Data-Efficient Large Language Model Training: A Survey

XINYANG LIU, Tianjin University, China
QIANG HU*, Tianjin University, China
YUJIE MA, Tianjin University, China
ZHENHENG TANG, Tianjin University, China
JIONGCHI YU, Singapore Management University, Singapore
TIANLIN LI, Beihang University, China
YAO ZHANG, Tianjin University, China
JUNJIE WANG, Tianjin University, China
HAO LIU, Tianjin University, China
YONGQIANG LYU, Tianjin University, China
YVES LE TRAON, University of Luxembourg, Luxembourg

Constructing large language models (LLMs) is labor-intensive and computationally unfriendly due to the requirement for large-scale and high-quality datasets. This paper presents a comprehensive survey of building LLMs with limited data to tackle the above challenges. It covers 116 papers, where 18 works focus on the pre-training process, and 98 works lie in the fine-tuning process. This survey: (i) unifies the problems and terminologies associated with data-efficient LLM training, (ii) systematically analyzes techniques proposed for identifying the most important data samples for LLM building, and (iii) highlights the pitfalls and research opportunities in this domain.

CCS Concepts: • **Computing methodologies** → *Natural language processing*; **Cost-sensitive learning**.

Additional Key Words and Phrases: data-efficient, LLMs, training, fine-tuning

ACM Reference Format:
Xinyang Liu, Qiang Hu*, Yujie Ma, Zhenheng Tang, Jiongchi Yu, Tianlin Li, Yao Zhang, Junjie Wang, Hao Liu, Yongqiang Lyu, and Yves Le Traon. 2026. Data-Efficient Large Language Model Training: A Survey. *J. ACM* 37, 4, Article 111 (January 2026), 35 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large Language Models (LLMs) have emerged as a promising avenue for artificial general intelligence (AGI), demonstrating remarkable capabilities across diverse application domains, including information retrieval [141], automated programming [70], and video generation [137]. Given this

* Corresponding author: Qiang Hu.

Authors' addresses: Xinyang Liu, Tianjin University, China, 3021244394@tju.edu.cn; Qiang Hu*, Tianjin University, China, qianghu@tju.edu.cn; Yujie Ma, Tianjin University, China, yujiema0911@tju.edu.cn; Zhenheng Tang, Tianjin University, China, m13596799901@163.com; Jiongchi Yu, Singapore Management University, Singapore, jcyu.2022@phdcs.smu.edu.sg; Tianlin Li, Beihang University, China, tianlin001@buaa.edu.cn; Yao Zhang, Tianjin University, China, zzyy@tju.edu.cn; Junjie Wang, Tianjin University, China, junjie.wang@tju.edu.cn; Hao Liu, Tianjin University, China, nanjierenxjtu@163.com; Yongqiang Lyu, Tianjin University, China, lyuyq@tju.edu.cn; Yves Le Traon, University of Luxembourg, Luxembourg, yves.letraon@uni.lu.

0004-5411/2026/1-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

success, both academia and industry have joined the race to develop more representative and powerful LLMs.

However, since the capability of LLMs is predicated on *scaling laws* [47], which link model capability to massive investments in training data and computational resources, training LLMs with competitive capability is challenging due to the two primary resource-intensive phases, pre-training and fine-tuning. Pre-training typically requires petabyte-scale, multi-source datasets drawn from diverse sources such as web crawls, books, source code, and scholarly articles, thereby incurring substantial computational overhead. Fine-tuning (e.g., instruction tuning, alignment) also requires high-quality domain-specific datasets to guide LLMs in learning this domain knowledge. Such large-scale, high-quality datasets incur huge costs for researchers and developers during both the preparation and model training phases, hindering the development of LLMs in practice, especially in resource-constrained institutions. For instance, data preparation of BigCodeBench [143] required a year-long collaboration involving 20 expert annotators to rigorously construct. Conversely, training Llama 3.1 405B consumed approximately 31 million H100 GPU hours to process 15.6 trillion tokens, resulting in a compute cost estimated at over $100 million [25].

To address these challenges, a promising research direction has been proposed and explored: data-efficient LLM training. The central hypothesis of data efficiency here is that the information density and relevance of the training data are the true drivers of the capability of LLMs. This perspective suggests that massive datasets often contain significant redundancy, noise, and low-utility samples that marginally contribute to learning while increasing computational costs. Therefore, data-efficient construction aims to identify the *minimal sufficient set* of datasets for LLMs to learn target distributions and capabilities.

In this context, this paper provides a comprehensive survey of data-efficient LLM training. We define it as the systematic methodology of building or adapting LLMs through the rigorous optimization of both pre-training and fine-tuning data utility. In total, we collect 116 works published between 17 Apr 2020 and 7 Aug 2025. Of these, 18 works focus on data efficiency in the pre-training process (including one work that targets both pre-training and fine-tuning stages), 98 works address the fine-tuning process. After excluding 9 benchmarks and empirical studies, we focus on the remaining 107 technique-oriented works. According to their design spirit, we further divide them into surrogate model-based methods (78) and surrogate model-free methods (29). Ultimately, the goal is to enable the construction of high-performance LLMs with significantly reduced data volume and computational expenditure.

Based on the analysis of existing works, we reveal pitfalls and provide future research directions. For example, of 87 surrogate-based works, only 17 works reported the external cost of preparing and using surrogate methods, hindering the practical deployment of their methods. Moreover, only 44 works conducted robustness evaluation of their trained LLMs. It is unclear if there is a trade-off between cost reduction and robustness degradation. In the future, research should focus more on multi-objective and distribution-aware optimization, as well as on theoretical guarantees.

Comparison with existing surveys. Some survey works have begun to address the resource constraints of machine learning. For example, Zha *et al.* [121] established the foundational principles of data-centric AI. However, their scope encompasses general machine learning modalities, such as computer vision and tabular data, and lacks a concentrated focus on the unique token-level dynamics and scaling behaviors inherent to LLMs. Subsequent work by Albalak *et al.* [3] provided a comprehensive review of data selection, specifically unifying methods across language model development stages under a taxonomy based on utility functions and selection mechanisms. Different from their work, we focus more on the two phases of pre-training and fine-tuning, systematically categorize each work based on their surrogate model dependency and data selection criteria, and provide a comparison of methods designed for these two phases. Luo *et al.* [68] explored efficient training from a data-centric perspective. While their analysis predominantly emphasizes post-training and alignment stages, it provides limited granularity on the foundational pre-training phase. To the best of our knowledge, our survey is the first one to systematically categorize data-efficient training and fine-tuning methods by (i) their reliance on external models versus heuristic approaches, and (ii) the specific sampling and filtering mechanisms employed at the method level. This structured approach clarifies the trade-offs between computational overhead and data quality, offering a comprehensive roadmap for building high-performance LLMs under constrained resources.

To summarize, the main contributions of this study are as follows:

- **Unified problem definition:** We establish a unified problem definition of data-efficient LLM training.
- **Systematic literature analysis:** We conduct a thorough survey of 116 articles. We reveal pitfalls in existing works and provide good practice guidance.
- **Vision:** We present promising research opportunities to facilitate this domain.

The remainder of this paper is organized as follows. Section 2 introduces key preliminaries. Section 3 defines the problem of data-efficient LLM training. Section 4 details our methodology for paper collection. Section 5 presents a systematic summarization and analysis of the collected literature, followed by a broader Discussion in Section 6. Section 7 highlights key research opportunities. Finally, Section 8 concludes the survey.

2 Preliminaries

2.1 Training of LLM

Pre-training. Based on pre-training, the model learns general knowledge from the large-scale data collected in the previous step. This process requires massive computational resources, often using supercomputers composed of thousands of high-performance GPUs and high-speed networks, and can take several weeks or even months to train the deep neural network parameters and build the foundation model. For instance, GPT-3 [9] was trained on over 1,000 NVIDIA GPUs over a prolonged period. To facilitate the pre-training, multiple techniques have been proposed with promising results, such as Zero Redundancy Optimizer (ZeRO) [85], Sparse Mixture-of-Experts (MoE) [27], FlashAttention [19], which are mainly designed to improve training efficiency, while some of them also benefit inference. The pre-trained models can be further fine-tuned for downstream tasks.

Fine-tuning. Fine-tuning is the process of adjusting the parameters of LLMs to adapt them to specific tasks. It does not train the model from scratch, but uses a pre-trained model as a starting point, preserving its existing general knowledge. On this basis, the model is further trained on domain- or task-specific datasets to optimize performance in particular scenarios, such as finance or healthcare. This approach reduces computational costs and allows leveraging cutting-edge models. For example, the famous LLM DeepSeek-Coder-V2 [139] is fine-tuned from DeepSeek-V2 [63].

2.2 Data Selection

Data Selection is the process of identifying and curating an optimal subset of data from a large source data pool. It can affect both the training and testing phases of the LLM construction pipeline. In this work, we focus on data selection for training and fine-tuning. That is, by training on a smaller, higher-utility subset, data selection aims to achieve a target level of model capability while drastically reducing the associated data labeling and computational costs. In the remaining parts, we use *data selection* as the unique terminology that encompasses several related processes:

- **Data Filtering/Curation**: The process of removing low-quality, redundant, toxic, or privacy-compromising data from a large corpus. This is often a technique in the pre-training stage.

- **Data Sampling**: The active selection of a subset of high-value data points from a larger pool. This is common in fine-tuning, where techniques like active learning [89] are used to identify the most informative instruction-response pairs.

