Data Management For 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 current research community still falls short in providing a systematic analysis of the effects of data management strategy selection, methodologies for evaluating curated datasets, and the ongoing pursuit of improved strategies. Consequently, the exploration of data management has attracted more and more attention among the research community. This survey provides a comprehensive overview of current research in data management within both the pretraining and supervised fine-tuning stages of LLMs, covering various noteworthy aspects of data management strategy design: data quantity, data quality, domain/task composition, etc. Looking toward 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., 2022), 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.

Organizing a well-suited training dataset using collected data, which we define as **data management**, is vitally important and challenging in both the pretraining and SFT stages of LLMs. 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 the techniques used in the construction of their pretraining dataset (Brown et al., 2020; Workshop et al., 2022; Touvron et al., 2023a), 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; Köpf et al., 2023; Wang et al., 2023c; Taori et al., 2023; Si 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 these challenges, it is necessary to conduct a systematic analysis of LLM data management, including the effect of data management strategy selection, the evaluation of curated datasets, and the latest pursuit of improved strategies. Therefore, this survey aims to provide a comprehensive overview of current research in LLM data management, as shown in Figure 1. Section 2 focuses on LLM pretraining data management, including the research on data quantity, data quality, domain composition, and data management systems. Sec-

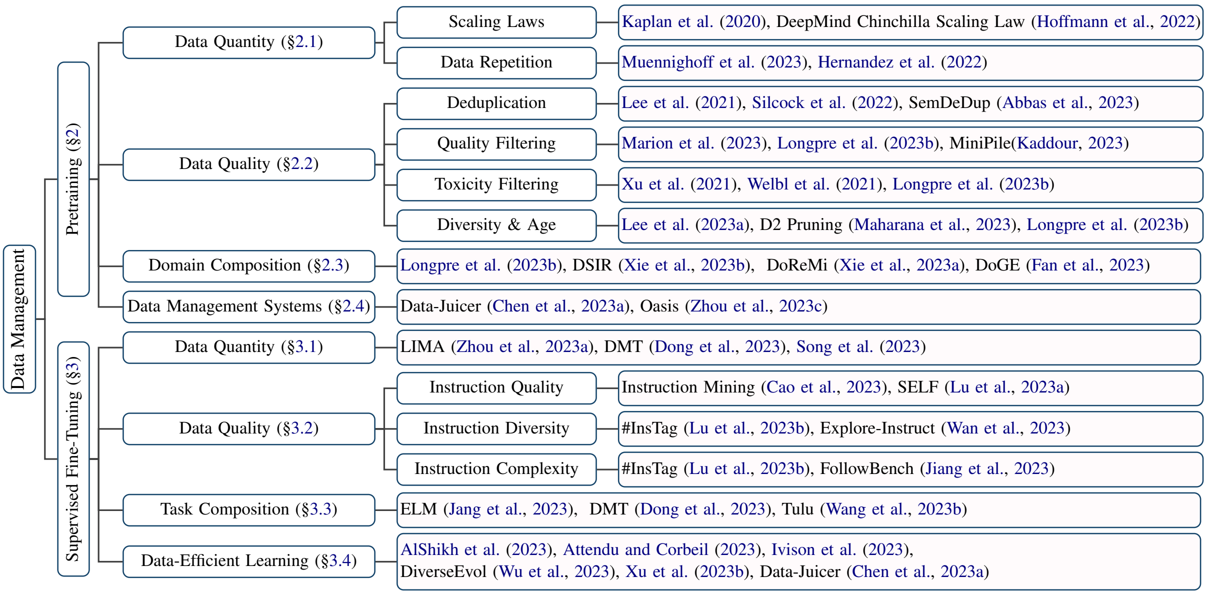
Data Management
Pretraining (§2)
Data Quantity (§2.1)
Scaling Laws
Kaplan et al. (2020), DeepMind Chinchilla Scaling Law (Hoffmann et al., 2022)
Data Repetition
Muennighoff et al. (2023), Hernandez et al. (2022)
Data Quality (§2.2)
Deduplication
Lee et al. (2021), Silcock et al. (2022), SemDeDup (Abbas et al., 2023)
Quality Filtering
Marion et al. (2023), Longpre et al. (2023b), MiniPile(Kaddour, 2023)
Toxicity Filtering
Xu et al. (2021), Welbl et al. (2021), Longpre et al. (2023b)
Diversity & Age
Lee et al. (2023a), D2 Pruning (Maharana et al., 2023), Longpre et al. (2023b)
Domain Composition (§2.3)
Longpre et al. (2023b), DSIR (Xie et al., 2023b), DoReMi (Xie et al., 2023a), DoGE (Fan et al., 2023)
Data Management Systems (§2.4)
Data-Juicer (Chen et al., 2023a), Oasis (Zhou et al., 2023c)
Supervised Fine-Tuning (§3)
Data Quantity (§3.1)
LIMA (Zhou et al., 2023a), DMT (Dong et al., 2023), Song et al. (2023)
Data Quality (§3.2)
Instruction Quality
Instruction Mining (Cao et al., 2023), SELF (Lu et al., 2023a)
Instruction Diversity
#InsTag (Lu et al., 2023b), Explore-Instruct (Wan et al., 2023)
Instruction Complexity
#InsTag (Lu et al., 2023b), FollowBench (Jiang et al., 2023)
Task Composition (§3.3)
ELM (Jang et al., 2023), DMT (Dong et al., 2023), Tulu (Wang et al., 2023b)
Data-Efficient Learning (§3.4)
AlShikh et al. (2023), Attendu and Corbeil (2023), Ivison et al. (2023),
DiverseEvol (Wu et al., 2023), Xu et al. (2023b), Data-Juicer (Chen et al., 2023a)

