# What Makes a High-Quality Training Dataset for Large Language Models: A Practitioners' Perspective

Xiao Yu
Huawei
Hangzhou, China
yuxiao25@huawei.com

Zexian Zhang
Wuhan University of Technology
Wuhan, China
zexianzhang@whut.edu.cn

Feifei Niu*
University of Ottawa
Canada
niufeifei@smail.nju.edu.cn

Xing Hu
The State Key Laboratory of Blockchain and Data Security, Zhejiang University
Hangzhou, China
xinghu@zju.edu.cn

Xin Xia
Huawei
Hangzhou, China
xin.xia@acm.org

John Grundy
Monash University
Australia
john.grundy@monash.edu

## ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable performance in various application domains, largely due to their self-supervised pre-training on extensive high-quality text datasets. However, despite the importance of constructing such datasets, many leading LLMs lack documentation of their dataset construction and training procedures, leaving LLM practitioners with a limited understanding of what makes a high-quality training dataset for LLMs. To fill this gap, we initially identified 18 characteristics of high-quality LLM training datasets, as well as 10 potential data pre-processing methods and 6 data quality assessment methods, through detailed interviews with 13 experienced LLM professionals. We then surveyed 219 LLM practitioners from 23 countries across 5 continents. We asked our survey respondents to rate the importance of these characteristics, provide a rationale for their ratings, specify the key data pre-processing and data quality assessment methods they used, and highlight the challenges encountered during these processes. From our analysis, we identified 13 crucial characteristics of high-quality LLM datasets that receive a high rating, accompanied by key rationale provided by respondents. We also identified some widely-used data pre-processing and data quality assessment methods, along with 7 challenges encountered during these processes. Based on our findings, we discuss the implications for researchers and practitioners aiming to construct high-quality training datasets for optimizing LLMs.

**ACM Reference Format:**
Xiao Yu, Zexian Zhang, Feifei Niu, Xing Hu, Xin Xia, and John Grundy. 2024. What Makes a High-Quality Training Dataset for Large Language Models: A Practitioners' Perspective. In *Proceedings of IEEE/ACM International Conference on Automated Software Engineering (ASE'24).* ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author


*ASE'24, Oct 27–Nov 01, 2024, Sacramento, CA, USA*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Large Language Models (LLMs) have demonstrated impressive performance across numerous domains, including natural language processing [28, 32, 35, 66, 77] and software engineering [10, 17, 25, 56, 66, 71, 78, 80]. This exceptional performance is primarily attributed to their self-supervised pre-training on extensive high-quality text datasets [24, 55, 55, 81]. LLM developers, therefore, invest significant effort into ensuring data quality: selecting data sources, preprocessing the collected data to ensure quality, and evaluating whether these datasets meet the requisite standards for training LLMs [1] [39]. However, despite the crucial role of constructing high-quality datasets in ensuring the performance of LLMs, many prominent LLMs either inadequately document their dataset construction procedures [9, 14] or provide only partial documentation, primarily outlining the data pre-processing operations employed [20, 37, 40, 49, 57, 69, 73, 79], without providing insight into the rationale behind their choices. Although some studies explore the effectiveness of specific data pre-processing operations, such as deduplication and removal of low-quality data [8, 12, 30, 34, 36, 38, 50, 63, 68], they typically examine only individual or a few data pre-processing methods in isolation, failing to offer a holistic understanding of the characteristics of high-quality training datasets for LLMs. Consequently, there are still limited insights into how LLM practitioners perceive the characteristics of a high-quality training dataset in practice.

To address this gap, we adopt a mixed-methods approach to gain insights into LLM practitioners' perceptions on high-quality training datasets for LLMs, as well as the common practices and challenges associated with data pre-processing and data quality assessment. We begin with semi-structured interviews involving 13 experienced LLM practitioners (11 from industry, two from academia)

[1] Training LLMs typically demands a substantial investment of time and computational resources. After applying data preprocessing methods to ensure data quality, it is often essential to perform additional rounds of data quality assessment to verify that the training data meets high-quality standards, thereby optimizing the efficiency and effectiveness of the training process.

with an average of 6.8 years of experience in LLMs, pre-training models, and/or deep learning models. From these interviews, we derive 18 characteristics of high-quality LLM training datasets, as well as ten potential data pre-processing methods and six data quality assessment methods. We then conduct an exploratory survey with 219 software practitioners from 23 countries across 5 continents to rate the identified characteristics according to their importance and provide rating rationales. Additionally, we ask them to specify the data pre-processing and data quality assessment methods they employed and the challenges encountered during these processes.

This work makes the following key contributions:

(1) We conduct an empirical study by interviewing 13 experienced LLM practitioners and then surveying 219 LLM practitioners to investigate their perceptions of high-quality training datasets for LLMs. We also identify common practices and challenges associated with data pre-processing and data quality assessment.

(2) We identify 13 important characteristics of high-quality datasets accompanied by practitioner rationales and highlight widely used data pre-processing and data quality assessment methods. We also identify seven common challenges encountered during these processes.

(3) We discuss potential implications for LLM researchers and practitioners, aiming to foster future developments in the field.

The rest of the paper is structured as follows: Section 2 discusses existing methods for data quality assurance in LLMs and data quality assessment methods. Section 3 describes the research methodology of our study. Section 4 presents the results obtained from our research. In Section 5 and Section 6, we discuss threats to validity and implications of our results for practitioners and future research. Section 7 concludes the paper.

## 2 RELATED WORK

***Data Quality Assurance for LLMs.*** Several prominent LLMs detail the data pre-processing methods employed in their respective papers or technical reports, such as removing machine-generated documents [27, 40, 69, 79], removing short texts [20, 27, 40, 41, 45, 47, 48, 73, 76, 79], removing unauthorized content [27, 48, 73], and deduplication [20, 27, 40, 41, 47, 48, 57, 73, 76, 79]. However, most studies do not offer comprehensive explanations for their choices or clarify the resulting impact on model performance.

Several studies investigate the effectiveness of existing or proposed data pre-processing methods, such as removing machine-generated documents and short texts [50, 63], deduplication [8, 12, 30, 34, 38, 50, 63, 68, 80], and remove toxic data [52, 63]. However, they typically examine only individual or a few data pre-processing methods, leading to a lack of a comprehensive perspective on what makes a high-quality training dataset for LLMs. In addition, they do not explore the impact of different extents of data pre-processing on model performance robustness, a challenge identified by our respondents.

***Data Quality Assessment.*** Prior to the era of LLMs, researchers proposed various quantitative metrics for data quality, such as accuracy [16, 19, 31, 53, 54, 58, 70, 72], completeness [16, 19, 31, 53, 54, 58, 70, 72], consistency [16, 19, 31, 53, 54, 58, 70, 72], timeliness [16, 19, 31, 53, 54, 58, 70, 72], relevance [16, 19, 31, 58], accessibility [16, 19, 31, 58, 72], format [16, 46, 53, 72], validity [46, 51, 59, 64], integrity [16, 19, 46, 64, 72], and uniqueness [16, 51, 59, 64]. Some studies [13, 33, 61] proposed to build a data quality classification model using labeled high-quality and low-quality data as training datasets and then employ this model to predict the data quality of the data under assessment. Ding et al. [23] utilized the K-Means clustering and Gaussian mixture models to group the labeled data, focusing on identifying outliers as potential noisy data.

Various data quality assessment tools have been proposed, such as Apache Griffin [1], Deequ [2], Great Expectations [3], and Qualitis [5], to analyze data quality in terms of accuracy, completeness, duplication, data format, and so forth. However, to date there has been no research exploring how LLM practitioners assess the quality of training data for their LLMs.

