Data Management For Training Large Language Models: A Survey

Anonymous ACL submission

Abstract

Data plays a fundamental role in training Large Language Models (LLMs). Efficient data management, particularly in formulating a well-suited training dataset, is significant for enhancing model performance and improving training efficiency during pretraining and supervised fine-tuning stages. Despite the considerable importance of data management, the underlying mechanism of current prominent practices are still unknown. Consequently, the exploration of data management has attracted more and more attention among the research community. This survey aims to provide a comprehensive overview of current research in data management within both the pretraining and supervised fine-tuning stages of LLMs, covering various aspects of data management strategy design. Looking into the future, we extrapolate existing challenges and outline promising directions for development in this field. Therefore, this survey serves as a guiding resource for practitioners aspiring to construct powerful LLMs through efficient data management practices.

1 Introduction

Large Language Models (LLMs) have shocked the natural language processing (NLP) community with their strong performance and emergent abilities (OpenAI, 2023; Touvron et al., 2023a; Wei et al., 2022). According to previous studies (Kaplan et al., 2020; Hoffmann et al., 2022b), LLMs' achievements depend heavily on self-supervised pretraining over processed vast volumes of text data. Recent research (Zhou et al., 2023a; Ouyang et al., 2022) further enhances LLMs' instruction-following ability and performance on downstream tasks through Supervised Fine-Tuning (SFT) on deliberately curated instruction datasets.

To construct suitable training datasets, data management is vitally important and challenging in both the pretraining and SFT stages of LLMs, which we define as following:

> **Data management:** the process of organizing a well-suited training dataset with collected data, including the data selection, combination and utilization strategies, and the evaluation of the chosen strategies.

In the pretraining stage, constructing datasets with high-quality data is essential for efficient training (Jain et al., 2020; Gupta et al., 2021). To equip LLMs with diverse and comprehensive abilities, heterogeneous dataset composition with mixtures of domains is also required (Gao et al., 2020; Longpre et al., 2023b; Shen et al., 2023). However, many prominent LLMs do not enclose (Anil et al., 2023; OpenAI, 2023) or only document (Brown et al., 2020; Workshop et al., 2022; Touvron et al., 2023a) the techniques used in the construction of their pretraining dataset, leaving the reasons and effects of choosing specific data management strategies absent. In the SFT stage, LLMs' performance and instruction-following abilities are primarily evoked by carefully constructed instruction datasets (Sanh et al., 2022; Ouyang et al., 2022). Although a handful of instruction datasets/benchmarks have been proposed (Wang et al., 2022, 2023c; Taori et al., 2023; Anand et al., 2023), practitioners still find it confusing about the effects of instruction datasets on the performance of fine-tuned LLMs, leading to difficulties in choosing proper data management strategies in LLM SFT practices. To address the sparsity problem of existing data, collecting data from multimodal source (Zhang et al., 2023a; Yang et al., 2023b) and model synthesis (Maini et al., 2024; Li et al., 2024a) rise as new trends.

To address these challenges, researchers try to discover and explore the underlying principles of data management. With more and more works been proposed to address different aspects, it is necessary to conduct a systematic discussion considering the whole picture. This survey aims to provide a comprehensive overview of current research in LLM data management and a guiding resource to practitioners attempting to build powerful LLMs with efficient data management practices.

In Section 2 and 3, we respectively discuss current research in the pretraining and SFT stages of LLMs, covering multiple aspects in data management like domain/task composition, data quality, data quantity, etc., as shown in Figure 3. However, there still lacks a well-established and acknowledged general data management pipeline. Hence, We hope our work can inspire future research to establish and analyze such general pipelines. With the vision that the development of data management should keep pace with that of LLMs' abilitites, we present more existing challenges and promising future directions in Section 4.

2 Pretraining of LLM

Data management is found to be important in the pretraining stage of many prominent LLMs (OpenAI, 2023; Touvron et al., 2023a; Wei et al., 2022). In this section, we will discuss works trying to explore data management in the pretraining stage of LLMs, including domain composition, data quantity and data quality, as shown in Figure 1(a). Strategies adopted by prominent pretrained models are listed in Table 1.

2.1 Domain Composition

Public available pretraining datasets (Gao et al., 2020) usually contain mixtures of data collected from multiple sources and domains. Many prominent models (Du et al., 2022; Gao et al., 2023; Zhang et al., 2023a) are also trained on a mixture of data from different domains. Figure 2 summarizes the revealed domain mixture ratios in the pretraining datasets of prominent models.

Early pretraining corpus mostly contain data with high diversity (Web and Wiki). With recent emphasis on the data quality and the requirement for advanced abilities, high quality text (Books and academic text) are integrated. Most recently, with improved importance of Coding LLM and essential finding that code-based pretraining can enhance reasoning capability of LLM, domain data like code and math take up higher ratio of the total pretraining data. A trend can be concluded that more and more domains are included to pretrain LLMs with more various and powerful abilities. The benefits of multi-domain composition are also proved in a recent study (Longpre et al., 2023b).

Proper domain mixture ratio is also important in the pretraining of LLMs. Early attempts usually found the ratio by elaborated experiments and intuitions (Gao et al., 2020; Du et al., 2022; Thoppilan et al., 2022). Recently, domain generalization techniques are leveraged to automatically assign domain weights to form a suitable target distribution, such as importance resampling (Xie et al., 2023b) and Group Domain Robust Optimization (Xie et al., 2023a). Contribution of each domain measured via gradients is also adopted to reweight domains (Fan et al., 2023). Xia et al. (2023) assign batch-level weights dynamically based on varying losses. Ye et al. (2024) propose data mixing laws to predict model performance with different mixing ratios.

Although proper domain composition is broadly acknowledged as beneficial in the pretraining of LLMs, some empirical analyses arrive at different conclusions and leave open questions for future research. For example, Longpre et al. (2023b) claim that the inclusion of diverse web domains may perform better than specific mixtures in certain tasks. *CodeGen2* (Nijkamp et al., 2023) studies programming and natural language mixtures and finds that models trained with mixtures do not perform better than but closely to domain-matched models given the same computing budget.

2.2 Data Quantity

It is recognized that the pretraining of LLMs needs large amounts of data. Scaling laws are proposed to depict the relationships between data quantity and model size. Repeatedly training on data is also studied due to data exhaustion.

2.2.1 Scaling Laws

Before the popularization of LLMs, the relationship between training dataset size and the performance of Transformer-based language models (Vaswani et al., 2017) had already attracted researchers' attention. Kaplan et al. (2020) find that the language model loss has a power-law relationship with training dataset size or model size, respectively, when not bottlenecked by each other and the training computing budget. They further depict the dependence between model size and training dataset size as:

$$L(N, D) = \left[\left(\frac{N_c}{N}\right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_c}{D}\right]^{\alpha_D} \tag{1}$$

where L is the language model test loss, D is the number of training tokens, N is the number of