Figure 1: Taxonomy of research in data management for pretraining and supervised fine-tuning of Large Language Models. For space limitation, only representative works are listed here. Please see the full taxonomy in Appendix C.

tion 3 discusses the data quantity, data quality, task composition, and data-efficient learning techniques in the SFT stage of LLMs. Looking into the future, Section 4 presents the existing challenges and promising future directions in training data management for LLMs. Through this survey, we are devoted to offering a guiding resource to practitioners attempting to build powerful LLMs with efficient data management practices.

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). Understanding the effects of these data management strategies is also crucial for building strong LLMs. In this section, we will discuss current works trying to disclose the working scheme of data management in the pretraining stage of LLMs.

2.1 Data Quantity

The amount of data required for efficient pretraining of LLMs is an ongoing research topic in NLP communities. First, scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) are proposed to depict the relationship between model size and training dataset size - with model size continuously increasing, the demand for more training data will also increase consistently. Then, the exhaustion of text data draws researchers' attention to data repetition in LLMs' pretraining (Muennighoff et al., 2023; Xue et al., 2023; Tirumala et al., 2023).

2.1.1 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) use Transformers and cross-entropy loss to study the empirical scaling laws for language model performance. They find that the model performance 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 N and training dataset size D as:

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

where L is the test loss, D is the number of training tokens, N is the number of 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 performance improves 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 $N_{opt} \sim$

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

$C^{0.73}$, showing that the model size should increase faster than the training dataset size.

Following the power-law relationship proposed by Kaplan et al. (2020), Hoffmann et al. (2022) conduct experiments on much larger language models and arrive at a new scaling law:

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

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.

2.1.2 Data Repetition

While Kaplan et al. (2020) and Hoffmann et al. (2022) focus on scaling law with unique data trained only for one epoch, Hernandez et al. (2022) address the issue about text overlap in the training dataset and study the scaling law with a small fraction of repeated data. They observe a strong *double descent phenomenon* (Nakkiran et al., 2021) caused by repeated data, where a peak of test loss appears in the middle range of repetition frequency, i.e., the number of epochs trained on repeated data. They also show that repeated data can cause a divergence from power-scaling law (Kaplan et al., 2020) on model sizes larger than 100M parameters.

According to the scaling law, more training data is required as the model size grows, raising concerns about the exhaustion of high-quality training data (Villalobos et al., 2022; Hoffmann et al., 2022). Addressing these concerns, several works study the consequence of repeatedly pretraining on the whole datasets for multiple epochs. Muennighoff et al. (2023) find that with constrained data and fixed computing budgets, repeatedly training on the whole dataset up to 4 epochs only causes trivial harm to test loss compared to training on unique new data. They also propose a scaling law on repeated training depicting the diminishing of returns with more repetition and larger model sizes. Xue et al. (2023) also observe a multi-epoch degradation in model performance and find that dataset size, model parameters, and training objectives are the key factors to this phenomenon. They further find that commonly used regularization techniques are not helpful in alleviating multi-epoch degradation, except for dropout. 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, whilst repeatedly training on randomly selected data cannot, suggesting a feasible way of repeating on intelligently selected data.