3 Problem Definition

In this section, we formally define the problem of data-efficient LLM training, including both pre-training and fine-tuning processes.

> Given an LLM M, a source data corpus D_{pool}, a test data corpus D_{test}, a budget constraint $Budget$, data-efficient LLM training is a problem of selecting a subset D_{select} of D_{pool} such that $|D_{select}| \leq Budget$ and $D_{select} = \arg\max_{D_i \subseteq D_{pool}} \varrho\,(M', D_{test})$ where M' is M trained with D_{select} and ϱ is a capability measurement.

.

In pre-training, M is the initialized LLM and D_{pool} is the collected pre-training raw data. In fine-tuning, M is a pre-trained LLM and D_{pool} is the domain-specific fine-tuning dataset.

4 Paper Collection

4.1 Survey Scope

Our survey focuses on data-efficient training of LLMs. Here, data-efficient indicates training or fine-tuning LLMs with a subset of the available data pool to reduce the data processing and computation costs, as detailed in Section 3. We use the following three selection criteria to collect papers. A paper must satisfy at least one of the criteria to be included in our survey.

(1) The paper proposes a new technique that targets data-efficient LLM construction.
(2) The paper empirically studies/analyzes existing data-efficient LLM construction techniques.
(3) The paper introduces benchmarks, datasets, or criteria that are specifically designed for data-efficient LLM construction.

As we focus on language models, some works that also lie in the data-efficient scope but focus on other tasks, e.g., vision tasks, are not included in our survey. For instance, seminal works on coreset selection [76], data selection via proxy models [16], and analysis of example forgetting [96] are primarily benchmarked on image classification tasks and are thus outside the scope of this survey.

4.2 Paper Collection Workflow

We follow previous survey works [13, 39, 45, 49, 132], to design our paper collection workflow. As illustrated in Figure 1, the workflow has three key steps: keywords-based paper search, multi-stage paper filtering, and snowballing. To conduct the first step, we design the following keywords.

- language model + active learning | subset/data selection | influence function | data sampling | data filtering | fewer data | label-efficient | data-efficient | annotation efficient

Based on the keywords, the automatic paper search process is employed to collect relevant works from five commonly used resources: Google Scholar, IEEE Xplore, ACM, Arxiv, and DBLP. This initial broad retrieval yields a total of 27,635 papers, with Google Scholar contributing the largest set (17,500 papers), followed by ACM (4,484 papers) and IEEE Xplore (3,371 papers).

After that, a rigorous multi-stage study selection protocol is applied to filter the papers. We first screen papers by title, significantly reducing the pool to 2,251 candidates. Subsequent filtering by venue quality further narrows the selection to 1,163 papers. We then perform a manual abstract review to verify alignment with the survey's scope, resulting in 297 papers. At the same time, we filter studies with fewer than eight pages to remove short papers, leaving 251 papers. After

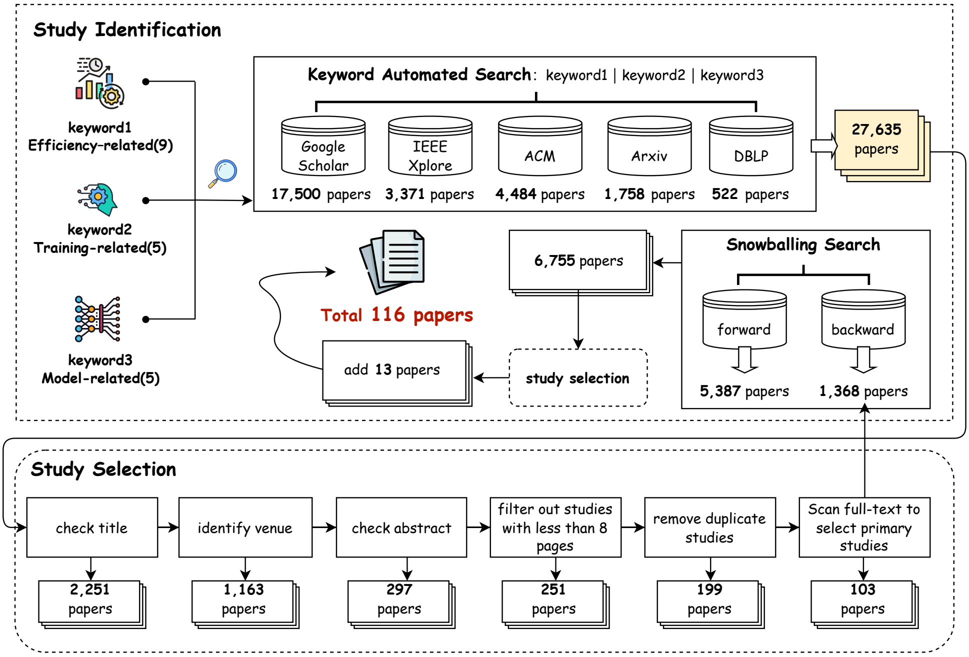


Fig. 1. Study collection and selection.

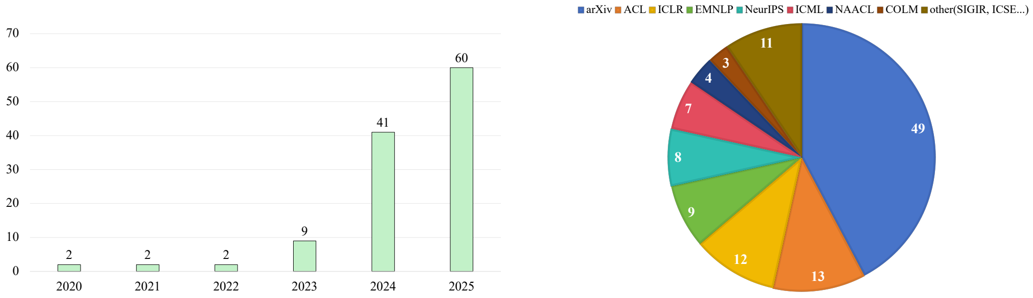
removing duplicate entries across databases, which leaves 199 unique papers, a full-text scan is conducted to identify the primary studies, resulting in a core set of 103 papers.

Finally, to mitigate the potential limitations of keyword-based retrieval and ensure the comprehensiveness of our survey, we employ snowballing, a complementary search strategy that iteratively expands the pool of literature by exploring the citation networks of the selected primary studies. This process follows the standard guidelines for performing forward and backward searches. Forward snowballing involves examining papers that cited our primary studies, identifying 5,387 candidates, while backward snowballing entails reviewing the reference lists of the selected papers, yielding 1,368 candidates. From this combined pool of 6,755 papers, we apply the same screening criteria used in the last phase. This rigorous expansion identifies 13 additional relevant papers. Consequently, this survey comprises a total of 116 papers.

4.3 Statistics of Collected Papers

Publication Trend. Figure 2a reveals a distinct trend of accelerating research activity. The initial period from 2020 to 2022 represents a preliminary phase in this specific domain, with a consistent but low volume of two papers identified in each year (2020, 2021, and 2022). A noticeable increase occurred in 2023, where the number of identified papers rose to 9. The most significant expansion is observed in the most recent years; the volume of literature grew to 41 papers in 2024 and peaked at 60 papers in 2025. This trajectory indicates that the vast majority of the research within our collected corpus (approximately 87%) has emerged within the last two years, accompanied by the boom of LLMs, underscoring the topic's rapidly growing urgency.

Venue Distribution. The publication venues of our covered studies are depicted in Figure 2b. 49 papers originate from arXiv, reflecting the fast-paced nature of LLM research, where rapid



(a) Distribution of papers by year.

(b) Publication venues distribution.

Fig. 2. Statistical overview of the collected papers.

dissemination often precedes formal peer review. Among peer-reviewed venues, the International Conference on Learning Representations (ICLR) and the Association for Computational Linguistics (ACL) are the leading contributors, publishing 12 and 13 papers, respectively. Other significant contributions are identified in EMNLP and NeurIPS, publishing 9 and 8 papers respectively, followed by ICML (7 papers). The remaining papers are distributed across NAACL (4 papers), COLM (3 papers), and other venues such as SIGIR and ICSE (in total of 11 papers). This distribution suggests that while the field relies heavily on pre-prints for speed, it is well-represented across the most prestigious conferences in both general machine learning and natural language processing.

5 Paper Summarization

We first categorize these works from two perspectives: the surrogate model dependency and the data selection criteria. Then, we introduce each work in detail.

Surrogate Model Dependency. Based on the usage of surrogate models, existing data-efficient techniques can be divided into surrogate-free methods and surrogate-based methods.

(1) Surrogate-Free Methods: The data selection process is self-contained and relies exclusively on the target model itself. Data priority is determined by analyzing the target model's internal states, such as its parameters, outputs, or training dynamics.
(2) Surrogate-Based Methods: These methods employ an external model distinct from the target model to generate data selection scores. Existing works have considered different types of surrogate models:
    - Representation: The surrogate functions as a fixed encoder (such as a pre-trained BERT) that generates embeddings to facilitate downstream selection strategies like clustering.
    - Scoring: The surrogate serves as a powerful large language model (e.g., GPT-4) that directly evaluates data samples based on perceived quality, complexity, or relevance.
    - Utility Prediction: The surrogate operates as a separate, often smaller, model explicitly trained to predict the utility of specific data samples for training the target model.