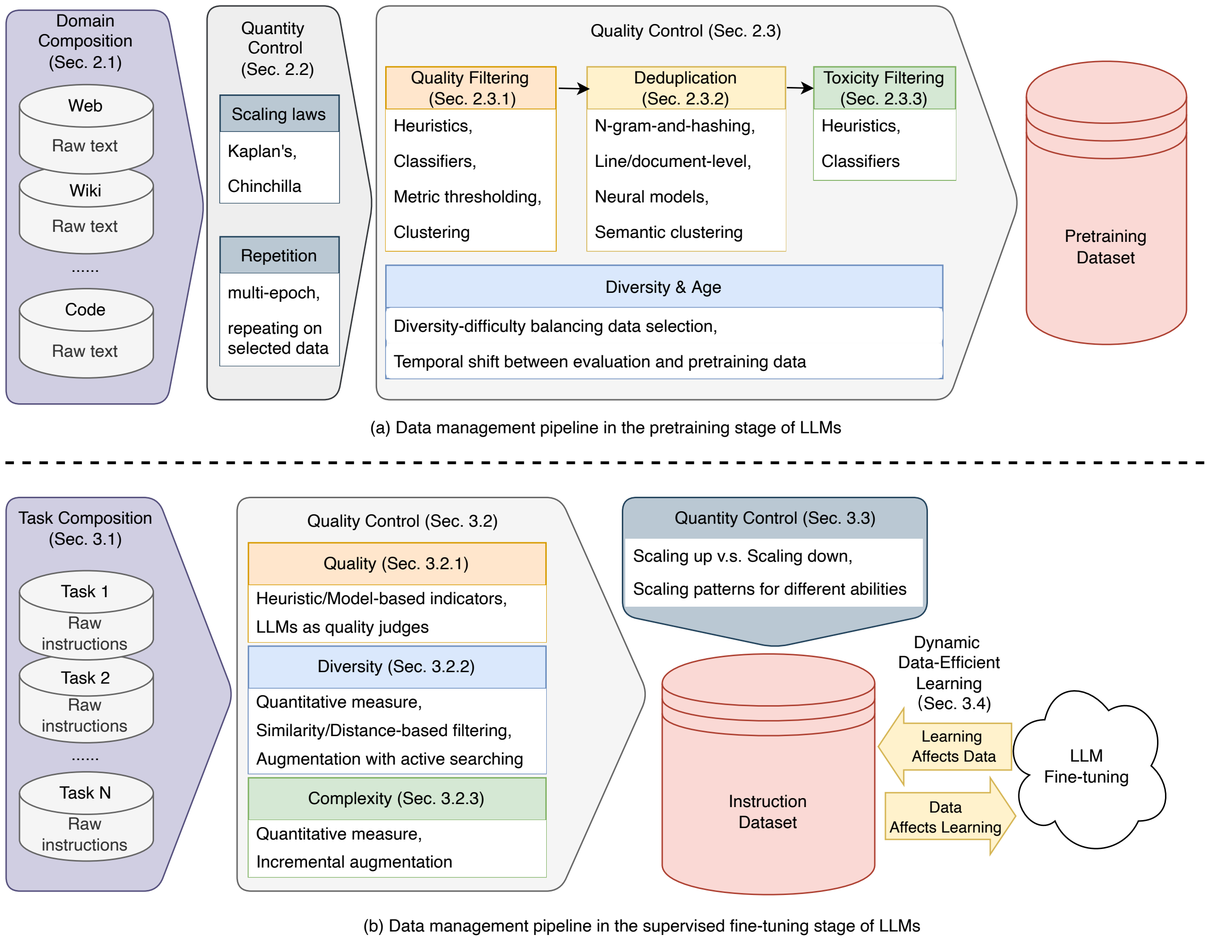

Figure 1: Data management pipelines for the pretraining and supervised fine-tuning of Large Language Models.

model parameters, α_D and α_N are the power-law components for the scaling of D and N, respectively, and D_c and N_c are constant numbers [1].

Fitting Equation 1, they conclude that model loss decreases predictably as long as the model size and training dataset size are scaled up simultaneously. Still, overfitting will happen if either of them is fixed while the other increases. Given fixed computing budget C, they analyze the optimal allocation of $D_{opt} \sim C^{0.27}$ and $N_{opt} \sim C^{0.73}$, showing that the model size should increase faster than the training dataset size.

Following Kaplan et al. (2020), Hoffmann et al. (2022b) conduct experiments on much larger language models and arrive at a new scaling law, usually called as *Chinchilla Scaling Law*:

$$L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta} \tag{2}$$

where they empirically fit $E = 1.69$, $A = 406.4$, $B = 410.7$, $\alpha = 0.34$ and $\beta = 0.28$. The optimal allocation of D_{opt} and N_{opt} are also analyzed as $D_{opt} \sim C^{0.54}$ and $N_{opt} \sim C^{0.46}$. Hence, they draw a different conclusion that model and training dataset sizes should scale roughly at the same rate with a larger computing budget. Su et al. (2024) dig deeper into Kaplan's scaling laws and provide more detailed instructions to fit the constants.

2.2.2 Data Repetition

While Kaplan et al. (2020) and Hoffmann et al. (2022b) both focus on scaling laws with unique data trained only for one epoch, Hernandez et al. (2022) study the scaling laws with a small fraction of repeated data in the training dataset and find that the text overlap may be harmful to model performance, causing a divergence from Kaplan's scaling law on model size larger than 100M parameters.

With the models grow larger and larger, data has becoming more and more demanding, raising concerns about the exhaustion of high-quality training data (Villalobos et al., 2022; Hoffmann et al., 2022b). Addressing these concerns, several works study the consequence of repeatedly pretraining on the whole datasets for multiple epochs. Scaling law

[1]The precise numerical values of D_c and N_c depend on vocabulary size and tokenization and do not have fundamental meaning.

Figure 2: The domain composition of prominent Large Language Models.

on repeated training is proposed to depict the diminishing of returns with more repetition and larger model sizes (Muennighoff et al., 2023) and shows a multi-epoch degradation phenomenon (Xue et al., 2023). Further analysis digs out that dataset size, model parameters, and training objectives are the key factors to this phenomenon, and classic regularization techniques may not be helpful, except for dropout (Xue et al., 2023).

There are still positive results in the research of data repetition. Muennighoff et al. (2023) find that repeatedly training on the whole dataset up to 4 epochs only causes trivial harm to test loss compared to training on unique new data. Instead of simply repeating over the whole dataset, Tirumala et al. (2023) show that repeatedly training on carefully selected data can outperform that on randomly selected new data, suggesting a feasible way of repeating on intelligently selected data.

Recently, pretraining with mixed real and synthesized data is adopted to meet the data exhaustion challenge (Javaheripi and Bubeck, 2023; Meta, 2024). It is also gaining more an more attention and develops into a new trend as data synthesize.

2.3 Data Quality

In the pretraining of LLMs, Quality control techniques of the pretraining datasets usually form an order (Rae et al., 2021; Nguyen et al., 2023; Tirumala et al., 2023; Gan et al., 2023), namely quality filtering, deduplication and toxicity filtering. Data diversity and age are also explored.

2.3.1 Quality Filtering

Public datasets like Common Crawl [2] and multilingual datasets (Kreutzer et al., 2022) usually contain low-quality data that hampers the training of LLMs. Hence, existing works usually perform quality filtering using hand-crafted heuristics (Yang et al., 2019; Raffel et al., 2020; Nijkamp et al., 2022), a trained classifier (Brown et al., 2020; Gao et al., 2020; Du et al., 2022; Touvron et al., 2023a; Wettig et al., 2024), metric thresholding (Wenzek et al., 2020; Muennighoff et al., 2023) or combinations of these techniques. Besides instance-level filtering, embedding clustering is also adopted to filter one cluster at a time (Kaddour, 2023).

Despite the reduction of training data quantity, quality filtering is usually proven to be beneficial in model performance improvement (Longpre et al., 2023b). Several carefully filtered high-quality datasets are proposed to train lightweight language models and achieve outstanding performances (Gunasekar et al., 2023; Li et al., 2023d; Javaheripi and Bubeck, 2023; Penedo et al., 2023). However, Gao (2021) finds that aggressive filtering might lead to performance degradation on a wide range of tasks for GPT-like LLMs due to the poor representativity of the filtering proxy objectives. To address this issue, Marion et al. (2023) comprehensively examine different data quality estimators and find that pruning datasets based on perplexity performs better than more complicated techniques like memorization. Gan et al. (2023) develop data-centric scaling laws and show that improving semantic and grammatical quality is more effective. However, there still lacks a well-established and theoretically efficient filtering strategy, leaving room for further exploration.