2.2 Data Quality

High-quality data is crucial in machine learning tasks (Jain et al., 2020; Gupta et al., 2021). In the pretraining of LLMs, quality assurance techniques are adopted and usually form a data management pipeline (Rae et al., 2021; Nguyen et al., 2023; Tirumala et al., 2023), including deduplication, quality filtering, and toxicity filtering. Other aspects like data diversity and data age are also studied.

2.2.1 Deduplication

Deduplication is widely used in many LLMs' 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) use N-gram similarity with MinHash (Broder, 1997) to detect duplications in training datasets and find that deduplication is beneficial in memorization 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). Silcock et al. (2022) compare it with two model-based neural approaches and conclude that neural approaches can significantly outperform traditional N-gram-and-hashing methods. Abbas et al. (2023) propose *SemDeDup* to remove semantic duplicates that lie closely in the pre-trained model's embedding space and apply clustering to reduce the searching computation.

2.2.2 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.,

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

2022), a trained classifier (Brown et al., 2020; Gao et al., 2020; Du et al., 2022; Touvron et al., 2023a), or threshold filtering using criteria like perplexity (Wenzek et al., 2020; Muennighoff et al., 2023). Kaddour (2023) constructs a subset of the Pile (Gao et al., 2020) called *MiniPile* by filtering out low-quality embedding clusters.

Quality filtering is usually proven to be beneficial in model performance improvement (Longpre et al., 2023b), despite the reduction of training data quantity and variety. Several carefully filtered high-quality datasets are proposed to train lightweight language models and achieve outstanding performances (Gunasekar et al., 2023; Li et al., 2023b; 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 examines three data quality estimators, i.e., perplexity, Error L2-Norm (EL2N), and memorization factor. Surprisingly, they find that pruning datasets based on perplexity and retaining the middle proportion of data performs better than more complicated techniques like memorization. However, no combination of pruning strategies seems to achieve consistently high performance.

2.2.3 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.

2.2.4 Diversity & Age

Some works focus on other aspects of data management in the pretraining stage of LLMs. For example, 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. They represent a dataset as an undirected graph with samples as nodes, difficulty scores as node properties, and distances in the embedding space as edge weights. Then, a forward and reverse message passing strategy is adopted to select a subgraph enveloping both diverse and difficult data samples.

Longpre et al. (2023b) explore the age of the evaluation dataset and conclude that the temporal shift between evaluation and pretraining data will lead to inaccurate performance estimation and the temporal misalignment might not be overcome by fine-tuning, especially for larger models.

2.3 Domain Composition

Public available pretraining datasets (Gao et al., 2020) usually contain mixtures of data collected from multiple sources and domains. Many prominent models (Thoppilan et al., 2022) are also trained on a mixture of data from different domains.

Efforts are made to explore the impact of domain mixtures on the pre-trained models' performance. Longpre et al. (2023b) experimentally conclude that domains with high quality (Books) and high diversity (Web) in the Pile (Gao et al., 2020) are broadly helpful. They also show that including as many data domains as possible is beneficial. Shen et al. (2023) arrive at the same point and emphasize the importance of global deduplication to remove overlaps among different domains. Longpre et al. (2023b) and Shen et al. (2023) all agree that specific mixtures may excel in evaluation benchmarks for targeted tasks, but the former 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.

Several methods are also proposed to find the proper domain composition weights. *DSIR* (Xie et al., 2023b) formulates the problem as a distribution matching problem between a large raw unlabeled dataset and some unlabeled target samples, which is solved using classic importance resampling approach (Rubin, 1988). Without knowledge of downstream tasks or target distributions, *DoReMi* (Xie et al., 2023a) trains a small proxy model using Group Domain Robust Optimization (Group DRO) (Oren et al., 2019; Sagawa* et al., 2020) to generate domain weights. Improved from *DoReMi* (Xie et al., 2023a), Fan et al. (2023) propose *DoGE* which reweights training domains to minimize the average validation loss across all training domains or on a specific unseen domain. A gradient-based generalization estimation function is adopted to measure the contribution of each domain to other domains. Then, higher weights are assigned to domains with higher contributions.