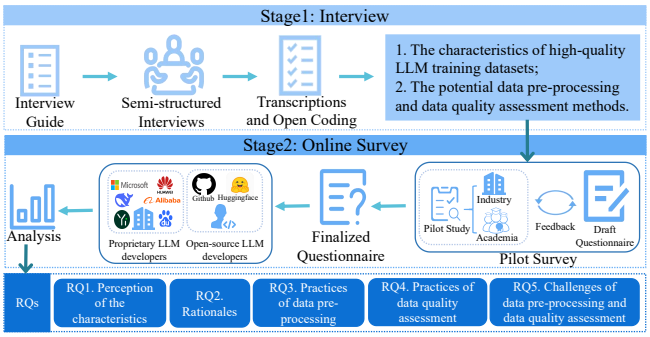


**Figure 1: Overview of the research methodology.**

## 3 METHODOLOGY

The methodology employed in this study follows a mixed-methods approach as shown in Figure 1 and consists of two stages. ***Stage* 1**: We conduct interviews with 13 LLM professionals to gather their insights into the characteristics of high-quality LLM training datasets, as well as the potential data pre-processing and data quality assessment methods. ***Stage* 2**: We carry out a large-scale online survey to ask LLM practitioners to assess the importance of the identified characteristics and provide their reasons. Additionally, we inquire about the data pre-processing and data quality assessment methods they employed, along with the challenges encountered during these processes. Each respondent spent 5-10 minutes to complete the survey. Both the interviews and survey received approval from the relevant institutional review board.

### 3.1 Stage 1: Interview

***Protocol:*** The first author conducted a series of face-to-face and in-depth interviews with 13 LLM professionals, each lasting between 40-60 minutes. The interviews are semi-structured and divided into three parts.

*Part 1:* We first ask some demographic questions, such as their job roles, educational background, years of experience with LLMs, pre-training models, and/or deep learning models, and team sizes. Furthermore, we explore practitioners' specialization in LLM categories (i.e., general LLMs or domain-specific LLMs) and their employed LLM training approaches, including training from scratch [2], continuing-pretraining [3], and fine-tuning.

[2]Training LLMs directly from randomly initialized parameters.
[3]Continuing to train model parameters based on a pre-trained model using additional large-scale unlabeled corpora.

*Part 2:* We ask two open-ended questions: (1) What do you consider to be essential characteristics of a high-quality training dataset for LLMs? (2) How do you or your team perform data pre-processing for quality assurance on LLM training datasets and evaluate their data quality?
*Part 3:* We prepare a list of candidate high-quality characteristics and data pre-processing as well as potential data quality assessment methods by thoroughly reviewing related papers or reports on data quality (e.g., [16, 21, 60]) and the data pre-processing and data quality assessment methods adopted in the LLM papers (e.g., [11, 57]). We select items not explicitly discussed and ask professionals to provide further insights on them.
***Interviewees:*** 13 LLM professionals are invited for interviews through our network in both the industry and academia. Eleven of these professionals are from global IT companies or LLM startups, including Microsoft, Huawei, Alibaba, Baidu, 01.AI, and DeepSeek, and occupy various roles such as data scientists, algorithm engineers, research scientists, and business analysts. Additionally, two professionals are university professors specializing in LLMs. Overall, they have an average of 6.8 years of experience in working with LLMs, pre-training models, and/or deep learning models (minimum: 2, median: 7, maximum: 11, standard deviation: 2.66). 60% of the professionals hold a Ph.D. degree. 30% of the professionals primarily focus on the application and research of general LLMs, while the remaining 70% concentrate on domain-specific LLMs. Regarding LLM training approaches, 23% of the professionals employ the training-from-scratch approach, 15% employ the continuing-pretraining approach, and 62% prefer the fine-tuning approach.
***Transcription and Open Coding:*** The first author transcribed and analyzed the interviews, using NVivo qualitative analysis software [4] for open coding to generate codes of the interview contents. The second author verified the initial codes created by the first author and provided suggestions for improvement. After incorporating these suggestions, the two authors separately analyzed the codes and sorted the generated cards into potential statements and answers. The overall Cohen's Kappa value between the two authors is 0.83, indicating substantial agreement. Disagreements were discussed to reach a common decision. To reduce bias, both authors reviewed and agreed on the final set of statements. Eventually, based on the results of the interviews, we identified 18 characteristics of high-quality LLM training datasets, as well as 10 potential data pre-processing methods and 6 data quality assessment methods.

## 3.2 Stage 2: Online Survey

***Design:*** We followed up our detailed interviews with a large-scale survey to confirm or refute the interview findings and potentially discover further insights about LLM dataset construction in practice. Our survey uses single/multiple-choice, Likert scale, and short-answer open questions. To account for respondents who may not understand or prefer not to answer, we include categories such as "Don't Know" or "Due to company privacy concerns, I prefer not to answer.". The survey consists of four sections:

(1) Demographics: We collected information about surveyed practitioners, including their country of residence, highest level of education, primary job role, years of experience with LLMs, pre-training models and/or deep learning, and team size.

(2) Practice of LLM training: We asked about practitioners' experiences with LLMs, including the main deep learning frameworks used (e.g., PyTorch, TensorFlow, and Keras.), types of LLMs involved (i.e., general LLMs and domain-specific LLMs), and approaches to training LLMs (i.e., training from scratch, continuing-pretraining, and fine-tuning). For practitioners who have used the fine-tuning approach, further details are requested, such as specific approaches to fine-tuning LLMs (i.e., full fine-tuning and parameter-efficient fine-tuning) and the fine-tuning data (i.e., instruct-tuning, Reinforcement Learning from Human Feedback (RLHF), task/domain-specific data fine-tuning).

(3) Perception of high-quality training datasets for LLMs: We gave the 18 characteristics of high-quality training datasets as shown in Table 1. Practitioners were then asked to rate these characteristics and provide explanations for why they perceive some as very important while others as less crucial.

(4) Practices and challenges of pre-processing and evaluating the quality of LLM training datasets: We asked practitioners what data pre-processing steps they perform on their datasets for quality assurance before training the LLM (e.g, removing duplicate data and low-quality documents). Next, we asked how practitioners evaluate the quality of their datasets (e.g., quantitative data quality metrics and data visualization). In addition, we ask about practitioners' reasons for adopting these data pre-processing and quality assessment steps, as well as the effectiveness of these steps. Finally, practitioners were asked about the challenges encountered during the LLM data pre-processing and data quality assessment stages.

At the end of the survey, practitioners are invited to share any additional thoughts they have about high-quality datasets for LLMs or our survey study. Prior to launching our survey, we conducted a pilot survey with five LLM professionals who were not part of the interviewees or surveyed practitioners. We sought feedback on the clarity and understandability of terms used as well as survey's length. Based on their feedback, we made minor adjustments to the draft survey and created a finalized version. To ensure accessibility for an international audience, we provide an English version of the survey on Google Forms. Additionally, in order to support practitioners from China, we translated the survey into Mandarin and made it available on a popular survey website [7] in China. English serves as an international lingua franca, while Mandarin is commonly spoken among many of our target audience.
***Participant Recruitment:*** To obtain a sufficient number of professionals from both industry and open-source LLM communities, we employed two strategies for participant recruitment: (1) We reached out to professionals within our social and professional networks employed in leading global IT companies or LLM startups like Microsoft, Huawei, Alibaba, Baidu, 01.AI, DeepSeek, and others. We asked for their assistance in disseminating our survey among their colleagues. Through this approach, we gather 102 responses. (2) We collected contributors' public email addresses from GitHub and Huggingface repositories, focusing on those involved in LLM projects. We then distributed an email to 10,891 potential developers containing a link to our survey. While we receive 59 auto-responses indicating the recipients' unavailability, we gather 121 responses. This response rate is slightly lower than that of other research surveys in software engineering (e.g., [42, 43, 62]), possibly due to privacy policies within some companies prohibiting employees

from disclosing details about their company's LLM development. Ultimately, we gathered a total of 223 survey responses. We filtered out three responses with completion times of less than two minutes and one response from an individual who is not a professional LLM practitioner. The survey results analysis presented in this paper is based on the remaining 219 valid responses. These respondents are from 23 countries spanning 5 continents, with China and the United States emerging as the top two countries with the highest number of respondents.