The choice between these two types of data selection involves critical trade-offs. Surrogate-based methods can offer superior selection precision by leveraging a more powerful model, but they introduce system complexity, computational overhead, and potential bottlenecks. Surrogate-free methods are more tightly integrated into the target model training loop but are bounded by the capabilities of the model.

Selection Criteria Categorization. Beyond the architectural choice of how a selection score is computed (i.e., surrogate dependency), existing methods can be further categorized into different groups based on what property of the data is being measured. The choice of method, or selection criterion, reflects the underlying hypothesis about what constitutes a *valuable* data sample. Specifically, we summarize five types of selection methods: training dynamics and influence-based methods, uncertainty methods, data diversity-based methods, intrinsic data quality-based methods, and heuristic and composite methods.

(1) Training dynamics and influence. These methods are derived directly from the model's interaction with the data during the training process [8, 106].
    - Optimization and gradient-based methods. This kind of method considers sample loss as a common indicator of training difficulty.
    - Gradient. Methods treat the model gradient as the selection reference, such as its norm or its alignment with the average batch gradient.
    - Data impact. Methods estimate the influence of data samples on the model's parameters or its performance on a separate validation set (e.g., via influence functions).
(2) Uncertainty. These methods prioritize data about which the model is least confident. The hypothesis is that resolving this uncertainty provides the highest information gain. Common methods include output entropy for classification [80, 125].
(3) Data diversity. This category focuses on the properties of the selected subset as a whole, rather than just individual sample scores [62, 98].
    - Diversity. Methods are designed to ensure the selected subset covers the largest portion of the feature space of the original dataset. This is often achieved by avoiding the selection of samples that are too similar to already-selected data.
    - Representativeness. Methods that select samples near the centroids of data clusters, to capture the core of the distribution.
(4) Intrinsic data quality. These methods assess the inherent quality of the data, often independent of the target model. This evaluation typically requires a powerful surrogate model to score textual attributes independent of the model's training state. [14, 136].
(5) Composite and multi-objective strategies. Some research introduces novel or composite methods. These methods often combine criteria or are custom-engineered heuristics designed to capture a specific notion of data *value* not covered by the categories above [73, 127].

The distribution of our collected works based on our taxonomy reveals a significant predominance of surrogate-based strategies, which account for 78 papers, compared to 29 surrogate-free approaches and 9 empirical benchmarks. We will introduce each work based on its surrogate model dependency and selection criteria separately, followed by 9 works that build benchmarks or conduct empirical studies to support data-efficient construction of LLMs.

5.1 Pre-Training

Table 1 presents the distribution of our collected works based on our taxonomy.

5.1.1 *Surrogate-Based Methods*

5.1.1.1 *Training Dynamics and Influence*

Brandfonbrener *et al.* [8] introduced conditional loss reduction filtering (CoLoR-Filter), a targeted data selection method derived from an empirical Bayes-inspired objective. CoLoR-Filter utilizes a pair of small auxiliary models as surrogates: a marginal (or prior) model trained on general data, and a conditional model obtained by fine-tuning this auxiliary model on a small dataset from the downstream task. Then, the data selection is based on the difference in log-likelihood between these two models, prioritizing examples where the conditional model achieves significantly lower loss than the marginal model. Yu *et al.* [120] addressed the problem that static data selection methods

Table 1. Statistics of different data selection studies (pre-training).

Surrogate Model Dependency	Selection Criteria Categorization	References
Surrogate-Free Methods	Training Dynamics and Influence	–
	Uncertainty	–
	Data Diversity	[86]
	Intrinsic Data Quality	–
	Composite and Multi-Objective Strategies	–
Surrogate-Based Methods	Training Dynamics and Influence	[8], [120], [140], [33]
	Uncertainty	[125]
	Data Diversity	[26]
	Intrinsic Data Quality	[32], [103], [59], [142], [50], [114], [73], [55]
	Composite and Multi-Objective Strategies	[88], [93], [54]

fail to capture the dynamically shifting data preferences of LLMs during pretraining, and proposed model-aware data selection with data influence models (MATES), a framework where a small data influence model continuously adapts to the pretraining model's evolving state to select the most effective data on-the-fly. The method first probes the main model locally to collect *oracle data influence*, quantified as the change in loss on a reference task after a single training step on a specific data point. This oracle data is then used to fine-tune the small data influence model for predicting the best data for the next training stage.

Gu *et al.* [33] designed MiniPLM, which enhances pre-training quality and efficiency through a teacher-guided Difference Sampling strategy. Unlike static filtering, it adaptively refines the data distribution by targeting the relative difficulty of instances, effectively filtering noise and down-sampling saturated simple patterns. By eliminating ineffective gradient updates from overly simplistic or noisy data at the early stages of pre-training, this approach significantly boosts the effective sample throughput. This active management of training dynamics exposes the student model to a more concentrated stream of diverse and complex knowledge, thereby facilitating faster convergence and better generalization across diverse model architectures. More recently, Zhu *et al.* introduced ToReMi [140], a two-stage, topic-aware framework that dynamically reweights pre-training data based on topical associations and training dynamics. In the first stage, it prioritizes challenging topics by increasing weights for high-loss categories to accelerate foundational knowledge acquisition. In the second stage, it attenuates weights for persistently high-loss topics to minimize noise interference. The framework employs LLM-based topic annotation through clustering and either generating custom labels or selecting from existing taxonomies.

*5.1.1.2 **Uncertainty***

Zhang *et al.* [125] proposed SIEVE, a system designed to solve the prohibitive cost of using high-performance models like GPT-4o for filtering web-scale LLM pretraining data. SIEVE's method involves a novel stream-based active learning algorithm that efficiently distills GPT-4o's filtering decisions into a lightweight classification model, specifically using a True Risk Minimizer (TRM) threshold to handle the severe class imbalance found in filtering tasks. Extensive experiments demonstrate that SIEVE matches GPT-4o's filtering accuracy, confirmed by both automatic and human evaluations, while reducing the cost by over 99%.

*5.1.1.3 **Data Diversity***

Fan *et al.* [26] proposed diversified file selection (DiSF), an algorithm that employs a pre-trained surrogate model as a feature extractor to map text samples into an embedding space, where it selects files that minimize the Frobenius norm of the feature covariance matrix to ensure diversity. DiSF addresses the diversity dilemma observed in prior domain-similarity-based selection methods, where focusing on similarity to high-quality domains improves domain-specific performance but degrades generic capabilities due to reduced feature diversity. By prioritizing decorrelated samples within the surrogate model's embedding space, DiSF effectively prevents this collapse and significantly enhances overall model capabilities across diverse benchmarks.

5.1.1.4 *Intrinsic Data Quality*

Li *et al.* [55] proposed a data pre-processing framework called generative deduplication (GD). This framework trains a generative model using a self-supervised keyword prediction task for only one epoch, integrating Time-dimensional Gaussian Noise (TGN) to increase training difficulty and limit the model's ability to learn from non-duplicate samples. During inference, texts for which the model accurately predicts the target keyword are identified as semantically duplicate and removed. Experiments show the GD framework effectively reduces training samples by 50.9% and time by 42.9% while improving downstream task performance.

Gu *et al.* [32] proposed PMP-based data selection (PDS), a framework that formulates pre-training data selection for language models as a generalized optimal control problem. PDS leverages Pontryagin's Maximum Principle to derive necessary theoretical conditions characterizing optimal data selection based on LM training dynamics, which are then approximately solved on a proxy dataset to compute quality scores used for training a data scorer model. PDS can reduce data demand by up to 1.8 times in data-constrained scenarios. Wettig *et al.* [103] introduced QuRating, a method for selecting pre-training data that captures human intuitions about text quality, moving beyond simple heuristics. The approach uses GPT-3.5-turbo to generate pairwise judgments on texts based on four abstract criteria: writing style, facts & trivia, educational value, and required expertise. Then, a *QuRater* model is trained on these 250,000 judgments for each criterion to learn scalar quality ratings, which are used to annotate the quality of a 260 B token corpus.

Similar to QuRating, to address the limitation of methods that rely on heuristics or single quality signals, Xu *et al.* [59] proposed flexible integration of data quality ratings for effective pretraining (FIRE). FIRE integrates ratings from multiple diverse data quality raters by first aligning their scores into a unified space and then integrating these aligned scores, weighting them by both rater reliability and orthogonality derived via PageRank. Evaluation demonstrated that FIRE is 2.5 times faster than random selection. Furthermore, to address limitations of single-dimensional data selection methods for LLM pre-training, which often overlook the multifaceted nature of data quality, Zhuang *et al.* [142] introduced Meta-rater, a multi-dimensional framework that first defines four new quality metrics: professionalism, readability, reasoning, and cleanliness (PRRC). The method then learns the optimal weightings for these and 21 other quality scores by training numerous small proxy models on differently weighted data subsets and fitting a regression model to their resulting validation losses.

As most of the existing methods overlook the specific domain nuances required for mathematical reasoning, Li *et al.* [50] designed mathematical data selection via skill graphs (MASS). MASS first constructs a skill graph by prompting a strong LLM to extract mathematical skills and their co-occurrence relationships from a high-quality reference dataset. This graph is then used to score data in a large target dataset by aggregating semantic similarities, prioritizing texts that cover important individual skills (nodes) and skill compositions (edges). Evaluation showed that MASS can reduce 50% to 70% of token usage. More recently, Maharana *et al.* [73] introduced D^2 PRUNING, which models the dataset as an undirected graph where nodes are initialized with difficulty scores and edges are weighted by embedding distance. The method uses forward message passing to update node scores based on their neighbors, balancing difficulty and density, and then uses reverse message passing during selection to down-weight neighbors of selected nodes to ensure diversity.