2.3.2 Deduplication

Deduplication is a necessary step in many LLMs' pretraining data management procedures and the preprocessing of many publicly available datasets (Brown et al., 2020; Workshop et al., 2022; Touvron et al., 2023a; Raffel et al., 2020). Lee et al. (2021) find that deduplication is beneficial in mem-

[2]https://commoncrawl.org/, a large text corpus contains raw web page data, metadata extracts, and text extracts.

orization mitigation, train-test overlap avoidance, and training efficiency improvement while keeping model perplexity. Kandpal et al. (2022) also show that deduplication can considerably lower the success rate of privacy attacks aiming at model memorization.

Among practices of deduplication, N-gram-and-hashing is the most commonly adopted technique (Lee et al., 2021; Borgeaud et al., 2022; Rae et al., 2021). It can operate at line-level (Touvron et al., 2023a), document-level (Hoffmann et al., 2022a; Li et al., 2022b) or combinations of them. Recently, neural models are experimentally proven to outperform traditional N-gram-and-hashing methods (Silcock et al., 2022). Addressing semantic deduplication, Abbas et al. (2023) propose *SemDeDup* to remove semantic duplicates that lie closely in the pretrained model's embedding space and apply clustering to reduce the searching computation.

2.3.3 Toxicity Filtering

Toxicity refers to the text content which is *"rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion"* (Gehman et al., 2020; Welbl et al., 2021). As raw text corpora usually contain toxic text (Luccioni and Viviano, 2021; Longpre et al., 2023b), toxicity filtering aims to remove text with undesirable toxic text in the pretraining datasets, further preventing LLMs from generating toxic utterances. Similar to quality filtering, heuristic and rule-based filtering (Lees et al., 2022; Gargee et al., 2022; Friedl, 2023) and N-gram classifiers (Raffel et al., 2020) are usually adopted as toxicity filters.

Although effective in model detoxifying, Longpre et al. (2023b) discover that toxicity filtering reduces the risk of toxic generation by sacrificing model generalization and toxicity identification ability. Moreover, Xu et al. (2021) and Welbl et al. (2021) find that training dataset detoxification leads to the marginalization of minority groups like dialects and minority identity mentions, posing challenges in building unbiased LLMs.

2.3.4 Data Diversity

Some works focus on other aspects of data management in the pretraining stage of LLMs. Lee et al. (2023a) show that the format diversities of publicly available pretraining datasets are high when measured by Task2Vec diversity coefficient (Miranda et al., 2022). Maharana et al. (2023) propose *D2 Pruning* to balance data diversity and difficulty in data selection by representing datasets as undirected graphs and adopting forward-and-reverse message passing strategy to select a subgraph enveloping both diverse and difficult data samples.

2.3.5 Data Age

In current practices, more recent LLMs are usually pretrained using newer data [3]. Some knowledge learned by pretrained LLMs could also be time-sensitive. Longpre et al. (2023b) study the impact of data age and find that the temporal shift between evaluation and pretraining data will lead to inaccurate performance estimation. This temporal misalignment might not be overcome by fine-tuning, especially for larger models.

2.4 Relations Among Domain Composition, Data Quantity and Data Quality

Recently, several scaling laws are proposed to explore the synergistic effect of different aspects on the pretrained model performance, such as the bivariate model performance prediction regarding data quantity and domain composition ratio (Ge et al., 2024a), the quality-quantity tradeoff under different computing budget (Goyal et al., 2024), and the positive correlation between data quality and model scale under the same data quantity (Bi et al., 2024). What's more, Shen et al. (2023) emphasize global deduplication to remove overlaps among different domains. Longpre et al. (2023b) claim that domains with high quality and diversity are more beneficial than other domains.

3 Supervised Fine-Tuning of LLM

Based on the general knowledge and capabilities learned in the pretraining stage, supervised fine-tuning (SFT) is proposed to further improve LLMs with instruction-following ability and alignment with human expectations (Wei et al., 2021; Sanh et al., 2022; Ouyang et al., 2022). Although LLMs fined-tuned with existing instruction datasets have achieved remarkable performance in various NLP tasks, the impacts of instruction data management on fine-tuned models are still under debate. The data management process in the SFT stage can be summarized as illustrated in Figure 1(b), including task composition, data quality control, data quantity control and dynamic data-efficient learning. Table 2 summarizes the data management practices of prominent fine-tuned LLMs.

[3] https://platform.openai.com/docs/models

3.1 Task Composition

Since LLMs have shown surprisingly emergent abilities in handling various NLP tasks, multitask fine-tuning appears to be promising to improve LLMs' generalization performance on unseen tasks. The benefits of increasing the number of tasks in SFT have been experimentally proven on models with different sizes ranging from 3B to 540B parameters (Wang et al., 2022; Sanh et al., 2022; Wei et al., 2021; Chung et al., 2022). With the scaling of tasks, the mixture ratio of data targeting different tasks is also found to be critical and usually decided by experiments and intuitions (Iyer et al., 2022; Longpre et al., 2023a). To enable LLMs to solve targeted tasks with specific skills, representation similarity (Ivison et al., 2023; Lee et al., 2024) and gradient similarity (Xia et al., 2024) is proposed to select relevant multitask subsets.

However, conflicts might exist among the many tasks. Dong et al. (2023) focus on task composition among mathematical reasoning, code generation, and general human-aligning abilities. They find that model abilities are improved when the mixed data amount is small but decreased otherwise. The negative impact of large amount mixing data might lie in the similarity degree of data format and data distribution among different SFT tasks. Wang et al. (2023b) also experimentally show that different instruction datasets may correspond to different specific abilities. And winning across all evaluations using a single dataset or combination seems to be challenging.

Divergent from compositing multiple tasks, some works claim that integration of LLMs tuned on single task data can outperform one LLM tuned on multiple tasks (Jang et al., 2023; Chen et al., 2023b). But fine-tuning more task-specific LLMs also means more resource consumption. How to efficiently equip LLMs with the ability to solve multiple tasks still demands more exploration.

3.2 Data Quality

Data quality is always a focal point in the SFT of LLMs, addressing instruction quality, diversity, and complexity. Here, we focus more on managing and analyzing existing instruction data instead of instruction generation methods discussed in previous surveys (Zhang et al., 2023b; Wang et al., 2023e).

3.2.1 Instruction Quality

Many researchers have found that the quality of instruction data is one of the most important factors in improving model performance (Chia et al., 2023; Zhou et al., 2023a; Ding et al., 2023). During the construction of instruction dataset, there is usually a filtering step to select high-quality instructions generated by models.

Heuristic- and model-based natural language indicators like perplexity and uncertainty are commonly adopted filtering criteria (Wang et al., 2023d; Cao et al., 2023; Bhatt et al., 2024). What's more, losses (Zhou et al., 2023b; Li et al., 2023b, 2024b) and output probabilities (Li et al., 2023a,e; Chen and Mueller, 2024; He et al., 2024b; Liu et al., 2024) of LLMs are adopted to compute more complex scores for data selection. Popular searching approaches like BlendSearch (Wang et al., 2020) are also leveraged to find high-quality instructions satisfying the criteria (Cao et al., 2023).