***Data Analysis:*** We analyze the survey results according to question type. For multiple-choice and single-choice questions, we report the percentage of each option selected. For Likert scale questions, we plot bar graphs to illustrate the distribution of Likert scores. In addition, we drop "Don't Know" and "Due to company privacy concerns, I prefer not to answer. " answers that form a negligible fraction (less than 1%) of all answers. For the open-ended questions, the first two authors independently analyzed these, categorizing them into specific characteristics, data pre-processing and quality assessment methods, or challenges encountered. The overall Cohen's Kappa value between the two authors is 0.86, indicating substantial agreement. Disagreements were discussed to reach a consensus.

## 4 RESULTS

We present the analysis results of our survey by answering the following key research questions:

- **RQ1:** How do practitioners perceive 18 candidate characteristics of high-quality training datasets for LLMs?
- **RQ2:** What are practitioners' rationales for perceiving a particular characteristic as important or unimportant?
- **RQ3:** What are the common data pre-processing operations used by practitioners for data quality assurance?
- **RQ4:** What are the common data quality assessment operations used by practitioners?
- **RQ5:** What challenges do practitioners frequently encounter during LLM data pre-processing and data quality assessment?

### 4.1 RQ1: Perception of the characteristics

Table 1 outlines the 18 characteristics of high-quality LLM training datasets, along with the distribution of surveyed practitioners' ratings regarding their importance and their average Likert scores (ranging from Not important=1 to Very Important=5). Notably, 13 out of the 18 characteristics have average Likert scores of 4.0 or higher. This suggests that these 13 characteristics are perceived as important and very important. Following prior works [22, 44, 75], we further investigate the ratings of various demographic groups as below:

- **Main job role:** Respondents who are data professionals, algorithm specialists, or others (such as business analysts and project managers).
- **Open-source LLM developers vs. Proprietary LLM developers:** Respondents who are based on their involvement in open-source LLM development or proprietary LLM development.
- **Experience level:** Respondents with low experience (ExpLow), i,e., we define as the 25% with the least experience in years (<3 years), with medium experience (ExpMed), i.e., 3-5 years, or with most experience (ExpHigh), i.e, we define as the 25% with the most experience in years (5-10 years).
- **Education level:** Respondents with a PhD degree or Bachelor's/Master's degree.
- **Type of LLMs:** Respondents focusing primarily on general LLMs or domain-specific LLMs.
- **Methods for constructing LLMs:** Respondents who employ the training-from-scratch/continue-pretraining approaches or fine-tuning approach.
- **Data of fine-tuning:** Respondents who primarily use instruct-tuning, RLHF, or task/domain-specific data fine-tuning.

Table 2 shows the rated importance of the 18 characteristics for different demographic categories of respondents, where Percentage represents the proportion of respondents within each demographic category. Note that because respondents may choose more than one option at the same time (e.g., some respondents will be involved in both data engineering and algorithmic engineering), the percentages in the demographic category will not add up to 100%.

All demographics give more "important" and "very important" than "not important" or "slightly important". Across all demographic groups, only a minority (less than 11%) give "not important" and "slightly important". More than 75% of respondents across all demographic groups rate the 18 characteristics as "very important" or "important", and about 49% - 62% of respondents rate these characteristics as "very important".

Following the practices in previous studies [18, 29, 75], we employ Fisher's Exact test [26] with Bonferroni correction [15] (when conducting multiple comparisons) to examine whether various demographic groups exhibit significant differences across individual characteristic and all 18 characteristics. We observe that for the importance of individual characteristics, there are no significant differences among various demographic groups [4]. For all 18 characteristics, there are also no significant differences among various demographic groups, except for open-source practitioners and proprietary LLM developers. Open-source practitioners perceive all 18 characteristics as significantly more important than proprietary LLM developers.

In line with practices in previous studies [65, 74, 75], we employ the Scott-Knott Effect Size Difference (SKESD) test [67] to divide the 18 characteristics into different groups based on their Likert scores. Note that we excluded three responses that chose "Don't know". The differences in the importance of the characteristics within the same group are negligible, while the differences in the characteristics between different groups are significant. Table 3 presents the characteristics ranked according to the SKESD test for all the respondents. Reliability, Relevance, Accuracy, and Compliance are considered the most important characteristics by the respondents, while Completeness, Balance, Absence of Duplicate Data, Consistency, and Timeliness are regarded as the least important.

[4] For the importance of individual characteristics among various demographic groups, please refer to our appendix [6].

**Table 1: The characteristics of a high-quality training dataset for LLM.**

| Characteristic | Distribution | Score |
|---|---|---|
| **Data Sources and Types** | | |
| **Wide Range of Sources:** The dataset includes diverse sources such as news articles, web texts, e-books, encyclopedias, scientific papers, social media posts, dialogue data, and code. (For general LLMs) | 17% 22% 56% | 4.25 |
| **Diversity:** The dataset covers various language styles, topics, and domains. (For general LLMs) | 8% 8% 30% 54% | 4.30 |
| **Relevance:** The data in the dataset should be closely related to specific questions, tasks, or goals. (For domain-specific LLMs) | 30% 70% | 4.70 |
| **Large-Scale Data:** The dataset contains a substantial amount of data, such as billions of tokens. | 6% 11% 31% 53% | 4.31 |
| **Reliability:** The data in the dataset should come from trustworthy sources and undergo proper collection, processing, and storage processes. | 19% 78% | 4.73 |
| **Data Content** | | |
| **Accuracy:** The data (or data annotations) in the dataset accurately reflect real situations without errors, omissions, or inaccuracies. | 22% 73% | 4.65 |
| **Knowledge Content:** The richness of useful information or knowledge contained in a text. For example, Wikipedia data has a higher knowledge content than social media content. | 8% 11% 25% 56% | 4.28 |
| **Absence of Toxic Data:** The dataset does not contain inappropriate content such as political discrimination, pornography, violence, politics, gender or racially biased, or vulnerable or defective code. | 9% 17% 34% 40% | 4.06 |
| **Absence of Low-Quality Documents:** The dataset excludes low-quality documents such as machine-generated documents, documents with grammatical errors, very short texts, texts without punctuation, URLs, spam, etc. | 6% 20% 23% 51% | 4.20 |
| **Absence of Duplicate Data:** The dataset does not include duplicate words, phrases, or duplicate documents. | 5% 8% 30% 24% 32% | 3.70 |
| **Timeliness:** For time-sensitive data (e.g., news data), the dataset should be regularly updated. | 17% 6% 17% 31% 31% | 3.53 |
| **Balance:** The dataset maintains relatively balanced sample sizes among different categories. | 6% 6% 31% 25% 33% | 3.75 |
| **Data Structure and Management** | | |
| **Consistency:** The data should be consistent, meaning the same data type should be represented in a consistent format, unit, and specification. | 8% 11% 16% 32% 32% | 3.70 |
| **Completeness:** The dataset should contain all necessary data fields and variables without missing values or empty fields. | 16% 16% 24% 41% | 3.84 |
| **Documentation:** The dataset should have detailed descriptions for other researchers or users to understand and utilize the dataset. | 16% 35% 49% | 4.32 |
| **Accessibility:** The dataset should be easily accessible and usable, with appropriate data structures and formats. | 14% 33% 53% | 4.39 |
| **Data Security and Compliance** | | |
| **Privacy Protection:** The dataset should not contain sensitive data such as personal identity, location information. | 6% 8% 19% 64% | 4.36 |
| **Compliance:** The dataset should be used according to proper licenses and copyrights and comply with data protection regulations. | 6% 25% 67% | 4.53 |

Not Important Slightly Important Moderately Important Important Very Important

## 4.2 RQ2. Rationales for importance

We use ✔or ✖to denote the rationale behind respondents perceiving a characteristic as important or unimportant. We include key rationales reported by survey respondents.

**Reliability.** Respondents emphasize the importance of reliability for several reasons: (1) Unreliable data may mislead LLMs; (2) Reliable data sources and data construction methods reduce the costs of validating data reliability; (3) Specific domains, such as network security, necessitate reliable data sources.