This joint optimization approach allows D^2 PRUNING to outperform state-of-the-art methods on vision and NLP datasets at pruning rates up to 70%.

Focusing on programming tasks, Yang *et al.* [114] proposed synthetic corruption informed pruning (SCIP), a method that first identifies the characteristics of low-quality code in an embedding space by analyzing synthetically corrupted data (e.g., removing brackets or altering conditionals). This analysis revealed that corrupted data tends to move to smaller embedding clusters or farther from cluster centroids. Based on this insight, SCIP prunes the dataset by ranking and removing a fraction of data points based on their small cluster size and large distance from the centroid. Experiments demonstrated that SCIP is twice as fast as basic pre-training.

5.1.1.5 *Composite and Multi-Objective Strategies*

Some works considered combining different selection criteria to boost data selection. Sachdeva *et al.* [88] proposed ASK-LLM, a method that combines quality and coverage for data selection. Here, quality is quantified by an instruction-tuned proxy LLM's probability of a *yes* response. For coverage, ASK-LLM introduces DENSITY sampling, which models the data distribution using kernel density sums and then applies Inverse Propensity Sampling (IPS) to select a diverse sample. In an extensive comparison of 19 samplers, the coverage-based DENSITY method successfully recovered full-data performance, while the quality-based ASK-LLM consistently outperformed full-data training, converging up to 70% faster even when rejecting 90% of the original dataset.

Shum *et al.* [93] proposed predictive data selection (PRESELECT). The core idea is to quantify a sample's *predictive strength*: the degree to which a model's compression loss on that sample correlates with the model's downstream benchmark performance. By training a simple fastText scorer on a small seed set of documents with high and low predictive strength, PRESELECT can scalably filter enormous datasets. Li *et al.* [54] considered combining data generation and data selection to search for better pre-training data, and designed a framework for rule-based data selection. Concretely, it uses an LLM to first generate a diverse pool of data quality rules. The core of the method involves rating a data batch against these rules to create score vectors, then applying a Determinantal Point Process (DPP) to select a final subset of independent rules based on the orthogonality of these vectors.

5.1.1.6 *Surrogate Model Analysis*

Table 2 summarizes the distribution of the surrogate model used in the above methods. The choice of surrogate models is quite diverse (evenly distributed). For representation-based surrogate models, the BERT family is the most popular. And the GPT family and the LLaMA family are most used as scoring and selection surrogate models, respectively.

5.1.2 *Surrogate-Free Methods*

Only one diversity based work did not rely on surrogate models for pre-training data selection. Renduchintala *et al.* [86] proposed INGENIOUS, which elects a diverse and representative subset by maximizing the Facility Location submodular function on selected data samples. In addition, it overcomes scalability challenges by partitioning the data into random blocks, reducing the similarity kernel's memory footprint. Experiments showed that INGENIOUS can reduce training costs by 72%.

Table 2. Usage of Surrogate Models

Surrogate Model Component	#Works
Representation Model	6
Scoring Model	8
Selection Model	7

5.1.3 *Dataset and Model Analysis*

This part analyzes the datasets and models that the researchers focused on during their evaluation of methods during pre-training.

Table 3. Summary tasks and their used works.

Task	References
Instruction-Response Pairs	[142], [32],
Text Classification	[73], [140], [86], [125]
Q&A Pairs	[8], [26], [32], [120], [93], [140],[50], [33]
Text Corpus	[86], [125], [120], [103], [32], [59], [26], [54], [114]
Programming	[55], [114], [93]

As shown in Table 3, existing works mainly focus on five types of tasks: Question & Answer (Q&A) Pairs, Text Corpora, Instruction-Response Pairs, Text Classification, and Code. Text Corpus is the most commonly used task for pre-training, accounting for 34.62% of all tasks. Moreover, the presence of specialized Code datasets (e.g., CodeSearchNet, The Stack) and Instruction-Response pairs (e.g., Alpaca, WizardLM) indicates that data selection research is expanding beyond general text pre-training into domain-specific and alignment-stage optimization [114, 142].

In terms of model architectures, the usage of models ranges from small LMs to state-of-the-art LLMs. The LLaMA family (including versions 1, 2, 3, and 3.1) emerges as the most ubiquitous testbed [54, 93, 142]. This is complemented by the frequent use of the Pythia suite [26, 120], which is often favored in research contexts for its transparency in training dynamics. The analysis also highlights a temporal span in model selection; researchers continue to utilize established baselines like BERT-base and GPT-2 [86, 140] alongside highly capable recent models such as Qwen-2.5, Mistral, and GPT-4o [50, 59]. This diversity indicates that while the field is rapidly adopting newer architectures, there remains a reliance on well-understood, smaller models.

5.2 Fine-Tuning

Table 4. Statistics of different data selection studies(fine-tuning).

Surrogate Model Dependency	Selection Criteria Categorization	References
Surrogate-Free Methods	Training Dynamics and Influence	[18], [133], [131], [102], [79], [138], [21], [111], [112], [53], [46], [124]
	Uncertainty	[80], [128], [134], [101]
	Data Diversity	[92], [2]
	Intrinsic Data Quality	[100]
	Composite and Multi-Objective Strategies	[99], [22], [44], [81], [74],[5], [115], [97], [126]
Surrogate-Based Methods	Training Dynamics and Influence	[108], [1], [60], [106], [4], [105], [90], [51], [136]
	Uncertainty	[87]
	Data Diversity	[7], [20], [117], [62], [98], [110], [118]
	Intrinsic Data Quality	[14], [130], [12], [71], [82]
	Composite and Multi-Objective Strategies	[119], [64], [94], [31], [69], [56], [127], [10], [78], [123], [11], [83], [95], [30], [15], [24], [122], [35], [6], [34], [84], [135], [57], [109], [38], [113], [23], [75], [52], [58], [29], [37], [28], [72], [104], [65], [129], [61], [67]

5.2.1 *Surrogate-Based Methods*

5.2.1.1 *Training dynamics and influence*

Xu *et al.* [108] introduced Bayesian data point selection (BADS), a framework that treats data selection as a posterior inference problem. BADS utilizes an auxiliary weighting network as a surrogate model to predict importance weights for training examples, treating both the main model parameters and these instance-wise weights as random variables. By inferring their joint posterior using stochastic gradient Langevin dynamics (SGLD) sampling, the method effectively aligns the training data distribution with a small, high-quality meta-set.

Zhou *et al.* [136] introduced data selection via implicit reward (DavIR), to quantify the *learnability* of a data point based on the relative reduction in loss between a pre-trained model and a reference model fine-tuned on the full dataset. Experiments demonstrated that models fine-tuned on only 6% of the dataset selected by DavIR achieve superior performance compared to models trained on the full dataset. Agarwal *et al.* [1] proposed NN-CIFT to leverage a surrogate neural network, termed InfluenceNetwork, to estimate data influence scores. Here, the influence score is quantified as the output of InfluenceNetwork, which takes the concatenated embeddings of the training and target data points as input. InfluenceNetwork is trained on a small subset of ground-truth influence values derived from LLMs and subsequently employed to predict influence scores for the remaining dataset. Empirical evaluations demonstrated that NN-CIFT achieves a data valuation cost reduction of up to 99%. Lin *et al.* [60] introduced LEAD, a framework to utilize multiple training information for data selection. Specifically, it first employs an offline dual-level clustering strategy, using fixed task-specific embeddings to group semantically similar instructions. During training, LEAD integrates a coarse-to-fine selection process: a multi-armed bandit (MAB) mechanism dynamically prioritizes these pre-computed clusters, and then combines instantaneous loss, historical smoothing, and gradient-based approximation to identify high-utility samples within the selected cluster. Evaluation demonstrated that LEAD improves model performance by up to 10.8% using only 2.5% data.

Addressing the high computational cost bottleneck of gradient-based data valuation, Ananta *et al.* [4] introduced quantized low-rank gradient similarity search (QLESS), a data selection framework that integrates absmax-based quantization with LoRA-based random projection to compress the gradient datastore used for influence estimation. Evaluation highlighted that QLESS reduces memory usage by up to 16× while maintaining downstream fine-tuning performance. Recently, Shen *et al.* [90] proposed safety-enhanced aligned LLM fine-tuning (SEAL) to use bilevel optimization to learn a data ranker for fine-tuning. This method addresses the critical problem where fine-tuning LLMs on task-specific data, even benign samples, can significantly compromise the model's pre-established safety alignment. SEAL formulates an optimization problem where the data selector is trained to up-rank safe, high-quality data and down-rank unsafe samples, ensuring the resulting fine-tuned model still fits a separate, safe alignment dataset.