2.4 Data Management Systems

Addressing the difficulty in pretraining data management, integrated data management systems are beneficial for LLM practitioners with different demands. Chen et al. (2023a) provide a data processing system *Data-Juicer* featuring the generation of diverse data recipes. They provide over 50 versatile data management operators and dedicated tools targeting users with different purposes. A timely feedback evaluation loop is also supported. Zhou et al. (2023c) also propose a pretraining data curation and assessment system *Oasis*, which can perform interactive rule filtering, debiased neural quality filtering, adaptive document deduplication, and holistic data assessment.

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). Many efforts have been made to construct instruction data using human crowd-sourcing (Wang et al., 2022; Köpf et al., 2023), self-instruct (Wang et al., 2023c; Taori et al., 2023) or adaptation of existing datasets (Si et al., 2023; Anand et al., 2023). 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.

3.1 Data Quantity

The explorations of the relationship between scaling instruction data quantity and fine-tuned model performance diverge in two directions. One branch of research focuses on scaling down the instruction data quantity to improve training efficiency (Zhou et al., 2023a; Chen et al., 2023b). For example, LIMA (Zhou et al., 2023a) carefully curates 1,000 high-quality samples and experimentally justifies their hypothesis that only limited instruction tuning data is needed to expose the knowledge and capabilities that the LLM has already acquired during pretraining. Chen et al. (2023b) observe that maybe a single instruction is sufficient for single task-specific LLM fine-tuning, and 16K samples with 1.9M tokens may be sufficient to train a model specialized in the natural language inference (NLI) task. Another branch of research argues that scaling up the instruction data quantity is crucial for success (Wei et al., 2021; Sanh et al., 2022).

Addressing this conflict, several works attempt to analyze the scaling patterns for different tasks or different model abilities. Ji et al. (2023) conduct an empirical study on 12 major real-world online user cases and show that scaling up the instruction data leads to continuous improvement in tasks such as extraction, classification, closed QA, and summarization while leading to little improvement in tasks such as math, code, and chain-of-thought. Disagree with Ji et al. (2023), 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 pre-trained models improve less with more instruction tuning data. Song et al. (2023) conduct experiments covering ten distinct in-domain abilities and three out-of-domain abilities, showing that the developments of most abilities are consistent with data scaling. Still, each ability develops at different paces during instruction tuning, while some abilities show completely different patterns.

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 data, there is usually a filtering step to select high-quality instructions generated by models. Wang et al. (2023d) use perplexity as the criterion to select the most appropriate instructions from the pool of candidate instructions generated by open-source models. Cao et al. (2023) propose an automatic data selector *Instruction Mining* to evaluate instruction data quality without human experts' interventions. They first hypothesize that the inference loss of a fine-tuned model on an evaluation set can serve as a proxy for data quality filtering objectives. Then, they use a set of heuristic-based and model-based natural language indicators to predict the inference loss without actually fine-tuning LLMs.

Instead of using indicators to filter low-quality instructions, several works (Li et al., 2023a; Lu et al., 2023a; Ye et al., 2023; Madaan et al., 2023) leverage the power of fine-tuned LLM itself to evaluate the quality of instructions. Li et al. (2023a) assign quality scores to augmented instructions using the language model and iteratively improve model prediction. Similarly, *SELF* (Lu et al., 2023a) and *Self-Refine* (Madaan et al., 2023) prompts LLM to provide self-feedback on their own responses in the iterative model evolution processes. Strong LLMs like ChatGPT are also adopted as quality judges during instruction collection (Ye et al., 2023).

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 (Zhou et al., 2023a; Ding et al., 2023; Taori et al., 2023). *Self-Instruct* (Wang et al., 2023c) adopts ROUGE-L similarity to filter out the newly generated instructions that are too similar to the existing ones. To better evaluate the instruction diversity of SFT datasets, *#InsTag* (Lu et al., 2023b) is proposed as an open-set fine-grained tagger using ChatGPT [3]. Specifically, it first prompts ChatGPT to provide tags for given queries in an open setting, then performs a normalization procedure to deal with the noise in the raw tagging. With the generated tags, they quantify instruction diversity as the unique tag coverage rate in the overall tag set. Popular open-set SFT datasets are analyzed using *#InsTag*, showing that larger dataset sizes tend to be more diverse and induce higher performance.

Diversity can be challenging in domain-specific tasks due to data constraints. Thus, Wan et al. (2023) propose an approach called *Explore-Instruct* to enlarge the data coverage through active exploration using LLMs. *Explore-Instruct* starts from representative domain user cases and searches the variations and possibilities by looking ahead into potential fine-grained sub-tasks and backtracking alternative branches in the search space.