✔*If the training dataset source is* ***unreliable****, it may* ***lead to incorrect outputs*** *from the model.*

✔*The data source should ideally be reliable and validated, as it's kinda* ***costly*** *for us* ***to validate a dataset****.*

✔*In the field of* ***network security****, we are more concerned about the accuracy and reliability of the data, as LLM-Sec applications are more concerned with a high degree of* ***integration and automation of applications with the existing security infrastructure****, where the reduction of noise is an important feature.*

**Relevance.** Respondents emphasize the importance of relevance for two main reasons: (1) It enhances the model's understanding of domain-specific knowledge, thereby improving its performance in downstream tasks; (2) It prevents the model from having to learn from irrelevant data, thus enhancing training efficiency.

✔ *Domain-specific data contains commonly used* ***professional terminologies and expressions*** *in the field. When trained on such data, the model can* ***generate more specialized responses in that domain****.*

**Table 2: The importance of all 18 characteristics for different demographic categories of respondents.**

| Demographic /Percentage | Distribution |
|---|---|
| All | 6% 13% 27% 52% |
| Data professional (30%) | 6% 13% 28% 49% |
| Algorithm specialist (81%) | 6% 13% 25% 53% |
| Other roles (30%) | 4% 6% 16% 32% 42% |
| Open-source developers (62%) | 5% 13% 29% 52% |
| Proprietary developers (38%) | 5% 6% 14% 23% 52% |
| PhD (27%) | 7% 12% 30% 49% |
| Master and Bachelor (73%) | 5% 14% 26% 53% |
| ExpLow (39%) | 5% 14% 30% 49% |
| ExpMed (32%) | 8% 13% 25% 51% |
| ExpHigh (29%) | 6% 12% 19% 60% |
| General LLMs (46%) | 5% 13% 27% 54% |
| Domain-Specific LLMs (89%) | 6% 13% 26% 52% |
| Training from scratch/Continue-pretraining (43%) | 6% 14% 22% 55% |
| Fine-tuning (92%) | 6% 12% 26% 53% |
| Instruct-tuning (68%) | 5% 11% 24% 58% |
| RLHF (27%) | 7% 12% 16% 62% |
| Task/domain-specific data (73%) | 6% 12% 25% 55% |

Not Important Slightly Important Moderately Important Important Very Important

**Table 3: The highly-ranked characteristics according to the SKESD Test.**

| Group | Characteristic |
|---|---|
| 1 | Reliability, Relevance, Accuracy |
| 2 | Compliance |
| 3 | Accessibility, Privacy Protection, Documentation, Large-Scale Data, Diversity |
| 4 | Knowledge Content, Wide Range of Sources, Absence of Low-Quality Documents |
| 5 | Absence of Toxic Data |
| 6 | Completeness, Balance, Absence of Duplicate Data, Consistency |
| 7 | Timeliness |

✔ *When the training data is directly relevant to the target tasks of the large model, it* ***avoids*** *the model having to* ***learn from irrelevant data****, thus* ***improving training efficiency****.*

**Accuracy.** Respondents emphasize the importance of accuracy due to two main reasons: (1) Accuracy ensures that decisions, analyses, and conclusions drawn from the training data are correct and reliable; (2) Accuracy significantly impacts the generalization ability and performance of the trained LLMs.

✔*Accuracy is often considered the most crucial data quality characteristic.* ***Inaccurate data can lead to incorrect decisions, analysis, and conclusions****, which can have severe consequences in various domains, such as finance, healthcare, or scientific research. Ensuring data accuracy is critical for* ***maintaining the trust and making reliable decisions****.*

✔*Accuracy: high-quality data is often the key to the model's success, and* ***problems*** *with a well-trained model are usually the result of* ***mislabelling facts or labeling with some unintentional paradigm*** *brought into play.*

✔*The accuracy of the labeling* ***affects the generalization ability*** *of the trained model and* ***its performance on a specific task****.*

✔*If the labeling quality in the dataset is low or noisy, the model may* ***learn incorrectly or perform poorly****.*

**Compliance.** Respondents consider compliance as important since it: (1) ensures that LLMs meet regulatory requirements; (2) prioritizes the legality of data collection processes, safeguarding against copyright infringement and violations of user privacy while adhering to applicable laws and regulations.

✔*Compliance is very important, in China, all large language models are* ***required to pass regulatory compliance audits****.*

✔*The first step is to ensure that the data is* ***legally compliant without compromising privacy****.*

✔*The need to ensure the legitimacy of the data collected,* ***avoid infringement of copyright and user privacy****, and* ***comply with relevant laws and regulations****.*

**Accessibility.** Respondents consider accessibility as important because easily accessible and usable training datasets facilitate other development teams in developing LLMs more easily and effectively.

✔*Accessibility is for* ***ease and efficiency of development****.*

✔*When making datasets publicly available, it is also important to ensure that the data is* ***accessible and easy to use with a clear data structure*** *so that other model development teams can benefit from our data.*

**Privacy Protection.** Respondents consider privacy protection as important because it prevents the generation of dialogues that could potentially expose user privacy, ensuring the confidentiality of personal information in model outputs.

✔*Privacy protection is to* ***avoid leaking some critical data that can easily lead to lawsuits****, especially for financial companies and areas with stricter regulations.*

✔*This effectively* ***prevents the model from generating privacy-revealing dialogues****.*

**Documentation.** Respondents consider documentation important because (1) it provides clarity on the domain and intended usage of the data, and (2) unclear descriptions may lead developers to misuse the dataset.

✔*Documentation Note: This is to make it* ***easier to determine the domain and role*** *to which the data applies.*

✔*If the data fields in the dataset lack detailed descriptions,* ***developers may misuse certain data fields****, leading to potential performance issues in the trained large language models.*

**Large-Scale Data.** Respondents value large-scale data because it provides the model with a diverse and extensive set of training samples, facilitating robust learning and improved performance.

✔*Data size: The dataset size is usually closely related to the model performance.* ***Larger datasets usually lead to better results*** *because the model has more training samples to learn from.*

✔*The abundance of data offers the model* ***a broader perspective of the underlying knowledge within the data****, thus reducing the risk of model overfitting.*

However, some respondents prioritize data quality over data size.

✖*Large-scale data: TextBook is all you need – Phi-1 and Phi-2 have demonstrated that* ***it is not the amount of data that counts****,* ***but the accuracy of the data*** *that improves the performance of the model.*

✖***Quality over quantity****: even if the dataset is small, if* ***the quality and relevance*** *of each data point are high, it can still significantly* ***improve model performance****.*

**Diversity.** Respondents consider diversity important for two main reasons: (1) it enhances the language model's ability to understand and learn from various language styles and patterns, and (2) it prevents LLMs from overfitting and becoming biased toward specific topics.

✔*The large language model is actually learning the language pattern, so the data diversity can* ***avoid the model is too fit to a certain local****, the more diverse the data, the trained model will be more like a human to express.*

✔*Diversity in datasets can help models to better understand and generalize to a variety of different language structures and contexts. Diverse datasets can* ***reduce overfitting and improve the generalization of models****.*

✔*Diverse data encompasses content from various backgrounds and perspectives, facilitating large language models in* ***reducing biases on specific topics****.*

**Knowledge Content.** Respondents emphasize the importance of knowledge content for two main reasons: (1) the high knowledge content better assists in solving domain-specific problems, and (2) enables LLMs to provide more accurate responses.

✔*In fields where high precision and reliability are required, such as the medical field, training data with high knowledge content makes it easier for large language models to* ***address specific needs within the field****.*

✔*Concrete, detailed and accurate knowledge is not relevant if the aim is to create general language understanding in LLMs, but may be* ***more relevant for domain-specific LLMs****.*

✔*When tackling complex tasks, large language models* ***require rich knowledge to provide more accurate answers****.*

**Wide Range of Sources.** Respondents emphasize the importance of a wide range of sources for two main reasons: (1) It enables LLMs to handle diverse formats, styles, and grammar, thereby enhancing their robustness. (2) Diverse data sources provide essential information for various downstream tasks.