In addition to domain-specific fine-tuning, multiple works have addressed the problem of data-efficient instruction tuning, which is a specific type of fine-tuning. Xia *et al.* [106] proposed low-rank gradient similarity search (LESS). The core idea of LESS is to use gradients as an influence function to guide the data selection, where cosine similarity is employed to normalize gradients, mitigating a performance-hurting bias against variable-length instruction data. Li *et al.* [51] designed selective reflection-tuning, a teacher-student collaboration paradigm for instruction selection. The pipeline involves two phases: a *Selective Instruction Reflection* phase, where the student model uses the instruction-following difficulty (IFD) score to select challenging instructions refined by the teacher, and a *Selective Response Reflection* phase, where the student uses a novel reversed-IFD (r-IFD) metric to select feasible and informative responses. After that, Wu *et al.* [105] propose ROSE, a reward-oriented instruction data selection framework. To fill the gap of misalignment in existing methods where minimizing next-token prediction loss often fails to correlate with actual task performance, ROSE employs a surrogate model initialized on a small random subset to estimate the gradient-based influence of training samples on a task-specific reward signal derived from direct preference optimization (DPO) loss. Experiments demonstrate that training on just 5% of data selected by ROSE achieves competitive results against the full dataset.

5.2.1.2 ***Uncertainty***

Ru *et al.* [87] introduced adversarial uncertainty sampling in discrete space (AUSDS) for active sentence learning. AUSDS leverages pre-trained language models to map discrete sentences into a continuous latent space, where adversarial attacks are employed on labeled batch data to efficiently identify regions near the decision boundary. Subsequently, k-nearest neighbor (KNN) search is performed within the latent representations of the entire unlabeled corpus to retrieve actual text samples residing in these high-uncertainty regions.

5.2.1.3 *Data Diversity*

Bashar *et al.* [7] proposed the mixed aspect sampling (MAS) framework, an active learning approach designed to mitigate the inherent linguistic bias of LMs by selecting a semantically diverse training set. MAS is an ensemble of two classifiers: 1) one focusing on linguistic aspects (using a fine-tuned LM), and 2) another on topical aspects (using a topic model). Based on this, it categorizes unlabeled instances into *agreed-minority, agreed-majority,* and *disagreed* groups. By applying cluster-based down-sampling to these groups, particularly prioritizing disagreed instances, MAS efficiently identifies the most informative and diverse samples for human annotation.

Das *et al.* [20] introduced DEFT-UCS, which leverages a pre-trained embedding model (Sentence-T5) as a surrogate to map input data into a latent space. This surrogate embedding allows the framework to identify and sample a representative core-set encompassing both easy and hard samples relative to cluster centroids. Based on DEFT-UCS, fine-tuning 32.5% of the original dataset can achieve comparable or improved accuracy compared to fine-tuning with the full dataset. Similar to DEFT-UCS, Yu *et al.* [117] proposed a diversity-centric framework utilizing k-means clustering coupled with an iterative refinement strategy. The method employs a surrogate scoring model to evaluate the quality gap between generated and reference responses, thereby dynamically re-weighting clusters to prioritize high-value data regions and filter out noise during training. Other works also leverage different diversity quantification metrics for data selection. Ling *et al.* [62] used semantic entropy of outputs as a diversity indicator, and Wang *et al.* [98] utilized a surrogate model to generate normalized weight gradients as data representations, which are then used to compute a novel metric called log determinant distance that measures the diversity. Yang *et al.* [110] proposed to employ Sparse AutoEncoders to measure data diversity based on activated features.

Addressing the problem that NLP models trained on data from one domain often fail to generalize to other domains, Yu *et al.* [118] suggested focusing on inter-dependencies between samples rather than just quality. They introduced three metrics to quantify the diversity of a dataset based on sentence embeddings: Max Dispersion (sum of pairwise Cosine distances), Convex Hull Volume, and Graph Entropy. To efficiently select a subset that maximizes these diversity metrics, the paper proposes the diversity advanced actor-critic framework, a reinforcement learning approach that uses an actor-critic model to select the most diverse samples.

5.2.1.4 *Intrinsic Data Quality*

Chen *et al.* [14] introduced ALPAGASUS to employ an LLM as an automatic evaluator to score each data sample based on its response to the instruction with input token *accuracy.* ALPAGASUS can provide a 5.7× training speedup based on their evaluation. Zhang *et al.* [130] introduced EDGE, a framework centered on a newly proposed guideline effectiveness metric. Here, the guideline indicates a summary of the key strategies or rules derived from an agent's interaction trajectory. EDGE evaluates the utility of guidelines using a surrogate model by comparing the cross-entropy loss of generations conditioned on prompts with and without guidelines. Using EDGE can reduce data requirements by up to 75% according to the evaluation. Moreover, Cao *et al.* [12]suggested using natural language indicators (e.g., reward model scores) to evaluate data quality for the selection, and Ma *et al.* [71] used neuronal activation states to quantify the data quality.

To address the problem of token-level noise in supervised fine-tuning, where uninformative, redundant, or harmful tokens within samples can degrade model performance, Pang *et al.* [82] proposed a generic token cleaning pipeline to evaluate each token's quality. Here, quality is measured by the loss disparity between a fixed base model and a reference model, utilizing either a static warm-up model (fine-tuned on a small high-quality subset) or a self-evolving approach where the latest updated model checkpoint serves as the reference model to refine the supervision signal. Different from previous works that focus on data quality, this work targets a finer-grained notion of quality at the token level.

5.2.1.5 Composite and Multi-Objective Strategies

Similar to pre-training-targeted methods, multiple fine-tuning methods integrate different selection criteria to facilitate the precision of data selection. These methods can be divided into 14 groups, as shown in Table 5.

Table 5. Overview of composite and multi-objective data selection strategies based on surrogate models during the fine-tuning phase

Combination of Metrics (Strategy)	**Reference**
Heuristics	[72]
Training Dynamics and Influence & Data Diversity	[69], [84], [113], [52], [61]
Training Dynamics and Influence & Intrinsic Data Quality	[30]
Training Dynamics and Influence & Heuristics	[34], [29], [28]
Uncertainty & Data Diversity	[119], [64]
Uncertainty & Intrinsic Data Quality	[94]
Uncertainty & Heuristics	[135]
Data Diversity & Intrinsic Data Quality	[31], [10], [78], [123], [11], [83], [15], [75], [58], [37], [104], [67], [95]
Data Diversity & Heuristics	[122], [35], [38], [65]
Intrinsic Data Quality & Heuristics	[57], [109]
Training Dynamics and Influence & Data Diversity& Intrinsic Data Quality	[23]
Training Dynamics and Influence & Data Diversity& Heuristics	[129]
Uncertainty & Data Diversity & Intrinsic Data Quality	[56], [127], [6]
Data Diversity & Intrinsic Data Quality & Heuristics	[24]

Heuristics. Witnessing the failure of existing internal computational dynamics-based data selection methods, Ma *et al.* [72] proposed monosemantic neuronal activation-based data selection (MONA) to use a Sparse Autoencoder to disentangle activations into sparse, monosemantic (single-concept) representations. This model-centric embedding is then combined with a Generalized Jaccard similarity metric to select source data that aligns with the target task's activation patterns. Experimental evidence stated that 5% of fine-tuning data is enough when using MONA.

Training Dynamics and Influence & Data Diversity. Lv *et al.* [69] proposed the code adaptive compute-efficient tuning (CodeACT) framework, which uses instruction-following score to measure the influence, followed by clustering methods to select a high-quality, complex, and diverse data subset. CodeACT can reduce training time by 78% and peak GPU memory usage by 27%. Lv *et al.*[84] introduced reinforced adaptive instruction selection (RAISE), a dynamic framework that models instruction selection as a sequential decision-making process optimized via reinforcement learning. It employs a trainable acquisition function to estimate the *dynamic value* of each instruction by processing state features such as instruction difficulty and semantics precomputed by an auxiliary model (e.g., Llama-3.1-8B-Instruct). Then, by dynamically selecting high-value, diverse data batches based on these surrogate-derived scores, RAISE outputs the optimal subset. RAISE achieved superior performance while using only 1% of the gradient update steps compared to full-dataset training.

Furthermore, Yang *et al.* [113] proposed SMALLTOLARGE (S2L), a scalable and effective data selection method for domain-specific supervised fine-tuning. S2L works by first training a small proxy model on the dataset, collecting the training loss trajectories for each sample, and then clustering these trajectories to group examples that exhibit similar training dynamics. By sampling from these clusters, S2L creates a small, representative subset to fine-tune a much larger target model. Motivated by the finding that weak and strong models consistently perceive instruction difficulty, Li *et al.* [52] proposed superfiltering. Superfiltering first filters data using IFD scores calculated by a small proxy model (e.g., GPT-2), and subsequently employs a facility location function to select a diverse subset from these high-quality candidates. Similarly, Lin *et al.* [61]

proposed DEALRec to identify influential samples for the efficient few-shot fine-tuning. DEALRec computes an influence score using a lightweight surrogate model, where the score is efficiently approximated via stochastic Hessian-Vector Products to estimate the change in empirical risk caused by removing a sample. To mitigate the capability gap between the surrogate and the LLM, DEALRec incorporates an effort score calculated as the gradient norm of the sample loss, and utilizes a stratified sampling strategy to explicitly address data diversity and coverage.

Training Dynamics and Influence & Intrinsic Data Quality. Gao *et al.* [30] proposed HINT, a novel framework consisting of two main components: hybrid pseudo-labeled data selection and noise-tolerant Training. The hybrid selection module filters low-quality data by combining training loss metrics with a retrieval-based method that checks for label similarity against known data. The noise-tolerant training module further mitigates the impact of remaining errors by employing a symmetric cross-entropy loss function and a consistency regularization objective.