In addition, LLMs are also queried to directly evaluate the quality of instructions. Fine-tuned LLMs are prompted to assign quality scores (Li et al., 2023c) or provide self-feedback (Lu et al., 2023a; Madaan et al., 2023) to their own responses to iteratively improve model prediction. Strong LLMs like ChatGPT (Ye et al., 2023; Chen et al., 2023c; Li et al., 2023a) or reward models (Du et al., 2023) are also adopted as quality judges during instruction data filtering. Recently, weak-to-strong strategy is introduced to select high-quality data with smaller and weaker models (Li et al., 2024c; Yang et al., 2024; Mekala et al., 2024).

3.2.2 Instruction Diversity

The intention and semantic diversity of instructions is another important factor that has shown positive effects on model performance improvement and robustness (Zhou et al., 2023a; Ding et al., 2023; Taori et al., 2023; Bukharin and Zhao, 2023). However, there is no well-acknowledged measurement to quantitatively indicate the diversity of an instruction dataset. *#InsTag* (Lu et al., 2023b) propose to measure instruction diversity using fine-grained tags generated by ChatGPT [4]. Specifically, it quantifies instruction diversity as the unique tag coverage rate in the overall tag set.

To maintain both diversity and data-efficiency in the instruction datasets, Rouge-L similarity (Wang et al., 2023c), embedding distance (Wu et al., 2023; Bukharin and Zhao, 2023; Huang et al., 2024) and scoring models (Ge et al., 2024b) are proposed to select instructions that are different from each other in literal, semantic and human-aligning level.

[4] https://chatgpt.openai.com/

Due to data constraints, diversity can be challenging in some domain-specific tasks. Thus, Wan et al. (2023) propose to enlarge the data coverage through active searching variations and possibilities of instructions using LLMs.

3.2.3 Instruction Complexity

Instruction complexity is found to be crucial in developing LLMs with complex instruction-following and reasoning abilities (Xu et al., 2023a; Luo et al., 2023b; Mukherjee et al., 2023; He et al., 2024a). Several works endeavor to quantify and evaluate instruction complexity. Using aforementioned tags, *#InsTag* (Lu et al., 2023b) quantifies complexity as the average tag number assigned to each query in a dataset. He et al. (2023) evaluate complex instruction with eight features addressing the length, contents, and formats of input texts and task descriptions.

It is also empirically showed that complexity enhancement is necessary for performance improvement (Zhao et al., 2023b). To increase the instruction complexity in SFT datasets, some works propose to incrementally augment existing instructions by adding nodes to semantic tree (Zhao et al., 2023b) or performing operations such as increasing reasoning, adding constraints, in-breadth evolving, deepening, and so on (Xu et al., 2023a; Luo et al., 2023b; Jiang et al., 2023b; Sun et al., 2024a).

3.3 Data Quantity

Different with the acknowledged scaling laws of pretraining data, explorations of the relationship between scaling instruction data quantity and fine-tuned model performance diverge in two directions. In the earlier stage, researchers follow the observations in the pretraining of LLMs and argue that scaling up the instruction data quantity is crucial for success (Wei et al., 2021; Sanh et al., 2022). Recently, more works claim that data quality is more important than data quantity in the SFT of LLMs, and propose to scaling down the instruction datasets with limited high-quality data (Zhou et al., 2023a; Chen et al., 2023b). However, Zhang et al. (2024) propose a power-based multiplicative joint scaling law, showing that increased fine-tuning data could lead to improved model performance after achieving good results with limited data.

Addressing this conflict, several works attempt to analyze the scaling patterns for different tasks or different model abilities. A consensus of these works is that different abilities have different scaling patterns and develop at different paces. Dong et al. (2023) find that general ability can be enhanced with about 1,000 samples and improves slowly after then, while mathematical reasoning and code generation improve consistently with the increasing of instruction data amount. Similarly, Yuan et al. (2023) observe a log-linear relation between instruction data amount and models' mathematical reasoning performance, but stronger pretrained models improve less with more instruction data. Surprisingly, the empirical study of Ji et al. (2023) on 12 major real-world online user cases draws to an exactly opposite point. Song et al. (2023) also show that some abilities have completely different patterns from others.

3.4 Dynamic Data-Efficient Learning

While works discussed above focus more on the static management of instruction datasets without interaction with model fine-tuning, some works try to combine data selection with model fine-tuning, achieving data-efficient learning in a dynamic way.

Training affects data. Some works propose to dynamically change the datasets along with the fine-tuning process. Attendu and Corbeil (2023) propose a dynamic data pruning method that periodically filters out unimportant examples during SFT using extended versions of EL2N metric (Paul et al., 2021; Fayyaz et al., 2022). AlShikh et al. (2023) predict the responses as "answer-like or not" by a binary classifier, in order to measure LLMs' instruction-following ability and serve as an early-stopping criterion. Kung et al. (2023) conduct active task searching to select informative tasks based on prompt uncertainty and fine-tune in a loop.

Data affects training. Instead of manipulating instruction datasets, some works propose special training strategies to accommodate the datasets. To mitigate forgetting and negative task impact, Yin et al. (2023a) and Wang et al. (2024) treat task selection as a replay strategy in continual learning scenarios; *DMT* (Dong et al., 2023) learns specialized and general abilities sequentially while keeping a small proportion of specialized data. To efficiently learn mixed-quality data acquired from LLMs with different level of abilities, *OpenChat* (Wang et al., 2023a) proposes *C-RLFT* strategy that considers different data sources as coarse-grained reward labels; Xu et al. (2023b), Sun et al. (2024a) and Kim and Lee (2024) propose to make the model progressively learn from easy to hard, respectively regarding different data quality, instruction complexity and task hardness.

3.5 Relations Among Task composition, Data Quality and Data Quantity

Similar as in the pretraining stage, different aspects of supervised fine-tuning data management can affect model performance jointly. Lu et al. (2023b) analyze popular open-set SFT datasets using *#InsTag* and show that larger dataset sizes tend to be more diverse and induce higher performance. Current research on data selection tends to uniformly consider instruction quality and diversity (Bukharin and Zhao, 2023; Xu et al., 2023c). Since different model abilities have different scaling patterns as discussed in Section 3.3, more efficient task composition strategies are required to build stronger multi-task LLMs.

In summary, we provide a list of takeaways in Appendix A. Some other aspects of data management are discussed in Appendix B.

4 Challenges and Future Directions

The exploration of data management and its impact on LLM pretraining and SFT is still an ongoing task. In this section, we point out several challenges and corresponding future directions in the research of training data management for LLMs.

General data management framework Although data management systems are proposed to compose various data recipes in either the pretraining or SFT stage of LLM (Chen et al., 2023a; Zhou et al., 2023c; Sun et al., 2024b), practitioners still need to spend efforts on organizing suitable datasets. A well-established general data management framework suitable for a broad range of applications is an urgent and worthy future direction in developing and promoting LLMs.

Beyond that, a more autonomous data management system is also needed to greatly save human efforts. To build such systems, LLMs might be leveraged and serve as different roles such as quality examinator, data augmentor, and so on.

Data debiasing and detoxifying Current pretraining corpora and instruction datasets might contain harmful information and social biases, which lead to negative social impacts and undesirable model behavior. With the application of LLMs keeps extending to more demanding fields, the fairness and harmlessness of LLMs will become more and more innegligible. Hence, as one way to build ideal LLMs without biases and harmful output, debiasing and detoxifying of pretraining and instruction data is an important research direction.

Multimodal data management Current research in data management mostly focuses on natural language processing. With the application of LLMs extending to modalities like vision, audio, etc., it is necessary to see the impacts of multimodal data management on the performance of fine-tuned multimodal LLMs.

Data management for LLM self-exploration The ability to actively explore the unknown environment and tasks is one of the future perspectives in LLM development. Learning from large-scale interaction data requires efficient data management system to construct suitable datasets.