3.2.3 Instruction Complexity

The complexity of instructions also attracts researchers' attention, especially in developing LLMs with complex instruction-following and reasoning abilities (Xu et al., 2023a; Luo et al., 2023; Mukherjee et al., 2023). 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.

To explore instruction complexity, Zhao et al. (2023b) propose *Tree-Instruct* to enhance the complexity of instruction data controllably. It treats the instruction as a semantic tree and constructs new complex instructions by adding nodes to the tree. Through experiments, they find that increased complexity can lead to continuing performance improvement. What's more, the improvement does not come from the increased number of tokens, as a few complex instructions still outperform diverse but simple instructions under the same token budget. They also show that curriculum instruction tuning ranging from easy to difficult might not be as helpful as expected, indicating the necessity of enhancing complexity. *Evol-Instruct* (Xu et al., 2023a; Luo et al., 2023) rewrites instructions step by step with operations such as increasing reasoning, adding constraints, in-breadth evolving, deepening, and complicating input with code and table. Similarly, Jiang et al. (2023) incrementally augment instructions with constraints on content,

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

situation, style, format, and example, proposing *FollowBench* to evaluate LLMs' constraint following ability.

3.3 Task Composition

Since LLMs have shown surprisingly emergent abilities in handling various NLP tasks, multitask fine-tuning appears 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).

Besides the scaling of the number of tasks, the mixture ratio of data targeting different tasks is also found to be critical (Iyer et al., 2022; Longpre et al., 2023a). Dong et al. (2023) focus on task composition among mathematical reasoning, code generation, and general human-aligning abilities. Compared with individual source data, They find that model abilities are improved when the mixed data amount is small but decreased otherwise. The results indicate that larger amounts of mixed data lead to conflicts among learning different abilities. They further vary the ratio of general and specialized data and conclude that the impact of data ratio might lie in the similarity degree of data format and data distribution among different SFT tasks.

Divergent from compositing multiple tasks together, 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). Jang et al. (2023) state that training expert LLMs to form an expert library is beneficial in negative task transfer avoidance, continually learning new tasks without catastrophic forgetting, and compositional abilities improvement. Wang et al. (2023b) conduct analysis on factual knowledge, reasoning, multilinguality, coding, and open-ended instruction following abilities of models trained with 12 instruction datasets and experimentally show that different instruction datasets may correspond to different specific abilities. What's more, winning across all evaluations using a single dataset or combination seems to be challenging.

3.4 Date-Efficient Learning

Addressing different aspects of instruction data management, a handful of works propose to fine-tune LLM more data-efficiently with subset selection or specially designed fine-tuning strategies.

Data Quantity AlShikh et al. (2023) introduce *Instruction Following Score (IFS)* to measure LLMs' instruction-following ability and serve as an early-stopping criterion. It is defined as the percentage of responses predicted as "answer-like" by a binary classifier. Based on observations of different scaling patterns for different abilities, Dong et al. (2023) propose *Dual-stage Mixed Fine-tuning (DMT)* strategy to learn specialized abilities and general abilities sequentially while keeping a small proportion of specialized data to prevent forgetting.

Data Quality Several works focus on selecting a subset of instruction data with the highest quality. Cao et al. (2023) adopt BlendSearch (Wang et al., 2020) to automatically select the best subset. AlpaGasus (Chen et al., 2023c) uses strong LLMs as auto-graders and selects data with scores above a threshold in the Alpaca dataset (Taori et al., 2023). 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). Without discarding data samples, OpenChat (Wang et al., 2023a) considers the general SFT data as a mixture of a small amount of expert data and a large amount of sub-optimal data without any preference labels. Then, *Conditioned-RLFT* strategy is proposed, which treats different data sources as coarse-grained reward labels and optimizes the LLM as a class-conditioned policy. To enhance instruction diversity in the chosen subsets, *DiverseEvol* (Wu et al., 2023) uses an iterative data sampling technique that selects new data points with maximized distances from any existing ones in model embedding space.

Task Composition Given a small amount of target task data, Ivison et al. (2023) select the relevant multitask subsets for fine-tuning according to the similarity between the pre-trained model's representations of the target and mixed task data. Similarly, *Dynosaur* (Yin et al., 2023a) treats task selection based on data representations as a replay strategy in continual learning scenarios to mitigate catastrophic forgetting issues and improve generalization to unseen tasks. Yue et al. (2023) build math generalist models *MAmmoTH* through instruction tuning on a unique hybrid of chain-of-thought and program-of-thought rationales in math.