✔ *The diverse data sources encompass texts with different formats, styles, and grammars, enabling the model to* ***learn how to handle these variations and enhance its robustness****.*

✔*Different data sources can* ***give us what we need for all sorts of language tasks****. Like, news articles are awesome for getting factual info for language models, scientific papers are perfect for academic stuff, social media posts can make conversations sound more natural, and code data helps improve code generation skills.*

✔*By acquiring data from a variety of sources, including different industries, languages, cultural backgrounds, and content types, it* ***ensures that the data is diverse and comprehensive, covering a wide range of situations in the current scenario****.*

**Absence of Low-Quality Documents.** Respondents consider the absence of low-quality documents as important because removing such documents (1) ensures that LLMs are trained on high-quality data, leading to more accurate and effective outputs, (2) and enhances the efficiency of LLMs training.

✔*Removing duplicate as well as low-quality data* ***reduces the model training cost*** *on one hand and* ***enhances the reliability of the model output*** *on the other.*

✔ ***Low-quality training documents can mislead large language models****, resulting in inaccurate output. Removing these low-quality documents can improve the quality of the model's output.*

However, some respondents suggest retaining a portion of low-quality training text to enhance the robustness of LLMs against such inputs.

✖*When people use large language models, their sentences might not always have perfect grammar. For example, with Google's Phi1 and Phi2 models, if you ask a question that's not well-phrased, the answers can sometimes be totally off. This might be* ***because the training data didn't include enough informal or poorly-structured dialogue, so the models aren't as good at handling those inputs****.*

✖*If you* ***remove too much of it****, you'll have problems with* ***the robustness of the model****.*

**Absence of Toxic Data.** Respondents consider the absence of toxic data as crucial, as eliminating such data mitigates the risk of the model generating harmful or offensive content during interactions and promotes a safer and more positive user experience.

✔*The accuracy, knowledge content, non-toxicity, and timeliness of the dataset directly* ***determine the strength and reliability of the trained model****.*

✔*This can effectively* ***prevent the model from generating toxic dialogues****.*

✔*Removing toxic data is essential not only for model performance but also for* ***ensuring ethical and responsible AI deployment****.*

However, some respondents suggest retaining a portion of toxic data and ensuring that LLMs are aware of their toxicity so that LLMs do not produce toxic responses.

✖*Some content involving pornographic and violent themes can be exposed to LLMs. This* ***enables the LLMs to recognize such content as inappropriate****. Consequently, when users inquire about pornographic or violent content, the models can avoid responding.*

**Completeness.** Respondents do not prioritize completeness as important for two main reasons: (1) It may lead to an excessive focus on exhaustive data collection, which might not always be necessary or feasible. (2) LLMs have the capacity to handle missing or incomplete data.

✖*In real life, some data fields might be missing because of privacy issues or incomplete sources. Making sure the dataset is complete* ***takes a lot of time and effort****, and* ***it's not always needed. LLMs can figure out and fill in the missing info*** *by learning and making guesses.*

✖ *Large language models typically have* ***the capability to handle missing or incomplete data****, thereby enhancing their robustness to incomplete data.*

However, some respondents perceive completeness as crucial.

✔*Data completeness refers to the absence of missing or incomplete values or records. Incomplete data can* ***result in biased analysis, misleading conclusions, and incomplete insights****.*

**Balance.** Most respondents do not consider data balance as important since (1) attempting to balance data distributions would reduce the diversity of data and overlook the natural variation in language usage across different contexts and domains; (2) Achieving perfect balance across all data categories may be impractical or unnecessary.

✖*Imposing strict balance requirements may also* ***overlook the natural variation in*** *language usage across* ***different contexts and domains****,* ***leading to a less representative dataset****.*

✖*Achieving perfect balance across all data categories may be* ***impractical or unnecessary****. Other factors, such as data quality, relevance, and coverage are more important.*

However, some respondents consider it to be important.

✔*The distribution of samples in the dataset is also* ***important for the model's performance and generalization ability****. If the data distribution is unbalanced or not representative of real-world situations, the model may be biased.*

**Absence of Duplicate Data.** Most respondents do not consider the absence of duplicate data as important because (1) Removing

duplicate or similar data may result in the loss of potentially important information or features; (2) LLMs are capable of handling such duplicate data.

✖*Deleting duplicate or similar data could potentially* ***lead to the loss of some potentially important information or features****, which could affect the performance of the model.*

✖*Duplicate data has little effect on the overall performance of LLMs. In some cases, duplicate data* ***may contribute to the robustness of the model by focusing on certain patterns or concepts****.*

✖*Most large language models already exhibit* ***a certain level of robustness*** *during training and are* ***capable of handling redundancy*** *in the training data.*

However, some respondents think that handling duplicate data is important to reduce LLMs' training time.

✔ *Duplicate data increases the size of the training set, causing the model to deal with a large number of similar samples, thereby* ***increasing training time and computational costs****.*

**Consistency.** Most respondents do not prioritize consistency as crucial because real-world data often varies in formats, and LLMs can handle inconsistencies to some extent.

✖*Format consistency of data may not be as important because often* ***real-world data exists in a wide variety of formats****, and models need to learn a wide variety of data representation formats to cope with a wide variety of usage scenarios.*

✖*Language models are* ***capable of handling minor inconsistencies*** *in the dataset.*

✖*LLMs are often* ***designed to be robust to noise and variability*** *in the input data.*

However, some respondents consider maintaining a certain level of data consistency essential for seamless integration and interoperability across systems or applications.

✔*Consistency ensures that data adheres to defined rules, formats, and constraints.* ***Inconsistent data can lead to integration issues, data duplication, and incorrect analysis****.* ***Maintaining data consistency is crucial for seamless data integration, sharing, and interoperability across different systems or applications****.*

**Timeliness.** Most respondents consider timeliness as not very important for two main reasons: (1) The domain model being trained does not necessitate recent data, diminishing the significance of timeliness consideration. (2) The timeliness requirement of LLMs can be fulfilled through alternative means such as external databases, web searches, and retrieval-augmented generation.

✖*Since the domain model I am training* ***does not need to be very current****, i.e., it does not have to consider recent data, timeliness is not very important.*

✖*Timeliness data* ***can be replaced with retrieval-augmented generation****.*

✖*Timely updating of timeliness data is also less important, and* ***would be better addressed through methods such as external databases or web searches****.*

Some respondents also indicate that timeliness is important, but it is challenging for LLMs to guarantee timeliness.

✔*Timeliness is* ***important in certain domains****, such as news. However, it is not feasible for large language models to frequently update using recent data due to the significant time cost involved, and OpenAI's models also* ***face difficulties in ensuring timeliness****.*

## 4.3 RQ3. Practices of data pre-processing

We investigated whether our surveyed practitioners perform data pre-processing for data quality assurance before training their LLMs.

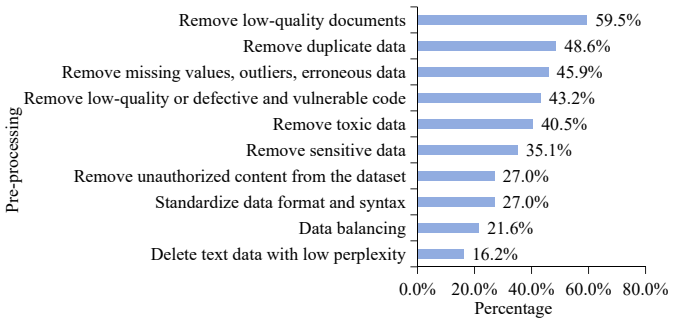


**Figure 2: The percentage of data pre-processing methods used by respondents.**

The results show that 3% of respondents do not perform any data pre-processing, while 7% indicate that their team does, but they are unable to provide specific details due to lack of knowledge or company privacy concerns. The remaining 90% of respondents explained the data pre-processing methods they use. Figure 2 summarises the main methods reported and the proportion of respondents who adopt each one. Because practitioners may employ multiple data preprocessing methods, the sum of these proportions does not equal 100%.