Efficient filtering for synthesized data As data annotation requires intensive human labors and existing data will be exhausted, automatically synthesizing new data using LLMs is newly proposed as a promising solution (Maini et al., 2024; Li et al., 2024a). In this process, efficient filtering for synthesized data is required to ensure its quality.

Fine-grained data ordering Some works start to pay attention to the ordering of data in both the pretraining (Gan et al., 2023; Guo et al., 2024) and SFT stage (Xu et al., 2023b; Yin et al., 2023a). It is shown that more fine-grained data ordering could be beneficial to model performance improvement.

Conflicted data separation In multi-task fine-tuning, negative impact of mixing data is observed and attribute to conflicts among different task data (Dong et al., 2023). Thus, separating and effectively learning from conflicted data samples is a challenging problem in multi-task learning.

5 Conclusions

This paper overviews the training data management of LLMs. We discuss the *pretraining* and *supervised fine-tuning* stages of LLM successively and summarize the up-to-date research efforts according to the data management process of each stage. Finally, we highlight several challenges and future directions for LLM training data management. We hope this survey can provide insightful guidance for practitioners and inspire further research in efficient training data management for the development of LLMs.

Limitations

In this survey, we provide an overview of training data management for LLMs. Despite our best efforts, there may still be several limitations remaining in our work.

The exploration of training data management expands across a wide range of datasets from different sources, models with different architectures and sizes, and tasks addressing the different abilities of LLMs. Due to the page limit, we do not include the technical details for each work, which may lead to certain confusion. Thus, we recommend interested researchers to read specific papers for more information.

As the research of LLMs develops vigorously, works are published or preprinted at a rapid speed. We tried our best to cover the up-to-date works proposed in the recent two years, but some works may be inevitably missed in this survey. We will continually pay close attention to the latest research developments to supplement our work.

In this work, we put our main efforts into training data management for LLMs. However, the management strategy for evaluation data are also important in the development of LLMs. Here, we leave discussion in this field in our future work.

References

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*.

Waseem AlShikh, Manhal Daaboul, Kirk Goddard, Brock Imel, Kiran Kamble, Parikshith Kulkarni, and Melisa Russak. 2023. Becoming self-instruct: introducing early stopping criteria for minimal instruct tuning. *arXiv preprint arXiv:2307.03692*.

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. `https://github.com/nomic-ai/gpt4all`.

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*.

Jean-michel Attendu and Jean-philippe Corbeil. 2023. NLU on data diets: Dynamic data subset selection for NLP classification tasks. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 129–146, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Gantavya Bhatt, Yifang Chen, Arnav M. Das, Jifan Zhang, Sang T. Truong, Stephen Mussmann, Yinglun Zhu, Jeff Bilmes, Simon Shaolei Du, Kevin Jamieson, Jordan T. Ash, and Robert Nowak. 2024. An experimental design framework for label-efficient supervised finetuning of large language models. *ArXiv*, abs/2401.06692.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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*, 33:1877–1901.

Alexander W. Bukharin and Tuo Zhao. 2023. Data diversity matters for robust instruction tuning. *ArXiv*, abs/2311.14736.

Yihan Cao, Yanbin Kang, and Lichao Sun. 2023. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*.

Daoyuan Chen, Yilun Huang, Zhijian Ma, Hesen Chen, Xuchen Pan, Ce Ge, Dawei Gao, Yuexiang Xie, Zhaoyang Liu, Jinyang Gao, et al. 2023a. Data-juicer: A one-stop data processing system for large language models. *arXiv preprint arXiv:2309.02033*.

Hao Chen, Yiming Zhang, Qi Zhang, Hantao Yang, Xiaomeng Hu, Xuetao Ma, Yifan Yanggong, and Junbo Zhao. 2023b. Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning. *arXiv preprint arXiv:2305.09246*.

Jiuhai Chen and Jonas Mueller. 2024. Automated data curation for robust language model fine-tuning. *arXiv preprint arXiv:2403.12776*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023c. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1286–1305.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *ArXiv*, abs/2311.15653.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. 2023. Doge: Domain reweighting with generalization estimation. *arXiv preprint arXiv:2310.15393*.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Mohammad Taher Pilehvar, Yadollah Yaghoobzadeh, and Samira Ebrahimi Kahou. 2022. Bert on a data diet: Finding important examples by gradient-based pruning. *arXiv preprint arXiv:2211.05610*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11737–11762.

Paul Friedl. 2023. Dis/similarities in the design and development of legal and algorithmic normative systems: the case of perspective api. *Law, Innovation and Technology*, 15(1):25–59.

Ruyi Gan, Ziwei Wu, Renliang Sun, Junyu Lu, Xiaojun Wu, Dixiang Zhang, Kunhao Pan, Ping Yang, Qi Yang, Jiaxing Zhang, et al. 2023. Ziya2: Data-centric learning is all llms need. *arXiv preprint arXiv:2311.03301*.

Leo Gao. 2021. An empirical exploration in quality filtering of text data. *arXiv preprint arXiv:2109.00698*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

SK Gargee, Pranav Bhargav Gopinath, Shridhar Reddy SR Kancharla, CR Anand, and Anoop S Babu. 2022. Analyzing and addressing the difference in toxicity prediction between different comments with same semantic meaning in google's perspective api. In *ICT Systems and Sustainability: Proceedings of ICT4SD 2022*, pages 455–464. Springer.

Ce Ge, Zhijian Ma, Daoyuan Chen, Yaliang Li, and Bolin Ding. 2024a. Data mixing made efficient: A bivariate scaling law for language model pretraining. *arXiv preprint arXiv:2405.14908*.

Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Hongxia Ma, Li Zhang, Hao Yang, and Tong Xiao. 2024b. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. *arXiv preprint arXiv:2402.18191*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. 2024. Scaling laws for data filtering–data curation cannot be compute agnostic. *arXiv preprint arXiv:2404.07177*.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Nitin Gupta, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal, et al. 2021. Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 4040–4041.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. 2024a. From complex to simple: Enhancing multi-constraint complex instruction following ability of large language models. *arXiv preprint arXiv:2404.15846*.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, et al. 2023. Can large language models understand real-world complex instructions? *arXiv preprint arXiv:2309.09150*.

Yexiao He, Ziyao Wang, Zheyu Shen, Guoheng Sun, Yucong Dai, Yongkai Wu, Hongyi Wang, and Ang Li. 2024b. Shed: Shapley-based automated dataset refinement for instruction fine-tuning. *arXiv preprint arXiv:2405.00705*.

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*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022a. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022b. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030.

Hui Huang, Bing Xu, Xinnian Liang, Kehai Chen, Muyun Yang, Tiejun Zhao, and Conghui Zhu. 2024. Multi-view fusion for instruction mining of large language model. *Information Fusion*, page 102480.

Hamish Ivison, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2023. Data-efficient finetuning using cross-task nearest neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9036–9061, Toronto, Canada. Association for Computational Linguistics.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2020. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3561–3562.

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Exploring the benefits of training expert language models over instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 14702–14729. PMLR.

Mojan Javaheripi and Sébastien Bubeck. 2023. Phi-2: The surprising power of small language models. Blog post.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 2023. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023a. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2023b. Followbench: A multi-level fine-grained constraints following benchmark for large language models. *arXiv preprint arXiv:2310.20410*.

Jean Kaddour. 2023. The minipile challenge for data-efficient language models. *arXiv preprint arXiv:2304.08442*.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. 2022. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3631–3643, Seattle, United States. Association for Computational Linguistics.