Others *LoBaSS* (Zhou et al., 2023b) introduces *learnability* as a new dimension of SFT data selection that data can be learned more effectively by the model are preferable and data lacking informative content or excessively demanding for the model should be avoided. The proposed *learnability* is further measured as the loss difference between fine-tuned and pre-trained models. Xu et al. (2023b) propose a contrastive post-training technique treating data acquired from LLMs with different levels of abilities as contrastive pairs. They also use a data curriculum scheme where the model learns progressively from the "easier" to the "harder " part. *Data-Juicer* (Chen et al., 2023a) also implements pipelines and operators for LLM fine-tuning.

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.

Comprehensive and Fine-grained Understanding As discussed in previous sections, many efforts have been made to understand the impacts of data management on different training stages addressing different aspects. While current studies contribute valuable pieces to the puzzle, a comprehensive understanding of the entire picture is still lacking. Moreover, explorations using different datasets and models on different tasks may lead to contradictory conclusions, e.g., the trade-off between quality and toxicity filtering (Longpre et al., 2023b), fine-tuning with a few high-quality data (Zhou et al., 2023a) v.s. data scaling (Wei et al., 2021), task composition (Wang et al., 2022) v.s. expert models (Jang et al., 2023), etc. Hence, more fine-grained understanding is required to solve these conflicts.

General Data Management Framework Although *Data-Juicer* (Chen et al., 2023a) and *Oasis* (Zhou et al., 2023c) propose data management systems to compose various data recipes in either the pretraining or SFT stage of LLM, practitioners still need to spend efforts on organizing suitable datasets. A general data management framework suitable for a broad range of applications to reduce data management costs is an urgent and worthy future direction in developing and promoting LLMs.

Data Curriculum Besides choosing better training data, data curriculum addressing the arrangement of data learning orders is also an important part of data management, e.g., learning from general abilities to target abilities or from easier tasks to harder tasks. There are a few works focusing on data curriculum in the training of LLMs (Xu et al., 2023b; Dong et al., 2023; Yin et al., 2023a). Although effective in practice, there is still a lack of analysis of data curriculum strategies.

Conflict Data Separation In the collection of training data, conflicts among the responses to the same query may exist. For example, given the same query, LLMs playing different roles may generate different responses. Mixing these samples together could lead to negative impacts on model performance because of the response conflicts. Thus, how to separate and effectively learn from these data samples is an interesting topic in the future.

Multimodal Data Management Current research in data management mostly focuses on natural language processing. With the application of LLMs extending to multimodalities like vision, audio, etc., the construction of multimodality datasets becomes more and more important. The proposed multi-modal LLMs usually construct their own instruction-tuning datasets collected from benchmark adaptation (Zhang et al., 2023a; Gao et al., 2023) or self-instruction (Pi et al., 2023; Yang et al., 2023b). The hybrid composition of language-only and multimodal data is also adopted in some works (Dai et al., 2023; Zhao et al., 2023c). It is interesting to see the impacts of multimodal data management on the performance of fine-tuned multimodal LLMs, e.g., the data scaling law in multimodal instruction fine-tuning, the quality-control techniques in multimodal dataset construction, and task balancing in multitask multimodal training.

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 into *data quantity*, *data quality*, and *domain/task composition* for each stage, as well as *data management systems* in the pretraining stage and *data-efficient learning* in the SFT 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.

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.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

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.

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*.

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*.

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*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

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.

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.

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.

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*.

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*.

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, 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*.

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. 2022. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030.

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*.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2023. 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.

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.

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*.

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.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Cheng-Jie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. 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. 2023a. 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. 2023b. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

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*.

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*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. 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*.

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.

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*.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2021. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.

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.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4227–4237.

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*.

Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. 2023. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*.

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.

D.B. Rubin. 1988. Using the sir algorithm to simulate posterior distributions. In *Bayesian statistics 3. Proceedings of the third Valencia international meeting, 1-5 June 1987*, pages 395–402. Clarendon Press.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

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*.

Qingyi Si, Tong Wang, Zheng Lin, Xu Zhang, Yanan Cao, and Weiping Wang. 2023. An empirical study of instruction-tuning large language models in chinese.

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*.

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*.

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. Supernaturalinstructions: 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.

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*.

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*.

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*.

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*.

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*.

