[14] A. Jarzębowicz and P. Weichbroth. 2021. A qualitative study on non-functional requirements: In agile software development. *IEEE Access* 9 (2021), 40458–40475.

[15] Marcos Kalinowski, Daniel Mendez, Görkem Giray, Antonio Pedro Santos Alves, Kelly Azevedo, Tatiana Escovedo, Hugo Villamizar, Helio Lopes, Teresa Baldassarre, Stefan Wagner, et al. 2025. Naming the pain in machine learning-enabled systems engineering. *Information and Software Technology* (2025), 107866.

[16] Christian Kastner. 2025. *Machine learning in production: from models to products*. MIT Press.

[17] Hiroshi Kuwajima, Hirotoshi Yasuoka, and Toshihiro Nakae. 2020. Engineering problems in machine learning systems. *Machine Learning* 109, 5 (2020), 1103–1126.

[18] Skylar Lei and Michael R Smith. 2003. Evaluation of several nonparametric bootstrap methods to estimate confidence intervals for software metrics. *IEEE Transactions on Software Engineering* 29, 11 (2003), 996–1004.

[19] R. Likert. 1932. *A Technique for the Measurement of Attitudes*. Number N° 136-165 in A Technique for the Measurement of Attitudes. Archives of Psychology. https://archive.org/details/likert-1932

[20] Johan Linåker, Sardar Muhammad Sulaman, Rafael Maiani de Mello, and Martin Höst. 2015. *Guidelines for conducting surveys in software engineering*. Technical Report. https://portal.research.lu.se/en/publications/8ac54dbe-b7ac-4244-9c43-0f0d157efa26

[21] Mark S Litwin and Arlene Fink. 1995. *How to measure survey reliability and validity*. Vol. 7. Sage, https://methods.sagepub.com/book/how-to-measure-survey-reliability-and-validity. doi:10.4135/9781483348957

[22] Clifford E Lunneborg. 2001. Bootstrap inference for local populations. *Drug information journal: DIJ/Drug Information Association* 35, 4 (2001), 1327–1342.

[23] Aryely Matos, Mario Patrício, Maria Isabel Nicolau, Edna Dias Canedo, Juliana Alves Pereira, and Anderson Uchôa. 2025. Data Privacy in Software Practice: Brazilian Developers' Perspectives. *Journal of Internet Services and Applications* 16, 1 (Jun. 2025), 291–311. doi:10.5753/jisa.2025.5302

[24] E. de S. Nascimento, I. Ahmed, E. Oliveira, M. P. Palheta, I. Steinmacher, and T. Conte. 2019. Understanding development process of machine learning systems: Challenges and solutions. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 1–6.

[25] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–28.

[26] Neoklis Polyzotis, Martin Zinkevich, Sudip Roy, Eric Breck, and Steven Whang. 2019. Data validation for machine learning. *Proceedings of machine learning and systems* 1 (2019), 334–347.

[27] Kevin Pontes, Arthur Cordeiro, Jessica da Silva, Jhefferson Souza, Emanuel Avila Cruz Pires, Marcelo Martins, Hugo Villamizar, and Anderson Uchôa. 2026. *Replication package for the paper: "Practitioners' Perceptions of Data Quality Requirements in Machine Learning-Enabled Systems: An Exploratory Survey"*. doi:10.5281/zenodo.21327346

[28] Maria Priestley, Fionntán O'donnell, and Elena Simperl. 2023. A survey of data quality requirements that matter in ML development pipelines. *ACM Journal of Data and Information Quality* 15, 2 (2023), 1–39.

[29] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 1776–1826.

[30] Neil Salkind. 2012. *Exploring Research* (8th ed.). Pearson Education. https://archive.org/details/exploringresearc0000salk_z7v0

[31] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[32] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015).

[33] Samuel Shapiro and Martin Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3/4 (1965), 591–611. https://www.jstor.org/stable/2333709

[34] Arumoy Shome, Luis Cruz, and Arie Van Deursen. 2022. Data smells in public datasets. In *Proceedings of the 1st international conference on ai engineering: Software engineering for ai*. 205–216.

[35] Julien Siebert, Lisa Joeckel, Jens Heidrich, Koji Nakamichi, Kyoko Ohashi, Isao Namba, Rieko Yamamoto, and Mikio Aoyama. 2020. Towards guidelines for assessing qualities of machine learning systems. In *international conference on the quality of information and communications technology*. Springer, 17–31.

[36] Hugo Villamizar, Marcos Kalinowski, Hélio Lopes, and Daniel Mendez. 2024. Identifying concerns when specifying machine learning-enabled systems: A perspective-based approach. *Journal of Systems and Software* 213 (2024), 112053. doi:10.1016/j.jss.2024.112053

[37] A. Vogelsang and M. Borg. 2019. Requirements engineering for machine learning: Perspectives from data scientists. In *27th IEEE International Requirements Engineering Conference Workshops (REW)*. IEEE, 245–251.

[38] Stefan Wagner, Daniel Méndez Fernández, Michael Felderer, Antonio Vetrò, Marcos Kalinowski, Roel Wieringa, Dietmar Pfahl, Tayana Conte, Marie-Therese Christiansson, Desmond Greer, et al. 2019. Status quo in requirements engineering: A theory and a global family of surveys. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 28, 2 (2019), 1–48.

[39] Stefan Wagner, Daniel Mendez, Michael Felderer, Daniel Graziotin, and Marcos Kalinowski. 2020. Challenges in survey research. In *Contemporary Empirical Methods in Software Engineering*. Springer, 93–125.

[40] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer. doi:10.1007/978-3-642-29044-2

[41] Zhi-Hua Zhou. 2021. *Machine learning*. Springer nature.