In Table 1, we rank characteristics based on their average importance as follows: Absence of low-quality documents, Absence of toxic data, Balance, Completeness, and Absence of Duplicate Data. Not surprisingly, the ranking of the percentage of data pre-processing methods employed by respondents is quite consistent with the ranking of their corresponding characteristics' importance (except for data balancing). Some respondents provided general descriptions of their specific data pre-processing workflows:

✍*Initially, data collection is conducted using a standardized format such as JSON, followed by coarse filtering (e.g., removing obviously erroneous or irrelevant data). Then, toxic and private data are filtered out, followed by handling anomalies and missing values in the data. Next, incomplete or defective code fragments are removed, and finally, data deduplication is performed. However, this is just a general process, as many data cleaning methods require custom heuristic strategies.*

✍ *I conduct a quality filter of the code, considering factors such as the number of stars in GitHub repositories and download counts for other open-source code. Additionally, I filter based on the code's language and the number of files in the code repository, excluding niche or unpopular languages and codes to ensure overall quality.*

Among these 10 data pre-processing methods, the practice of deleting text data with low perplexity stands out as more common among proprietary LLM developers (49%) compared to open-source LLM developers (21%). However, for the remaining 9 methods, the difference in usage between proprietary and open-source LLM developers is generally not substantial (typically within 20%). The prevalent use of deleting text data with low perplexity among proprietary developers may be attributed to their access to advanced tools and pre-annotated high-quality datasets for data pre-processing. As one proprietary LLM developer explains:

✍ *We follow the data cleaning approach of CCNet web data. We train a 5-gram Kneser-Ney model on a high-quality dataset from the same domain and use it to compute the perplexity of the evaluation text. Text with low perplexity is then removed.*

## 4.4 RQ4. Practices of data quality assessment

We investigated whether our surveyed practitioners conduct data quality assessments before training their LLMs. Our results indicate that 11% of respondents do not perform any data quality assessments, while 20% state that their team does but cannot provide specific details due to either lack of knowledge or company privacy concerns. The remaining 69% of respondents elucidate the data quality assessment methods they employ. Figure 3 summarises these methods and the proportion of respondents who utilize each one. Manual assessment, quantitative data quality metrics, and data visualization emerge as the most commonly used methods.

Among these 6 data quality assessment methods, "Train a smaller-scale LLM using both preprocessed and unprocessed data separately. If the LLM trained on the preprocessed data exhibits better performance, then the preprocessed dataset is considered to be of higher quality" stands out as more common among proprietary LLM developers (43%) compared to open-source LLM developers (14%). However, for the remaining 5 methods, the difference in usage between proprietary and open-source LLM developers is generally not substantial (typically within 20%). The prevalent use of this method among proprietary developers may be attributed to the fact that proprietary companies have larger parameter sizes for their LLMs and need to ensure data quality before training LLMs. Sandbox experiments with smaller-scale models are a more intuitive method for evaluating data quality. Some respondents share general details about their quality assessment methods:

✍ *1. Directly print out the data distribution to see the diversity of the data. 2. Directly print out the data and look at dozens of cases to see if the length of the text description, the logic is reasonable, whether there is a large amount of duplicated data, and if it is question-and-answer data, to see the reasonableness and diversity of the question-and-answer.*

✍ *The data amount of the fine-tuned model is not very high, and manual sampling is used to check that the data meets the training requirements.*

✍ *We first use an automated evaluation model, followed by inspecting the sampled data classified as high and low quality. Based on the identified misclassified data, we summarize the reasons for misjudgment, further refine the evaluation criteria, and update the automated evaluation model accordingly. Finally, before training the large language model, we conduct sandbox experiments using models with smaller parameter sizes.*

✍ *We utilize labeled high-quality data and other data to establish a classification model, determining whether the data under evaluation is of high quality.*

✍ *I use ChatGPT to classify whether the code snippet is of high quality.*

✍ *Sandbox experiments with small-scale large language models allow for the observation of the relationship between data quality and validation set's perplexities.*

✍ *During data quality assessment, tools are used to check whether the data format is correct and if all data fields are complete.*

✍ *Internally, the company maintains a data quality dashboard that utilizes quantitative metrics to assess data security, consistency, completeness, and other aspects.*

## 4.5 RQ5. Challenges of data pre-processing and data quality assessment

We asked practitioners about the significant challenges that they face during the LLM data pre-processing and data quality assessment phases. We categorize seven main types of challenges and present their frequencies using the multiplication symbol (*X*).

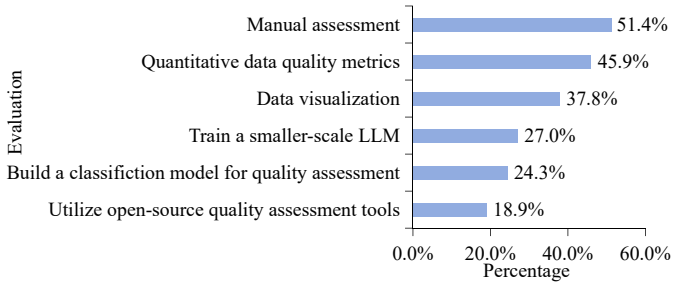


**Figure 3: The percentage of data quality assessment steps uesd by respondents.**

***Lack of uniform, standardized, and systematized data pre-processing methods (24 X).*** A few respondents highlighted that the absence of a uniform pre-processing procedure complicates the handling of diverse datasets.

✍ *I need to handle data from different fields and sources, each with varying data structure formats and data quality. Currently, there is no unified pre-processing procedure for different datasets.*

✍ *Lack of standardized tools and metrics to do this data cleaning. Often poor quality or poorly documented internal datasets make working with data hard. Data silos and teams not being willing to share data make job hard.*

***Unsure of the extent of data pre-processing (21 X).*** Several respondents expressed uncertainty regarding the optimal extent to which low-quality, toxic, and duplicate data should be removed, and questioned the feasibility of transforming low-quality data into high-quality data to increase the dataset size.

✍ *The extent of data pre-processing is an issue, too much processing may have negative impacts, too little processing may leave the dataset containing too much low-quality data, and some detailed research is needed to indicate the positive and negative impacts of each type of processing.*

✍ *When cleaning code, filtering directly based on the number of stars may result in the loss of valuable code files. Additionally, some domains inherently have fewer datasets, and cleaning further reduces their quantity. It would be beneficial to explore methods that could convert low-quality data into high-quality data.*

✍ *During the deduplication process, methods such as exact matching and fuzzy matching are used. Fuzzy matching might lead to excessive deduplication, reducing the diversity of the samples. Additionally, applying these matching methods in large datasets can be very time-consuming.*

✍ *Filtering out toxic data can reduce the likelihood of large language models generating harmful content, but it may also diminish the models' ability to recognize toxic content. Finding a balance is necessary.*

***Subjectivity and inefficiency of manual data quality assessment (19 X).*** Some of our respondents highlighted that manual data quality assessment is highly subjective, relying heavily on individual judgment, which can result in inconsistencies. Furthermore, manual quality assessment is not only inefficient but also the assessments of sampled subsets may not accurately reflect the overall dataset.

✍ *Assessing the quality of a dataset is a challenge as it involves subjective judgement and domain knowledge.*

✍ *Manual assessments are too slow for large datasets, we need automated solutions.*

✍ *When conducting a manual assessment of a subset of samples, ensuring the representativeness of the samples is crucial, but it's often difficult to achieve manually.*

***Difficulties in quantifying data quality metrics (23 X).*** Several respondents highlighted key challenges associated with quantifying data quality metrics. Quantitative metrics, often customized by users, fail to measure the semantic quality of the text. Each domain may have unique metrics for determining high-quality data, complicating the establishment of universal metrics.

✎ *In domain-specific datasets, it is difficult to have a quantitative measure of what constitutes high-quality data and what constitutes low-quality data.*

✎ *The lack of standardized assessment metrics also adds to the complexity of the process.*

✎ *Metrics should ensure both the clarity and relevance of the data, as well as their coverage of diversity and comprehensiveness.*

✎ *Automated quality assessment is primarily rule-based and heavily relies on user customization. It is ineffective in assessing the quality of text on a semantic level.*

✎ *The knowledge content of data cannot be quantitatively measured.*

***Difficulty in data quality assessment for unstructured data (9 X).*** Some respondents have pointed out that it is more challenging to assess the quality of unstructured data.