Training Dynamics and Influence & Heuristics. Guo *et al.* [34] introduced preference-oriented data Selection (ProDS). ProDS utilizes a surrogate model warmed up via supervised fine-tuning and direct preference optimization to compute gradients that quantify the alignment of training samples with bidirectional preferences (positive and negative). These preferences identify data samples that actively improve response quality rather than just mimicking features. By employing an annealing algorithm to synthesize these surrogate-derived preference scores, ProDS selects subsets that allow models to outperform those trained on full datasets. After that, Feng *et al.* [28] introduced a Whitened Feature Distance (WFD) metric, which applies Cholesky whitening and normalization to gradient features to mitigate dominant feature bias and follows a similar schedule for data selection.

Recently, Fu *et al.* [29] proposed T-SHIRT, which utilizes the target LLM as a surrogate evaluator to compute a difficulty score. T-SHIRT isolates and assesses only the informative tokens (those with high loss) to accurately measure data quality while ignoring noise. Furthermore, it employs a hierarchical selection strategy that prioritizes samples whose local neighbors exhibit both high average S-difficulty scores and low variance, thereby ensuring selected data is both high-quality and robust. Extensive experiments showed that T-SHIRT can achieve superior performance by using only 5% of the data.

Uncertainty & Data Diversity. Yu *et al.* [119] introduced PATRON, which employs the frozen pre-trained language model equipped with cloze-style prompts as a zero-shot surrogate model. This surrogate generates task-aware pseudo-labels and uncertainty estimates for unlabeled data, which are then refined through a kernel-based uncertainty propagation method using SimCSE embeddings to filter out noise and bias. To ensure diversity in the selected batch, PATRON utilizes a partition-then-rewrite strategy that iteratively optimizes sample selection within clustered embedding spaces.

At the same time, He *et al.* [64] proposed DEITA, a framework designed to identify important instruction tuning data by measuring samples across three dimensions: complexity, quality, and diversity. To achieve this, DEITA integrates an evolution-based method to train specific scorers that predict complexity and quality based on ChatGPT rankings, subsequently employing a score-first, diversity-aware selection strategy that prioritizes high-scoring samples while filtering out semantically redundant examples using embedding distance.

Uncertainty & Intrinsic Data Quality. Si *et al.* [94] introduced NOVA, a novel data filtering framework combining familiarity and quality. Here, familiarity is quantified via two key metrics: internal consistency probing and semantic equivalence identification. And the data quality score is derived from an expert-aligned reward model. Using an averaged rank, NOVA selects the final data subset for supervised fine-tuning. Extensive experiments demonstrated that NOVA significantly reduced both factuality and faithfulness hallucinations across multiple benchmarks.

Uncertainty & Heuristics. To address the critical problem of knowledge conflicts in domain-specific instruction-tuning, Zhong *et al.* [135] proposed a knowledge-aware data selection (KDS). KDS identifies that existing data selection methods, which are often data-centric (focusing on quality or diversity), fail in specialized domains since their selections contradict the LLM's pretrained knowledge. KDS quantifies these conflicts by first generating multiple responses from the LLM for each data sample and then applying two novel metrics: knowledge alignment, which uses an NLI model to measure the alignment between the LLM's responses and the reference answer, and knowledge consistency, a reference-free metric that measures the semantic uncertainty among the multiple generated responses.

Data Diversity & Intrinsic Data Quality. Researchers prefer to consider both data diversity and quality when designing their data selection methods. Ge *et al.* [31] proposed clustering and ranking (CaR), a two-step data selection method. CaR first ranks instruction pairs using a compact instruction pair quality scoring model that is aligned with human-expert preferences. Second, it employs k-Means clustering to preserve dataset diversity. Bukharin *et al.* [10] designed quality-diversity instruction tuning (QDIT), an algorithm for automatically selecting small, effective instruction tuning datasets. QDIT quantifies diversity using the facility location function based on instruction embeddings and combines this score linearly with an external quality score derived from a pre-trained reward model or a high-capacity LLM, using a greedy algorithm to select the subset that maximizes this combined objective.

Specifically focusing on mathematical reasoning tasks, Ni *et al.* [78] introduced the quality-aware diverse selection (QaDS) strategy. QaDS integrates a K-center Greedy approach to ensure data diversity. For data quality, it estimates the quality from the positive influence samples exert on each other during a "one-shot inference" process. This work addresses the lack of validated data selection methods and the unclear understanding of optimal data composition for mathematical reasoning. Zhang *et al.* [123] proposed Quad. Quad first uses an embedding model to generate vector representations for the input data, clusters the dataset, and then employs a Multi-Armed Bandit framework, treating each cluster as an arm to iteratively balance selecting high-quality clusters with selecting less-visited clusters. To measure quality, it introduces a novel influence function that incorporates attention layers for richer semantic detail and uses the Kronecker product to accelerate the computationally expensive Hessian matrix calculation. At the same time, Pang *et al.* [83] proposed DS^2 (Diversity-aware Score curation method for Data Selection), a novel pipeline designed to address the inaccuracies and biases inherent in using LLMs as cost-effective data raters for instruction tuning. DS^2 models these error patterns by estimating a score transition matrix, which is derived without ground-truth labels by leveraging k-NN statistical information. DS^2 then uses this matrix to correct the initial LLM scores and selects data by first prioritizing these curated quality scores, and secondarily ranking by a long-tail diversity score to ensure sample variety.

Cai *et al.* [11] introduced InfiAlign. The core idea of InfiAlign is an automated data selection pipeline that selects alignment data by evaluating open-source corpora along multidimensional metrics, including diversity (semantic and domain-level), difficulty (using response length as a proxy), and quality (verified by LLMs and rule-based verifiers). This pipeline is followed by a curriculum-guided self-supervised fine-tuning stage and a preference-based DPO stage. Chen *et al.* [15] proposed MIG, an automatic data selection method for instruction tuning that iteratively samples data to maximize information gain in a semantic label graph. The graph models labels as nodes and their semantic relationships as edges. Data quality is integrated via quality scores by a surrogate quality scorer, while diversity is quantified through information distribution and propagation across the graph. Leveraging the submodularity of the information metric, MIG employs a greedy algorithm to efficiently balance quality and diversity. Similarly, Mirza *et al.* [75]

proposed Combination++, which quantifies the quality by the response of a task-specific model, and quantifies the diversity by clustering.

Liang *et al.* [58] proposed Synth-Empathy. The core of it is a novel quality selection step that uses a discriminator model, fine-tuned on the high-quality empatheticdialogues (ED) dataset, to filter the synthetic data based on embedding similarity, followed by a diversity selection step using the K-Center Greedy algorithm. He *et al.* [37] proposed TACOS. It leverages an LLM to assign open-domain tags to instructions, which are then normalized and clustered to capture diversity. Subsequently, TACOS uses an LLM to perform pairwise comparisons of data samples within each cluster, establishing consistent relative quality criteria instead of isolated scoring.

Different from previous works, Wu *et al.* [104] formulated data selection for supervised fine-tuning as a set cover problem and introduced GraphFilter, a method that balances data quality and diversity. The method utilizes a surrogate quality scorer to assign an informativeness score to each sentence, which is then multiplicatively combined with term frequency-inverse document frequency (TF-IDF) to create a unified priority function. The algorithm iteratively selects the sentence with the highest priority, removes its n-grams from the graph, and updates priorities. Lu *et al.* [67] proposed INSTAG, an automatic open-set fine-grained tagging framework that uses ChatGPT to assign semantic and intentional tags to queries in SFT datasets, which are then processed through a systematic normalization procedure. Using the resulting 6.6K tags, the method defines diversity as the tag coverage rate and complexity as the average number of tags per query.

Song *et al.* [95] propose IterSelectTune, an iterative training framework designed to address the problem of efficiently selecting high-quality instruction data for LLM finetuning, which typically requires extensive use of models like GPT-4. IterSelectTune first applies K-means clustering to the source set to derive a Diverse Subset, uses a small BERT-based classifier to automatically identify hard instructions—those where the base LLM's response quality is inferior to the original. This classifier is iteratively trained to mimic GPT-4's judgments on small, diverse data batches, minimizing cost and human involvement. By finetuning a LLaMA2-7B model on just 20% of the source data selected by this method, the resulting model consistently outperformed the model trained on the full dataset across multiple benchmarks, including MT-Bench and AlpacaEval 2.0.

Data Diversity & Heuristics. Zhang *et al.* [122] proposed EQUAL, employing an embedding model fine-tuned via contrastive learning as a surrogate to align document features with QA pair features to enable the calculation of an optimal transport (OT) score. This OT score efficiently estimates the distributional similarity between document clusters and a reference set to guide a multi-armed bandit selection strategy. By leveraging this surrogate-enhanced scoring to prioritize high-value clusters, EQUAL reduces computational costs by 5-10 times. After that, PASER has been proposed by He *et al.* [35]. PASER designs a capability degradation score, which is calculated using Jensen-Shannon divergence (JSD) between the original and pruned models' output distributions, as a surrogate metric to quantify capability loss. In this way, PASER allows adaptive allocation of data budgets to the most severely affected clusters identified via semantic-structural clustering.