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*.

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*.

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023c. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*.

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 Other Aspects of Data Management For LLMs

A.1 Social Bias

Besides the marginalization of minority groups caused by data detoxifying mentioned in Section 2.2.3, 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.

A.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.

A.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., 2022; 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.

B 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.

C Taxonomy

The full taxonomy of research discussed in this survey is illustrated in Figure 2

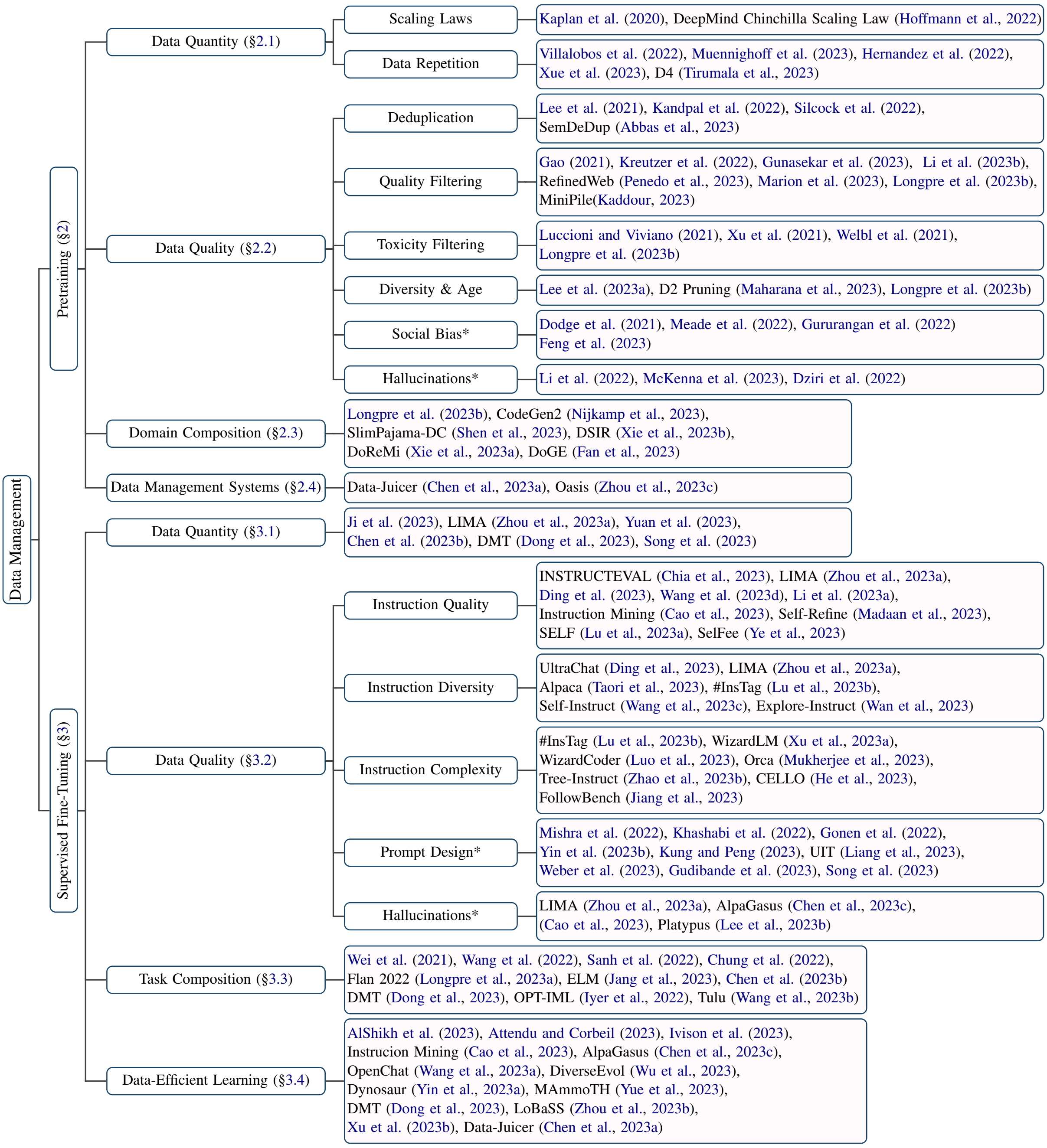
Data Management
Pretraining (§2)
Data Quantity (§2.1)
Scaling Laws
Kaplan et al. (2020), DeepMind Chinchilla Scaling Law (Hoffmann et al., 2022)
Data Repetition
Villalobos et al. (2022), Muennighoff et al. (2023), Hernandez et al. (2022), Xue et al. (2023), D4 (Tirumala et al., 2023)
Data Quality (§2.2)
Deduplication
Lee et al. (2021), Kandpal et al. (2022), Silcock et al. (2022), SemDeDup (Abbas et al., 2023)
Quality Filtering
Gao (2021), Kreutzer et al. (2022), Gunasekar et al. (2023), Li et al. (2023b), RefinedWeb (Penedo et al., 2023), Marion et al. (2023), Longpre et al. (2023b), MiniPile(Kaddour, 2023)
Toxicity Filtering
Luccioni and Viviano (2021), Xu et al. (2021), Welbl et al. (2021), Longpre et al. (2023b)
Diversity & Age
Lee et al. (2023a), D2 Pruning (Maharana et al., 2023), Longpre et al. (2023b)
Social Bias*
Dodge et al. (2021), Meade et al. (2022), Gururangan et al. (2022) Feng et al. (2023)
Hallucinations*
Li et al. (2022), McKenna et al. (2023), Dziri et al. (2022)
Domain Composition (§2.3)
Longpre et al. (2023b), CodeGen2 (Nijkamp et al., 2023), SlimPajama-DC (Shen et al., 2023), DSIR (Xie et al., 2023b), DoReMi (Xie et al., 2023a), DoGE (Fan et al., 2023)
Data Management Systems (§2.4)