✎ *In the quality assessment phase, it is still difficult to assess the quality of unstructured data, and the quality can only be judged by the performance of the trained model.*

✎ *Unstructured data, such as documents encompasses a variety of formats and types, each requiring different methods for data processing and evaluation.*

***Interpretability and false positive rate of built classification models for quality assessment (11 X).*** Some respondents noted a lack of interpretability in the results produced by classification models for quality assessment, and these models are prone to mislabeling high-quality data as low-quality.

✎ *When establishing a classification model for quality assessment, interpretability is low, and the reasons for identifying data as dirty are unclear. For data in new domains or not covered in the training set, the data quality score may be low, while for data from the same distribution, the score may be biased towards high quality.*

✎ *When using a classification model for quality assessment, it is easy to misclassify high-quality data as low-quality. For some partially repairable low-quality data, they are directly filtered out, which does not support the repair of low-quality data.*

***The lack of theoretical proof of small-scale model sandbox experiments (8 X).*** Several respondents emphasized the absence of theoretical validation behind utilizing small-scale model sandbox experiments for data quality assessment.

✎ *When evaluating data quality through sandbox experiments with small-scale models, there is no theoretical proof that the performance of small-scale models can be extrapolated to large-scale models.*

✎ *Sometimes, when I use a dataset to train a model with a relatively small number of parameters and evaluate the quality of the dataset based on the performance of the small-scale model, the data may improve the performance of the small-scale model but not the large-scale model.*

# 5 THREATS TO VALIDITY

Our research exclusively focuses on the characteristics of high-quality training data for single-modal LLMs designed specifically for text data, without considering multi-modal LLMs used in computer vision and speech applications. This decision arises from the current prevalence of single-modal LLMs compared to multi-modal models, making it easier to engage more practitioners specializing in single-modal LLMs for the completion of our questionnaire. Importantly, we intentionally target practitioners working with single-modal LLMs in our survey, and explicitly state in the questionnaire that our study is solely focused on this specific area.

We surveyed 219 practitioners from 23 countries spanning 5 continents. Our respondents included professionals from global IT companies, and start-ups specializing in LLMs, as well as contributors to open-source LLM projects. However, our findings may not fully capture perspectives of all professionals working with LLMs. This limitation arises from several factors: LLMs are relatively recent, and training such models requires high-performance GPUs and advanced expertise. Additionally, some companies may have privacy policies preventing employees from disclosing details about their LLM development. As a result, we collect 219 responses for this study. In the future, we aim to gather more responses to further enhance the generalizability of our conclusions.

Due to limitations in the length of our questionnaire, our survey can only provide a broad perspective on LLM practitioners' perceptions regarding the characteristics of high-quality datasets. Each characteristic may correspond to intricate and nuanced data preprocessing or collection operations. We plan to conduct in-depth interviews or case studies to delve deeper into each characteristic, aiming to gain a more comprehensive understanding.

# 6 IMPLICATIONS

## 6.1 Implications for practitioners

Given the rapid emergence and evolution of LLMs over the past two years, newcomers to the field may not fully grasp the intricate processes required to construct high-quality training datasets essential for optimizing LLM performance. Our study has systematically identified several important characteristics of high-quality datasets. Each characteristic has been rated for its perceived importance by respondents, and comprehensive rationales have been provided to explain why certain characteristics are considered more important than others. Opinions on many of these characteristics vary among our respondents, often highlighting trade-offs and specific considerations. For example, in the case of the characteristic "Absence of Toxic Data", some respondents suggest retaining a portion of toxic data and ensuring that LLMs are aware of its toxicity, so that LLMs do not produce toxic responses unintentionally. We present a balanced view of the debate to help practitioners weigh these characteristics and consider their specific contexts when deciding whether to incorporate them. We recommend novice practitioners prioritize characteristics considered most important for high-quality datasets, along with widely used data pre-processing and quality assessment methods. Focusing on these key areas can help reduce complexities associated with data pre-processing and quality assessment, streamlining the dataset construction process and enhancing training data quality. This targeted strategy not only facilitates efficient dataset construction but also lays a solid foundation for the development of effective LLMs for newcomers. In addition, we identified a number of common data pre-processing steps and data quality assessment approaches used by LLM practitioners. LLM practitioners can review these for applicability for their own LLM development and quality assurance.

## 6.2 Implications for future research

Our results highlight the potential for researchers to tackle the challenges practitioners encounter in data pre-processing and quality assessment. (1) The development of standardized and systematized workflows for data pre-processing can streamline practitioners' processes, ensuring consistency and reproducibility across various projects. (2) Conducting systematic studies to analyze how different data pre-processing techniques, implemented at varying extents, influence the robustness of LLMs' performance, offers valuable insights and guidance for practitioners. (3) Investigating the interpretability of built classification models for quality assessment provides valuable insights into guiding data pre-processing efforts. (4) Proposing more abstract quantitative metrics for evaluating data quality at the semantic level, such as knowledge content, enhances the understanding of data quality beyond surface characteristics. (5) Exploring methods for assessing data quality in unstructured data domains broadens the understanding of data quality assessment practices. (6) Researching the theoretical proof of small-scale model sandbox experiments for data quality assessment can offer valuable insights into the effectiveness and generalizability of such assessment methods. (7) The development of automated data repair methods capable of transforming low-quality data into high-quality data addresses the challenge of insufficient high-quality datasets, particularly in domains where high-quality data is scarce.

## 7 SUMMARY

Through a mixed-methods approach involving semi-structured interviews and a large-scale survey with LLM practitioners, we have identified key characteristics of high-quality LLM training datasets, widely-used data pre-processing and quality assessment methods, and common challenges encountered in these processes. Our findings offer valuable insights that can assist LLM practitioners in understanding the intricate processes required to construct high-quality training datasets. Moreover, we highlight opportunities for LLM researchers to develop solutions that better support practitioners in alleviating the challenges encountered during the construction of high-quality datasets. To facilitate replication of our study, the interview guide and questionnaire is available in our replication package [6].

## ACKNOWLEDGEMENTS

Grundy is supported by ARC Laureate Fellowship FL190100035.

## REFERENCES


[1] 2024. Apache Griffin. https://griffin.apache.org/.
[2] 2024. Deequ. https://github.com/awslabs/deequ.git.
[3] 2024. Great Expectations. https://github.com/great-expectations/great_expectations.
[4] 2024. Nvivo qualitative software. https://lumivero.com/products/nvivo/.
[5] 2024. Qualitis. https://github.com/WeBankFinTech/Qualitis.
[6] 2024. Supplemental Materials. https://doi.org/10.6084/m9.figshare.25928863.
[7] 2024. wenjuanxing software. https://www.wjx.cn.
[8] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540* (2023).
[9] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
[10] Toufique Ahmed and Premkumar Devanbu. 2023. Better patching using LLM prompting, via Self-Consistency. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 1742–1746.
[11] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A Survey on Data Selection for Language Models. arXiv:2402.16827 [cs.CL]
[12] Miltiadis Allamanis. 2019. The adverse effects of code duplication in machine learning models of code. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*. 143–153.
[13] Gabriel Amaral, Alessandro Piscopo, Lucie-Aimée Kaffee, Odinaldo Rodrigues, and Elena Simperl. 2021. Assessing the quality of sources in Wikidata across languages: a hybrid approach. *Journal of Data and Information Quality (JDIQ)* 13, 4 (2021), 1–35.
[14] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).
[15] Richard A Armstrong. 2014. When to use the B onferroni correction. *Ophthalmic and Physiological Optics* 34, 5 (2014), 502–508.
[16] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–52.
[17] Lenz Belzner, Thomas Gabor, and Martin Wirsing. 2023. Large language model assisted software engineering: prospects, challenges, and a case study. In *International Conference on Bridging the Gap between AI and Reality*. Springer, 355–374.
[18] Tamara Bondar, Hala Assal, and AbdelRahman Abdou. 2023. Why do Internet Devices Remain Vulnerable? A Survey with System Administrators. In *Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb 2023). NDSS*.
[19] Matthew Bovee, Rajendra P Srivastava, and Brenda Mak. 2003. A conceptual framework and belief-function approach to assessing overall information quality. *International journal of intelligent systems* 18, 1 (2003), 51–74.
[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[21] Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2022. The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529* (2022).
[22] Jeffrey C Carver, Oscar Dieste, Nicholas A Kraft, David Lo, and Thomas Zimmermann. 2016. How practitioners perceive the relevance of esem research. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. 1–10.
[23] Junhua Ding, Xinchuan Li, Xiaojun Kang, and Venkat N Gudivada. 2019. A case study of the augmentation and evaluation of training data for deep learning. *Journal of Data and Information Quality (JDIQ)* 11, 4 (2019), 1–22.
[24] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. 2021. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740* (2021).
[25] Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M Zhang. 2023. Large language models for software engineering: Survey and open problems. *arXiv preprint arXiv:2310.03533* (2023).
[26] Ronald A Fisher. 1922. On the interpretation of $\chi$ 2 from contingency tables, and the calculation of P. *Journal of the royal statistical society* 85, 1 (1922), 87–94.
[27] Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen-tau Yih, Luke Zettlemoyer, and Mike Lewis. 2022. Incoder: A generative model for code infilling and synthesis. *arXiv preprint arXiv:2204.05999* (2022).
[28] Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383* (2024).
[29] Junxiao Han, Shuiguang Deng, David Lo, Chen Zhi, Jianwei Yin, and Xin Xia. 2021. An empirical study of the landscape of open source projects in Baidu, Alibaba, and Tencent. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 298–307.
[30] Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. 2022. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487* (2022).
[31] Thomas N Herzog, Fritz J Scheuren, and William E Winkler. 2007. *Data quality and record linkage techniques*. Vol. 1. Springer.
[32] Yucheng Hu and Yuxing Lu. 2024. RAG and RAU: A Survey on Retrieval-Augmented Language Model in Natural Language Processing. *arXiv preprint arXiv:2404.19543* (2024).
[33] Fang Ji, Heqing Zhang, Zijiang Zhu, and Weihuang Dai. 2021. Blog text quality assessment using a 3D CNN-based statistical framework. *Future Generation*