Jisu Kim and Juhwan Lee. 2024. Strategic data ordering: Enhancing large language model performance through curriculum learning. *arXiv preprint arXiv:2405.07490*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Po-Nien Kung and Nanyun Peng. 2023. Do models really learn to follow instructions? an empirical study of instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.

Po-Nien Kung, Fan Yin, Di Wu, Kai wei Chang, and Nanyun Peng. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. *ArXiv*, abs/2311.00288.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Alycia Lee, Brando Miranda, and Sanmi Koyejo. 2023a. Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data. *arXiv preprint arXiv:2306.13840*.

Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. 2023b. Platypus: Quick, cheap, and powerful refinement of llms. *arXiv preprint arXiv:2308.07317*.

Changho Lee, Janghoon Han, Seonghyeon Ye, Stanley Jungkyu Choi, Honglak Lee, and Kyunghoon Bae. 2024. Instruction matters, a simple yet effective task selection approach in instruction tuning for specific tasks. *arXiv preprint arXiv:2404.16418*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8424–8445.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3197–3207.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong

Zhang, et al. 2024a. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*.

Haoran Li, Yiran Liu, Xingxing Zhang, Wei Lu, and Furu Wei. 2023a. Tuna: Instruction tuning using feedback from large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024b. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. *ArXiv*, abs/2402.10110.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024c. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *ArXiv*, abs/2402.00530.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023b. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *ArXiv*, abs/2308.12032.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Cheng-Jie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022a. How pre-trained language models capture factual knowledge? a causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1720–1732.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023c. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023d. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022b. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. 2023e. One shot learning as instruction data prospector for large language models. *ArXiv*, abs/2312.10302.

Shihao Liang, Kunlun Zhu, Runchu Tian, Yujia Qin, Huadong Wang, Xin Cong, Zhiyuan Liu, Xiaojiang Liu, and Maosong Sun. 2023. Exploring format consistency for instruction tuning. *arXiv preprint arXiv:2307.15504*.

Liangxin Liu, Xuebo Liu, Derek F Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024. Selectit: Selective instruction tuning for large language models via uncertainty-aware self-reflection. *arXiv preprint arXiv:2402.16705*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023a. The flan collection: Designing data and methods for effective instruction tuning. In *ICML*.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2023b. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*.

Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Fei Mi, Baojun Wang, Weichao Wang, Lifeng Shang, and Qun Liu. 2023a. Self: Language-driven self-evolution for large language model. *arXiv preprint arXiv:2310.00533*.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023b. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models.

Alexandra Sasha Luccioni and Joseph D Viviano. 2021. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023b. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. 2023. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. *arXiv preprint arXiv:2310.07931*.

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1878–1898.

Dheeraj Mekala, Alex Nguyen, and Jingbo Shang. 2024. Smaller language models are capable of selecting instruction-tuning training data for larger language models. *arXiv preprint arXiv:2402.10430*.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Brando Miranda, Patrick Yu, Yu-Xiong Wang, and Sanmi Koyejo. 2022. The curse of low task diversity: On the failure of transfer learning to outperform maml and their empirical equivalence. *arXiv preprint arXiv:2208.01545*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2023. Codegen2: Lessons for training llms on programming and natural languages. *arXiv preprint arXiv:2305.02309*.

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. In *The Eleventh International Conference on Learning Representations*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

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*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and Eric Xing. 2023. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*.

Emily Silcock, Luca D'Amico-Wong, Jinglin Yang, and Melissa Dell. 2022. Noise-robust de-duplication at scale. In *The Eleventh International Conference on Learning Representations*.

Chiyu Song, Zhanchao Zhou, Jianhao Yan, Yuejiao Fei, Zhenzhong Lan, and Yue Zhang. 2023. Dynamics of instruction tuning: Each ability of large language models has its own growth pace. *arXiv preprint arXiv:2310.19651*.

Hui Su, Zhi Tian, Xiaoyu Shen, and Xunliang Cai. 2024. Unraveling the mystery of scaling laws: Part i. *arXiv preprint arXiv:2403.06563*.

Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, and Ruohui Huang. 2024a. Conifer: Improving complex constrained instruction-following ability of large language models. *arXiv preprint arXiv:2404.02823*.

Yiding Sun, Feng Wang, Yutao Zhu, Wayne Xin Zhao, and Jiaxin Mao. 2024b. An integrated data processing framework for pretraining foundation models. *arXiv preprint arXiv:2402.16358*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. *arXiv preprint arXiv:2308.12284*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Fanqi Wan, Xinting Huang, Tao Yang, Xiaojun Quan, Wei Bi, and Shuming Shi. 2023. Explore-instruct: Enhancing domain-specific instruction coverage through active exploration. *arXiv preprint arXiv:2310.09168*.

Chi Wang, Qingyun Wu, Silu Huang, and Amin Saied. 2020. Economic hyperparameter optimization with blended search strategy. In *International Conference on Learning Representations*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. 2024. Inscl: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. *arXiv preprint arXiv:2403.11435*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Yue Wang, Xinrui Wang, Juntao Li, Jinxiong Chang, Qishen Zhang, Zhongyi Liu, Guannan Zhang, and Min Zhang. 2023d. Harnessing the power of david against goliath: Exploring instruction data generation without using closed-source models. *arXiv preprint arXiv:2308.12711*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023e. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. *arXiv preprint arXiv:2310.13486*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2447–2469.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. 2023. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2023. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023a. Doremi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023b. Data selection for language models via importance resampling. *arXiv preprint arXiv:2302.03169*.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2390–2397.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Canwen Xu, Corby Rosset, Luciano Del Corro, Shweti Mahajan, Julian McAuley, Jennifer Neville, Ahmed Hassan Awadallah, and Nikhil Rao. 2023b. Contrastive post-training large language models on data curriculum. *arXiv preprint arXiv:2310.02263*.

Yang Xu, Yongqiang Yao, Yufan Huang, Mengnan Qi, Maoquan Wang, Bin Gu, and Neel Sundaresan. 2023c. Rethinking the instruction quality: Lift is what you need. *ArXiv*, abs/2312.11508.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. 2023. To repeat or not to repeat: Insights from scaling llm under token-crisis. *arXiv preprint arXiv:2305.13230*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023a. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023b. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*.

Yu Yang, Siddhartha Mishra, Jeffrey N Chiang, and Baharan Mirzasoleiman. 2024. Smalltolarge (s2l): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. *arXiv preprint arXiv:2403.07384*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Yunhua Zhou, Jun Zhan, and Xipeng Qiu. 2024. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint arXiv:2403.16952*.

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. 2023. Selfee: Iterative self-revising llm empowered by self-feedback generation. Blog post.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023a. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. *arXiv preprint arXiv:2305.14327*.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023b. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3063–3079, Toronto, Canada. Association for Computational Linguistics.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023a. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yingxiu Zhao, Bowen Yu, Binyuan Hui, Haiyang Yu, Fei Huang, Yongbin Li, and Nevin L Zhang. 2023b. A preliminary study of the intrinsic relationship between complexity and alignment. *arXiv preprint arXiv:2308.05696*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023a. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Haotian Zhou, Tingkai Liu, Qianli Ma, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023b. Lobass: Gauging learnability in supervised fine-tuning data. *arXiv preprint arXiv:2310.13008*.

Tong Zhou, Yubo Chen, Pengfei Cao, Kang Liu, Jun Zhao, and Shengping Liu. 2023c. Oasis: Data curation and assessment system for pretraining of large language models. *arXiv preprint arXiv:2311.12537*.

A Takeaways

In the pretraining stage of LLMs:

- The coverage of more domains and proper domain mixture ratio are important. Recently, researchers try to automatically find the proper domain mixture weights, which still show room for improvement.