He *et al.* [38] proposed SHED. Similar to previous works, SHED first uses model-agnostic clustering to group the data and select representative samples, and then uses a proxy-based Shapley calculator to assign a quality score to each cluster based on that proxy's contribution. Finally, optimization-aware sampling builds the curated dataset from these clusters. Liu *et al.* [65] introduced task-specific data selection (TSDS), which uses the feature space of surrogate model encodings to compute a distribution alignment loss. In this way, TSDS quantifies and minimizes the discrepancy between the selected subset and the target distribution. To further enhance sample quality, the method introduces a novel regularizer utilizing kernel density estimation that explicitly encourages diversity and mitigates the adverse effects of near-duplicates.

Intrinsic Data Quality & Heuristics. Li *et al.* [57] proposed style consistency-aware response ranking (SCAR). SCAR ranks data based on two perspectives: linguistic form (lexical and syntactic choices) and instructional surprise (the predictability of a response given an instruction). Specifically, SCAR trains a ranking model to automatically prioritize instruction-response pairs based on these two types of information. Specifically targeting long chain-of-thoughts instruction tuning, Yang *et al.* [109] proposed SELECT2REASON. Based on the insight that high-utility data exhibits more *rethinking behaviors* like self-correction, it employs a joint ranker that combines question difficulty that quantified using an LLM-as-a-Judge, and the length of the reasoning trace.

Training Dynamics and Influence & Data Diversity & Intrinsic Data Quality. Do *et al.* [23] designed SPaRFT, a self-paced framework that adaptively optimizes training data selection. The method employs a surrogate difficulty estimator to calculate an initial difficulty score for each training example based on its empirical solve rate. Then, these scores are concatenated with semantic embeddings to form clusters for data reduction. By utilizing these surrogate-derived difficulty scores to cluster data and subsequently using a multi-armed bandit algorithm to sample from these clusters based on real-time model performance dynamically, SPaRFT achieves state-of-the-art reasoning accuracy while using up to 100× fewer training samples than standard methods.

Training Dynamics and Influence & Data Diversity & Heuristics. Zhang *et al.* [129] introduced STAFF, considering three types of features for data selection. STAFF begins by calculating a speculative score, derived from the gradient norm of a small model in the same family, to estimate the difficulty of learning each sample. To ensure accuracy, STAFF partitions data into regions and computes a verification result, defined as the ratio of the target model's score to the speculative score for a sampled subset. Finally, a selection budget is dynamically allocated to each region, ensuring that critical data is prioritized while maintaining diversity through stratified sampling. Empirical results showed that STAFF reduces selection time by up to 70.5%.

Uncertainty & Data Diversity & Intrinsic Data Quality. Several works considered combining uncertainty, diversity, and data quality for data selection. CROWDSELECT is the representative one proposed by Li *et al.* [56]. It leverages multi-LLM wisdom to analyze responses from multiple LLMs and their reward scores. It proposes difficulty, separability, and stability metrics derived from this multi-LLM perspective to capture more comprehensive characteristics of instruction. CROWDSELECT addresses the limitation that single-signal methods fail to capture the complexity of instruction-following across diverse fields and may overlook valuable nuances. Bai *et al.* [6] proposed a pre-trained language model-based active learning approach specifically for sentence matching. It addresses the problem that previous active learning methods for NLP primarily rely on entropy-based uncertainty criteria and ignore language characteristics. The proposed method enhances standard uncertainty sampling by introducing three new linguistic criteria derived from a PLM: noise, coverage, and diversity.

More recently, Zhang *et al.* [127] proposed D3, a data selection framework for sample-efficient LLM instruction tuning. D3 involves a scoring step that defines sample distinctiveness (diversity), introduces a novel uncertainty-based prediction difficulty score to measure sample challenge while mitigating context-oriented generation diversity. Moreover, it employs a teacher LLM to assess sample sufficiency and reliability. Subsequently, a selection step formulates a D3 weighted Core-Set objective to jointly optimize these three data value aspects and identify the most valuable subset, allowing for iterative refinement using model feedback. D3 addresses the problem that existing selection methods often rely on single metrics or heuristic staging, failing to comprehensively assess data value or balance criteria effectively for identifying potent subsets from large, potentially low-quality or redundant datasets.

Data Diversity & Intrinsic Data Quality & Heuristics. Finally, Du *et al.* [24] proposed a model-oriented data selection (MoDS) approach for instruction tuning. MoDS iteratively selects data based on three criteria: quality (using a quality evaluation model), coverage (using a K-Center Greedy algorithm for diversity), and necessity (identifying data where an initial fine-tuned model performs poorly). Experimental results show that a model fine-tuned with only 0.2% instruction pairs selected by MoDS performed better than a model fine-tuned on the full original dataset.

5.2.1.6 *Surrogate Model Analysis*

Table 6 summarizes the distribution of the surrogate model used for data-efficient LLM fine-tuning. Different from the pre-training phase, representation-based models are more frequently used as surrogates. In detail, for representation-based surrogate models, the BERT family is the most popular one. The LLaMA family is most used as scoring and selection surrogate models.

5.2.2 *Surrogate-Free Methods*

Table 6. Usage of Surrogate Models

Surrogate Model Component	#Works
Representation Model	37
Scoring Model	26
Selection Model	23

5.2.2.1 *Training dynamics and influence*

Dai *et al.* [18] proposed balanced and influential data selection (BIDS). BIDS employs an iterative selection algorithm to choose training examples that provide the strongest influence on the most underrepresented task currently in the selected subset. Here, influence is quantified via the cosine similarity of gradient projections derived from training and validation examples, followed by column-wise normalization to calibrate influence scales across tasks. The authors demonstrated that a 15% subset selected by BIDS outperforms full-dataset training. Zhao *et al.* [133] introduced gradient-based graph instruction selection (G2IS), a novel method that utilizes principal component analysis (PCA) on validation set gradients to identify core knowledge representations and employs a gradient walk algorithm to select data points. Evaluation showed that G2IS can reduce 99% training cost while keeping model performance.

Some other works also use gradients as guidance for the data selection. For example, Zhao *et al.* [131] proposed collaborative learning under exemplar scoring (CLUES) to compute the trace of accumulated gradient inner products as a measure of influence on training dynamics. Wang *et al.* [102]proposed a two-stage method, ClusterUCB. ClusterUCB first groups the training data pool into clusters based on the intuition that samples with similar gradients will have similar training influences, and then frames the inter-cluster selection as a multi-armed bandit problem, using a modified upper confidence bound algorithm with a cold start to allocate a constrained computing budget. The authors showed that ClusterUCB can reduce 80% of training budgets. Deng *et al.* [21] introduced gradient trajectory pursuit (GTP), which formulates the selection as an L0-norm regularized objective, efficiently finding a data subset whose gradient trajectory matches a target trajectory within a compressed subspace. Surprisingly, experimental results showcase that using 0.5% of the data can reach the full-dataset performance based on GTP.

Moreover, Kang *et al.* [46] argue that existing data selection methods, which simply match the target distribution, are suboptimal for models that have already been pre-trained. They proposed a gradient-based method, gradients of optimal transport for data selection (GOT-D), which identifies unlabeled data for an initial pre-fine-tuning stage. The key idea is to use the gradients of the optimal transport distance to select samples that most effectively nudge the pre-training distribution closer to the target distribution. Yang *et al.* [111] proposed refined contribution measurement with in-context learning, a technique that quantifies the fine-grained contribution of individual samples by measuring the reduction in perplexity they induce on a diverse assessment set when used as in-context demonstrations. To ensure scalability, a lightweight selection module is trained on these computed scores, enabling the curation of large datasets with strictly linear inference complexity.

Loss is a commonly used indicator for data selection. Nikdan *et al.* [79] proposed to assign weights to training samples based on their second-order influence on a target distribution's loss. The method calculates model-specific weights for both gradient descent and Adam optimizers, and uses a landmark-based approximation with novel Jacobian-vector product embeddings to efficiently scale the influence computation.

More recently, Zhang *et al.* [124] proposed GRAPE, a supervised fine-tuning framework grounded in the hypothesis that data closely aligned with a pre-trained base model distribution yields superior performance compared to generic high-quality synthetic data. GRAPE gathers multiple candidate responses for a set of instructions and selects the specific response that maximizes the conditional probability under the target model itself, thereby ensuring the training data fits the model's intrinsic knowledge distribution without requiring external reward models or gradient computations.

Most works designed different techniques with promising results, but lacked theoretical guarantees. To address this, Zhou *et al.* [138] proposed HYPERINF, a novel method that combines the strong convergence guarantees of Schulz's iterative algorithm with a generalized Fisher Information Matrix as a low-rank Hessian approximation. HYPERINF reduces computational and memory costs on LoRA-tuned models to be constant and independent of the rank. After that, Yang *et al.* [112] introduced CRest, a scalable framework with rigorous theoretical guarantees. CRest addresses the challenges of training non-convex models by dynamically modeling the loss landscape as a series of quadratic functions and extracting core-sets for each quadratic sub-region to approximate the full gradient. To ensure convergence for stochastic gradient methods like SGD, it iteratively selects multiple mini-batch core-sets from larger random subsets of data, yielding nearly-unbiased gradients with small variance. Additionally, CRest excludes learned examples from the selection pipeline to further enhance efficiency.

Li *et al.* [53] introduced the IFD metric, calculated using the self-model briefly trained on a small data subset. This model then quantifies the difficulty of an instruction by comparing the model's response generation loss with and without the instruction's context. Experiments demonstrated that selecting only a small fraction (5-10%) of data with high IFD scores leads to models that match or outperform those trained on the full datasets.