Data-Juicer (Chen et al., 2023a), Oasis (Zhou et al., 2023c)
Supervised Fine-Tuning (§3)
Data Quantity (§3.1)
Ji et al. (2023), LIMA (Zhou et al., 2023a), Yuan et al. (2023), Chen et al. (2023b), DMT (Dong et al., 2023), Song et al. (2023)
Data Quality (§3.2)
Instruction Quality
INSTRUCTEVAL (Chia et al., 2023), LIMA (Zhou et al., 2023a), Ding et al. (2023), Wang et al. (2023d), Li et al. (2023a), Instruction Mining (Cao et al., 2023), Self-Refine (Madaan et al., 2023), SELF (Lu et al., 2023a), SelFee (Ye et al., 2023)
Instruction Diversity
UltraChat (Ding et al., 2023), LIMA (Zhou et al., 2023a), Alpaca (Taori et al., 2023), #InsTag (Lu et al., 2023b), Self-Instruct (Wang et al., 2023c), Explore-Instruct (Wan et al., 2023)
Instruction Complexity
#InsTag (Lu et al., 2023b), WizardLM (Xu et al., 2023a), WizardCoder (Luo et al., 2023), Orca (Mukherjee et al., 2023), Tree-Instruct (Zhao et al., 2023b), CELLO (He et al., 2023), FollowBench (Jiang et al., 2023)
Prompt Design*
Mishra et al. (2022), Khashabi et al. (2022), Gonen et al. (2022), Yin et al. (2023b), Kung and Peng (2023), UIT (Liang et al., 2023), Weber et al. (2023), Gudibande et al. (2023), Song et al. (2023)
Hallucinations*
LIMA (Zhou et al., 2023a), AlpaGasus (Chen et al., 2023c), (Cao et al., 2023), Platypus (Lee et al., 2023b)
Task Composition (§3.3)
Wei et al. (2021), Wang et al. (2022), Sanh et al. (2022), Chung et al. (2022), Flan 2022 (Longpre et al., 2023a), ELM (Jang et al., 2023), Chen et al. (2023b) DMT (Dong et al., 2023), OPT-IML (Iyer et al., 2022), Tulu (Wang et al., 2023b)
Data-Efficient Learning (§3.4)
AlShikh et al. (2023), Attendu and Corbeil (2023), Ivison et al. (2023), Instrucion Mining (Cao et al., 2023), AlpaGasus (Chen et al., 2023c), OpenChat (Wang et al., 2023a), DiverseEvol (Wu et al., 2023), Dynosaur (Yin et al., 2023a), MAmmoTH (Yue et al., 2023), DMT (Dong et al., 2023), LoBaSS (Zhou et al., 2023b), Xu et al. (2023b), Data-Juicer (Chen et al., 2023a)

Figure 2: Taxonomy of research in data management for pretraining and supervised fine-tuning of Large Language Models (LLM).