*Computer Systems* 116 (2021), 365–370.
[34] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*. PMLR, 10697–10707.
[35] Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. 2023. Large Language Models versus Natural Language Understanding and Generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*. 278–290.
[36] Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. 2022. The stack: 3 tb of permissively licensed source code. *arXiv preprint arXiv:2211.15533* (2022).
[37] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. (2023).
[38] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499* (2021).
[39] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032* (2023).
[40] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161* (2023).
[41] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science* 378, 6624 (2022), 1092–1097.
[42] Jenny T Liang, Chenyang Yang, and Brad A Myers. 2024. A large-scale survey on the usability of ai programming assistants: Successes and challenges. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–13.
[43] Jenny T Liang, Thomas Zimmermann, and Denae Ford. 2022. Understanding skills for oss communities on github. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 170–182.
[44] David Lo, Nachiappan Nagappan, and Thomas Zimmermann. 2015. How practitioners perceive the relevance of software engineering research. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. 415–425.
[45] Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, et al. 2023. Paloma: A Benchmark for Evaluating Language Model Fit. *arXiv preprint arXiv:2312.10523* (2023).
[46] maketing evolution. 2022. What is Data Quality? Definition & Dimensions. https://www.marketingevolution.com/marketing-essentials/data-quality.
[47] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474* (2022).
[48] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. A conversational paradigm for program synthesis. *arXiv preprint arXiv:2203.13474* 30 (2022).
[49] Nostalgebraist. 2022. chinchilla's wild implications. https://www.alignmentforum.org/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications.
[50] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2024. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems* 36 (2024).
[51] Qlik. 2024. Data Quality. https://www.qlik.com/us/data-governance/data-quality.
[52] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).
[53] Thomas C Redman. 1997. *Data quality for the information age*. Artech House, Inc.
[54] Thomas C Redman. 2001. *Data quality: the field guide*. Digital press.
[55] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. 2022. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2584–2594.
[56] Xiaoxue Ren, Xinyuan Ye, Dehai Zhao, Zhenchang Xing, and Xiaohu Yang. 2023. From Misuse to Mastery: Enhancing Code Generation with Knowledge-Driven AI Chaining. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 976–987.
[57] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950* (2023).
[58] Monica Scannapieco and Tiziana Catarci. 2002. Data quality under a computer science perspective. *Archivi & Computer* 2 (2002), 1–15.
[59] Robert Sheldon. 2024. What is data management and why is it important? https://www.techtarget.com/searchdatamanagement/definition/data-quality.
[60] Fatimah Sidi, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A Jabar, Hamidah Ibrahim, and Aida Mustapha. 2012. Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval & Knowledge Management*. IEEE, 300–304.
[61] Vanessa Simard, Mikael Rönnqvist, Luc Lebel, and Nadia Lehoux. 2023. A Method to Classify Data Quality for Decision Making Under Uncertainty. *ACM Journal of Data and Information Quality* 15, 2 (2023), 1–27.
[62] Edward Smith, Robert Loftin, Emerson Murphy-Hill, Christian Bird, and Thomas Zimmermann. 2013. Improving developer participation rates in surveys. In *2013 6th International workshop on cooperative and human aspects of software engineering (CHASE)*. IEEE, 89–92.
[63] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint arXiv:2402.00159* (2024).
[64] Myles Suer. 2023. What Is Data Quality and Why Is It Important? https://www.alation.com/blog/what-is-data-quality-why-is-it-important/.
[65] Jie Tan, Daniel Feitosa, and Paris Avgeriou. 2021. Do practitioners intentionally self-fix technical debt and why?. In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 251–262.
[66] Ze Tang, Jidong Ge, Shangqing Liu, Tingwei Zhu, Tongtong Xu, Liguo Huang, and Bin Luo. 2023. Domain Adaptive Code Completion via Language Models and Decoupled Domain Databases. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 421–433.
[67] Chakkrit Tantithamthavorn, Shane McIntosh, Ahmed E. Hassan, and Kenichi Matsumoto. 2017. An Empirical Comparison of Model Validation Techniques for Defect Prediction Models. 1 (2017).
[68] Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2024. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems* 36 (2024).
[69] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[70] Yair Wand and Richard Y Wang. 1996. Anchoring data quality dimensions in ontological foundations. *Commun. ACM* 39, 11 (1996), 86–95.
[71] Junjie Wang, Yuchao Huang, Chunyang Chen, Zhe Liu, Song Wang, and Qing Wang. 2024. Software testing with large language models: Survey, landscape, and vision. *IEEE Transactions on Software Engineering* (2024).
[72] Richard Y Wang and Diane M Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33.
[73] Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi DQ Bui, Junnan Li, and Steven CH Hoi. 2023. Codet5+: Open code large language models for code understanding and generation. *arXiv preprint arXiv:2305.07922* (2023).
[74] Pavlína Wurzel Gonçalves, Gül Calikli, Alexander Serebrenik, and Alberto Bacchelli. 2023. Competencies for code review. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–33.
[75] Xin Xia, Zhiyuan Wan, Pavneet Singh Kochhar, and David Lo. 2019. How practitioners perceive coding proficiency. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 924–935.
[76] Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. 1–10.
[77] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* 18, 6 (2024), 1–32.
[78] Quanjun Zhang, Chunrong Fang, Yang Xie, Yaxin Zhang, Yun Yang, Weisong Sun, Shengcheng Yu, and Zhenyu Chen. 2023. A Survey on Large Language Models for Software Engineering. *arXiv preprint arXiv:2312.15223* (2023).
[79] Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568* (2023).
[80] Zibin Zheng, Kaiwen Ning, Yanlin Wang, Jingwen Zhang, Dewu Zheng, Mingxi Ye, and Jiachi Chen. 2023. A survey of large language models for code: Evolution, benchmarking, and future trends. *arXiv preprint arXiv:2311.10372* (2023).
[81] Jie Zhu, Jiyang Qi, Mingyu Ding, Xiaokang Chen, Ping Luo, Xinggang Wang, Wenyu Liu, Leye Wang, and Jingdong Wang. 2023. Understanding self-supervised pretraining with part-aware representation learning. *arXiv preprint*

*arXiv:2301.11915* (2023).