- Large amount of data is widely considered critical, and proper data repetition may also bring positive impacts to model performance.

- Data quality control is necessary usually form an order, namely quality filtering, deduplication and toxicity filtering.. However, over-aggressive quality and toxicity filtering may lead to performance degradation and social biases, which is still under-explored.

- Data diversity and temporal misalignment also have impacts on model performance, which call for future study.

In the supervised fine-tuning stage of LLMs:

- Multitask fine-tuning is widely adopted nowadays. However, conflictions may exist among tasks and hinders the model abilities. Hence, dealing with negative task confliction is also calling for better answers. Ensembling multiple single-task experts instead of training one multitask model also arises as an new trend.

- Quality control are usually achieved through heuristics, human evaluation or LLMs as quality judges. Instruction diversity and complexity are also beneficial and enhanced by several works. The exploration of more diverse and complex instructions is still an open question.

- Works have shown that the SFT of LLM rely more on data quality than data quantity. However, digging deeper into the influence of data quantity, some researchers find that the learning of different tasks may require different amount of data.

- Instead of keep instruction datasets unchanged during fine-tuning, works propose to adjust the datasets dynamically through fine-tuning. Special fine-tuning strategies are also continually shown up to utilize the instruction data more efficiently.

B Other Aspects of Data Management For LLMs

B.1 Social Bias

Besides the marginalization of minority groups caused by data detoxifying mentioned in Section 2.3.3, several works (Kurita et al., 2019; Nangia et al., 2020; Meade et al., 2022; Feng et al., 2023) find that pre-trained LLMs can capture social biases contained in the large amounts of training text. Evaluating on the C4.EN (Raffel et al., 2020) dataset, Dodge et al. (2021) recommend documenting the social biases and representational harms as well as excluded voices and identities in large web text corpora. Using a dataset of U.S. high school newspaper articles, Gururangan et al. (2022) also argue that the quality filters used for GPT-3 (Brown et al., 2020) prefer newspapers published by larger schools located in wealthier, educated, and urban ZIP codes, leading to a language ideology. Feng et al. (2023) conduct a comprehensive case study focusing on the effects of media political biases in the pretraining corpus on the fairness of hate speech detection and misinformation detection w.r.t. partisan leanings and how it is propagated to language models even further to downstream tasks.

As addressed in previous research, there is still a large gap between current prominent LLMs and ideal LLMs without social biases. Many questions are worth exploring, such as how to mitigate the potential biases in pretraining datasets, the existence of bias in the SFT datasets, and whether it is feasible to reduce social bias through SFT.

B.2 Prompt Design

Current instructions are either heuristically designed by human (Wang et al., 2022; Köpf et al., 2023) or synthetically generated by prominent models (Peng et al., 2023; Ding et al., 2023). The choice of prompts might cause significant model performance variation (Gonen et al., 2022; Weber et al., 2023). Early attempts include manual reformulation of prompts into the ones easier to follow for language models (Mishra et al., 2022), and choosing prompts with the lowest perplexity to get the most significant gains in model performance (Gonen et al., 2022). Recently, Liang et al. (2023) develop a format transfer framework *UIT* to transfer instructions from different datasets into unified formats automatically.

Some works focus on studying the impact of prompt phrasing. Khashabi et al. (2022) surprisingly find that the discretized interpretation of continuous prompts is not always consistent with the discrete prompts describing the same task as heuristically expected. Yin et al. (2023b) find that removing the descriptions of task output, especially the label information, might be the only reason for performance degradation. They also propose an automatic task definition compression algorithm to remove almost half or more of the tokens while improving model performance. Kung and Peng (2023) also remove all semantic components in task definitions but the output space information. They achieve comparable model performance using the modified task definitions and delusive examples containing incorrect input-output mappings. Based on their experiment results, they cast doubts on the performance gain of fine-tuned models and state that the model may only learn superficial patterns during instruction tuning.

Besides the choice of phrasing, the generation source of prompts is another factor in prompt design. Gudibande et al. (2023) raise questions on fine-tuning a weaker language model on outputs of a stronger model and find that the imitation model might adapt to mimic the stronger model's style but not its functionality. Similarly, Song et al. (2023) also observe that human-designed data can outperform synthetically generated data from GPT-4 (OpenAI, 2023) to a relatively large extent.

B.3 Hallucinations

Despite their strong power, LLMs are notorious for their hallucinations, i.e. the generation of input-, context- or fact-conflicting contents (Zhang et al., 2023c). Several works in hallucination trace down the occurrence of hallucination to the lack of pertinent knowledge and the internalization of false knowledge from the pretraining corpora (Li et al., 2022a; McKenna et al., 2023; Dziri et al., 2022). To mitigate hallucination, the curation of pretraining corpora is adopted by many LLMs, mainly focusing on the extracting of high-quality data, e.g., GPT-3 (Brown et al., 2020), Llama 2 (Touvron et al., 2023b), and Falcon (Penedo et al., 2023). The manually curated (Zhou et al., 2023a) and automatically selected (Chen et al., 2023c; Cao et al., 2023; Lee et al., 2023b) high-quality instruction data are also experimentally shown to be effective in reducing hallucination during the SFT stage. It can be seen from the previous research that data management in both the pretraining and SFT stages can be a promising solution to hallucination.

C Related Surveys

As LLMs draw more and more attention, a handful of surveys have been published or preprinted addressing different aspects of their development. Related to our work, several of them also include parts of the data preparation process in the pretraining or SFT of LLM. Zhao et al. (2023a) review the development of LLMs and the latest advancements covering a wide range of topics. Yang et al. (2023a) also provide an overview of the LLM evolution and discuss the related techniques from model, data, and downstream tasks. Also concentrating on data, Zha et al. (2023) introduce data-centric AI and its related tasks and methods for general machine learning models instead of LLMs. Zhang et al. (2023b) survey the instruction tuning of LLMs and its related methodologies, data construction, applications, and so on. Wang et al. (2023e) review the technologies aligning LLMs with human expectations including data collection, training methodologies, and model evaluation.

Unlike previous surveys, this survey provides a systematic and detailed overview of data management at both the pretraining and SFT stages of LLMs. We focus on the proper organization of training datasets and discuss recent research addressing the effects of different data management strategies, the evaluation of curated training datasets, and the latest advances in training data management strategies, providing a guiding resource for practitioners aiming to build powerful LLMs through efficient data management.

D Comparison of Data Management Strategies Used by Representative LLMs

We provide two summary tables, Table 1 for pretrained LLMs and Table 2 for SFT LLMs, with better and clearer comparison of the data management strategies used by current representative LLMs.

E Taxonomy

The full taxonomy of research discussed in this survey is illustrated in Figure 3

Pretrained LLMs	Open-sourced	Quantity	Deduplication	Quality Filters	Toxicity Filters	Domian Composition
T5 (Raffel et al., 2020)	√	750GB	N-gram	Heuristic	Heuristic	99% Web, < 1% Wiki
GPT-3 (Brown et al., 2020)		499B tokens	MinHash, LSH	Classifier		82% Web, 16% Books, 3% Wiki
GLaM (Du et al., 2022)		1.6T tokens		Classifier		46% Web, 28% Dialog, 20% Books, 6% Wiki
LaMDA (Thoppilan et al., 2022)		1.56T words				50% Dialog, 25% Web, 12.5% Wiki, 12.5% Code
Chinchilla (Hoffmann et al., 2022a)		1.4T tokens	N-gram, Doc-level	Heuristic	Heuristic	65% Web, 30% Books, 4% Code, 1% Wiki
AlphaCode (Li et al., 2022b)		715.1GB	Doc-level	Heuristic		100% Code
GLM (Zeng et al., 2022)	√	400B tokens				50% Pile, 50% Chinese Web data
BLOOM (Workshop et al., 2022)	√	1.61TB text	SimHash, Substring clustering	Heuristic	Heuristic	60% Web, 10% Books, 10% Code, 10% Academic, 5% Dialog, 5% Wiki
PaLM (Anil et al., 2023)		780B tokens	Levenshtein distance	Heuristic, Classifier	Classifier	50% Dialog, 28% Web, 13% Books, 5% Code, 4% Wiki
LLaMA (Touvron et al., 2023a)	√	1.4T tokens	Line-level, Book-level	Heuristic, Classifier	Classifier	82% Web, 4.5% Code, 4.5% Wiki, 4.5% Books, 2.5% Academic, 2% Dialog
Mistral (Jiang et al., 2023a)	√	-	-	-	-	-
phi-1/1.5 (Gunasekar et al., 2023) (Li et al., 2023d)	√	7B tokens		Classifier		99% Academic, <1% Code
phi-2 (Javaheripi and Bubeck, 2023)	√	1.4B tokens		Classifier		
GPT-4 (OpenAI, 2023)		-	-	-	-	-
LLaMA 2 (Touvron et al., 2023b)	√	2.0T tokens			Heuristic	
QWen (Bai et al., 2023)	√	3T tokens	Exact Match, MinHash, LHS	Heuristic, Classifier	Classifier	Web, Books, Codes, Academic
Deepseek LLM (Bi et al., 2024)	√	-	-	-	-	-

Table 1: The data management strategies used by representative pretrained models. The blank units mean no specific design of corresponding strategies according to the original papers. The '-' means the data management process is not released. Part of the data is adopted from (Longpre et al., 2023b)

SFT LLMs	Dataset	Quantity	Quality Control	Diversity Control	Complexity Enhancing	No. of Tasks	Task Balancing
Tk-Instruct (Wang et al., 2022)	NIv2	5M	Heuristic Human			1616	Limited instances per task
Flan-T5 (Longpre et al., 2023a)	Flan 2022	15M		Input Inversion		1836	Experiments intuitions
OPT-IML (Iyer et al., 2022)	OPT-IML Bench	18M				2000	Experiments
Alpaca (Taori et al., 2023)	Alpaca	52K	Heuristic	ROUGE-L similarity		80	
Vicuna (Chiang et al., 2023)	ShareGPT	70K	Heuristic				
LIMA (Zhou et al., 2023a)	LIMA	1K	Heuristic Human	Heuristic, Human			
Dolly (Conover et al., 2023)	dolly-15k	15K	Human				
Orca (Mukherjee et al., 2023)	sampled Flan 2022	5M			Chat-GPT/ GPT-4 augmentation		
WizardLM (Xu et al., 2023a) WizardCoder (Luo et al., 2023b) WizardMath (Luo et al., 2023a)	WizardLM WizardCoder WizardMath	250K		Evol-Instruct	Evol-Instruct		
AlpaGasus (Chen et al., 2023c)	AlpaGasus	9K	Chat-GPT grading				
Platypus (Lee et al., 2023b)	Open-Platypus	25K	Dedup, Heuristic				
OpenChat (Wang et al., 2023a)	ShareGPT	6K	C-RLFT				
MAmmoTH (Yue et al., 2023)	MathInstruct	260K				7 math fields	Combining CoT and PoT

Table 2: The data management strategies used by representative supervised finetuned models. The blank units mean no specific design of corresponding strategies according to the original papers. "NIv2" is the abbreviation for "Super-NaturalInstructions". "Dedup" is the abbreviation for "Deduplication".

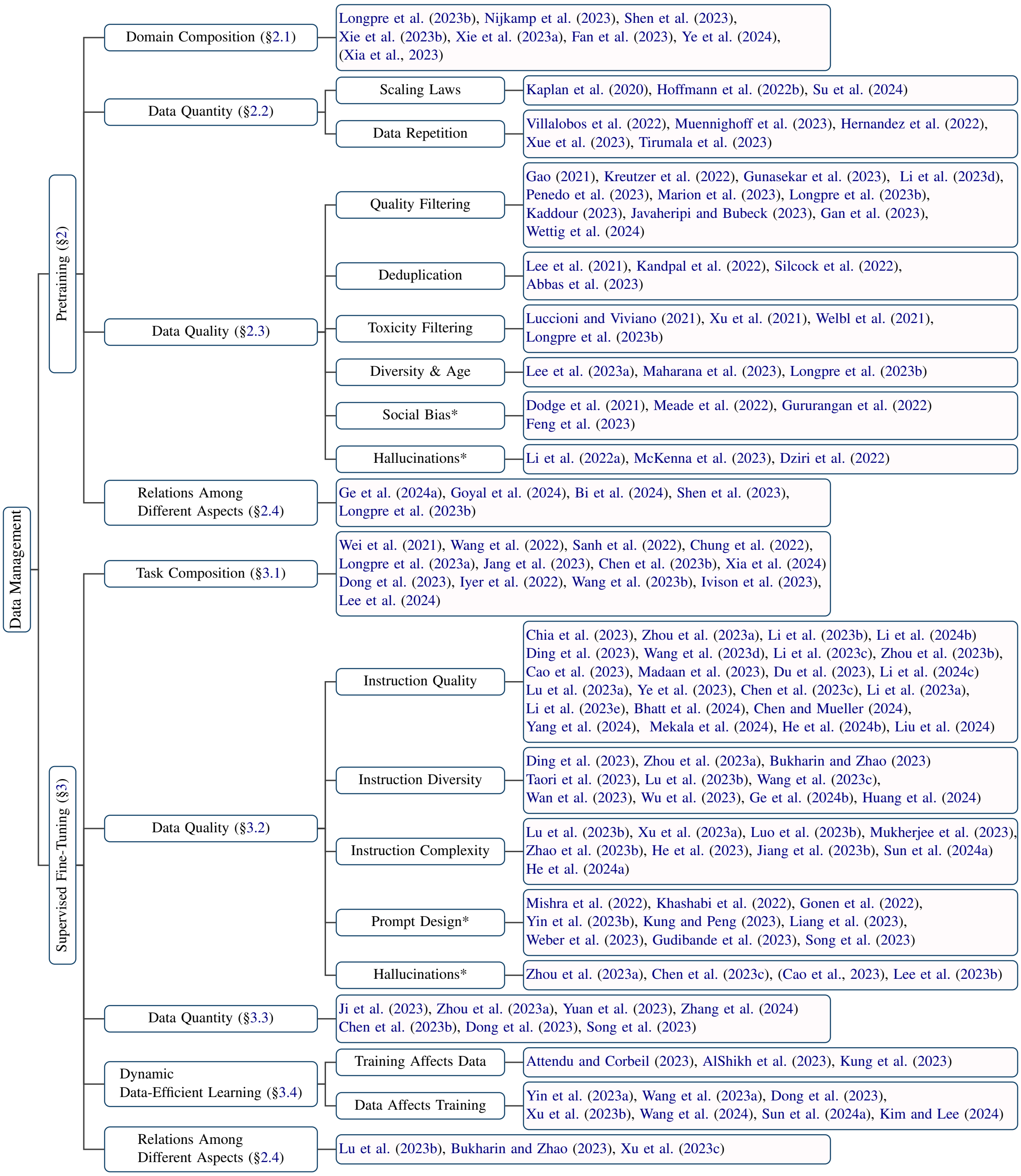

Figure 3: Taxonomy of research in data management for pretraining and supervised fine-tuning of Large Language Models (LLM).