5.2.2.2 ***Uncertainty***

Osband *et al.* [80] proposed employing an epinet, a small additional network, to form an epistemic neural network (ENN) capable of estimating its own uncertainty, specifically distinguishing epistemic from aleatoric uncertainty. Then the data selection is based on the predicted uncertainty score. Zhang *et al.* [128] proposed constraint LoRA with dynamic active learning (STAR) to utilize the uncertainty score for data selection. STAR introduces a dynamic uncertainty measurement that linearly interpolates between the uncertainty of the frozen base model and the fine-tuned model over time, while simultaneously employing a hybrid regularization strategy consisting of Monte-Carlo dropout and L2 norm weight decay to prevent model over-confidence. Zhao *et al.* [134] proposed UFO-RL, a framework that uses a novel uncertainty estimation technique (average Log-Softmax) to efficiently identify and select data instances of intermediate difficulty where the model exhibits high learning potential. Moreover, Wang *et al.* [101] analyzed chain-of-thought (CoT) reasoning through the lens of token entropy, discovering that a small part of tokens exhibit high entropy, which act as critical decision points that steer the reasoning path. Based on the finding, the authors proposed a framework that restricts policy gradient updates exclusively to the top 20% high-entropy tokens, achieving the token-selection purpose.

5.2.2.3 ***Data Diversity***

Shi *et al.* [92] proposed a unified framework combining diversity-aware self-Signal dilution (DASD) and convergent adaptive weighted sampling (CAWS). The DASD method calibrated the model by clustering reasoning paths and quantifying diversity as the sum of semantically distinct members to dilute the confidence scores of redundant answers for robust training targets. This calibrated foundation enables CAWS, an adaptive inference algorithm that aggregates confidence-weighted votes to halt generation once a candidate achieves dominance and stability over a sliding window, to dynamically terminate inference. Consequently, the approach achieves over 70% reduction in inference costs while maintaining high accuracy, such as reaching 92.6% on GSM8K with fewer than 16 samples on average compared to the standard baseline of 64. Agarwal *et al.* [2] proposed data-efficient language model instruction fine-tuning (DELIFT). Inspired by in-context learning that quantifies the informational gain of a sample by measuring its effectiveness in improving model predictions for other samples when used as a prompt. DELIFT selects diverse and non-redundant subsets tailored to the specific fine-tuning stage using submodular functions.

5.2.2.4 ***Intrinsic Data Quality***

Wang *et al.* [100] proposes Data Whisperer, an efficient, training-free data selection method for task-specific LLM fine-tuning. Data Whisperer leverages the model's own predictive capabilities to score the utility of each training sample with few-shot in-context learning. To mitigate ICL's order sensitivity, it introduces a context-aware weighting mechanism that refines these utility scores based on their attention scores. Data Whisperer can lead the model to achieve superior performance on the full dataset using just 10% of the data.

5.2.2.5 ***Composite and Multi-Objective Strategies***

Similar to the surrogate-based methods, several surrogate-free techniques integrate multiple selection criteria for the data selection to guarantee more data properties.

Heuristics. Tu *et al.* [97] introduced ResoFilter, a novel filtering approach that shifts focus from simple data selection to quality refinement by analyzing the interaction between data and model parameters during fine-tuning. The selection process utilizes a characteristic intensity metric, which is calculated by performing a one-shot fine-tuning for each data sample and measuring the resulting weight shifts in the feed-forward networks of the model's deeper layers. Samples with lower weight differences are prioritized as they represent high-consistency data that aligns better with the model's pre-existing knowledge. Empirical evaluations showed that ResoFilter achieved performance comparable to full-dataset training on mathematical reasoning tasks using only 50% of the available data.

Training Dynamics and Influence & Heuristics. Wang *et al.* [99] proposed a framework that evaluates data by generating adversarial instruction samples through character, word, and sentence-level attacks. The authors introduced the adversarial instruction-following difficulty (AIFD) score to measure the difficulty of aligning responses to adversarial inputs, and the adversarial instruction output embedding consistency (AIOEC) metric to assess embedding stability when model responses are unreliable. Mei *et al.* [74] proposed a group-level optimal transport-guided coreset selection (GORACS). A proxy optimization objective that leverages optimal transport distance and gradient norms to efficiently bound the generalization gap without full model evaluation has been introduced. This objective is optimized via a two-stage initialization-then-refinement algorithm (ITRA) that first selects a greedy initial subset and then refines it through iterative sample exchanges guided by efficient marginal improvement estimation. Evaluation demonstrated that GORACS can reduce fine-tuning time cost by approximately 80%.

Training Dynamics and Influence & Data Diversity. Jia *et al.* [44] designed ITERIT, an iterative data selection framework for instruction tuning. The method employs a complexity score, which measures IFD via perplexity ratios and is updated per training epoch, alongside a diversity score that utilizes TF-IDF on response n-grams. Specifically, the TF-IDF calculation determines the importance of n-grams by balancing their frequency within a single response against their rarity across the selected dataset, dynamically down-weighting features that have already been covered to ensure a diverse selection. By iteratively updating these scores and greedily selecting a subset that maximizes the combined metric, the approach facilitates a collaborative loop where the model guides data selection. For the same purpose, Pan *et al.* [81] proposed G-DIG. It first utilizes an influence function that quantifies the gradient-based similarity between training samples and the seed dataset to select training examples that have a positive influence on the model. It then enhances diversity by clustering the gradients of these selected examples and resampling uniformly from the clusters to maximize the variety of influences on the model.

Considering both gradients and diversity, Zhang *et al.* [126] proposed task-agnostic gradient clustered coreset selection (TAGCOS). TAGCOS uses sample gradients as data representations, applies K-means clustering to these gradients to manage data diversity, and then uses an optimal matching pursuit (OMP) algorithm within each cluster to select a subset that best approximates the gradient of the entire cluster. Evaluation demonstrated that selecting only 5% of the data via TAGCOS achieved performance close to that of training on the full dataset.

Uncertainty & Data Diversity. Arabelly *et al.* [5] introduced weighted task diversity, a label-efficient sampling strategy that leverages existing task categorizations and model uncertainty to optimize data selection for instruction tuning. The selection process allocates a minimum budget to every task to ensure broad coverage, while distributing the remaining budget based on the inverse of the base model's confidence. Here, confidence is measured as the average product of token-level probabilities. Finally, it prioritizes tasks where the pre-trained model exhibits higher uncertainty.

Uncertainty & Heuristics. Yang *et al.* [115] introduced P3, an adaptive framework for LLM fine-tuning that dynamically optimizes data usage by tailoring training content to the model's evolving competence. To achieve this, the authors introduced a policy-driven difficulty metric that evaluates sample complexity based on the model's real-time generation probabilities, which drives a self-paced learning mechanism to progressively expose the model to more challenging data. This selection process is further refined using a determinantal point process to ensure high diversity within the chosen subsets, thereby enhancing generalization.

Training Dynamics and Influence & Uncertainty & Intrinsic Data Quality & Heuristics. Ding *et al.* [22] proposed decomposed difficulty-based data selection (3DS), a model-centric data selection framework designed to align data selection with the target model's internal knowledge distribution. 3DS consists of two stages: first, a prompt-driven selection filters data based on the model's own quality assessment, followed by a difficulty-based selection using three metrics: instruction understanding difficulty, which measures the model's comprehension of the input instruction; response confidence difficulty, which gauges the uncertainty of the model's generated response; and response correctness difficulty, which evaluates the difficulty of producing the ground truth answer. These metrics are calibrated by an attention-based importance weighting mechanism to focus on semantically significant tokens. Results showed that selecting just 5,000 samples from a pool of 1.9 million achieves an average accuracy improvement of 2.97% over baselines.

5.2.3 Benchmark or Empirical Study

Among our reviewed works in the fine-tuning phase, in addition to those proposed novel techniques for data selection, 9 works built benchmarks or conducted comprehensive studies to analyze existing methods and support future research in this domain.

Some works focus on the scenario of active learning for language models, where active learning is a long-term studied problem in the machine learning community. Hu *et al.* [40] established the first benchmark for active code learning to evaluate sample-efficient training strategies systematically. The key findings reveal that, first, for clustering-based methods, model output vectors are surprisingly the most effective features for data selection, outperforming code tokens and embeddings in

62.5% of cases. Second, while active learning proves highly effective for classification tasks, it is notably ineffective for the non-classification task of code summarization, resulting in a performance gap of over 29.64%. Jacobs *et al.* [42] presented a comprehensive empirical study evaluating the effectiveness of uncertainty-based active learning (AL) strategies. The study compares three primary query functions: *Variation Ratio*, *Predictive Entropy*, and *Bayesian Active Learning by Disagreement (BALD)*. Furthermore, it investigates the impact of query-pool size and introduces three heuristics (*RET*, *RECS*, and *SUD*) designed to mitigate the known issues of redundancy and outlier selection in AL. The paper concludes that while standard uncertainty-based AL is an effective strategy for sample-efficient fine-tuning of BERT, its performance gains are more modest compared to its application on older NLP models or in computer vision.

Yin *et al.* [116] investigated data selection for LLM fine-tuning under strict compute constraints. It formalizes this practical challenge as a cost-aware optimization problem, modeling the trade-off between the initial compute spent on selecting data and the subsequent training efficiency gains. The core problem addressed is that while sophisticated data selection methods can improve performance per training step, their own computational cost might make them suboptimal when total compute is limited. The study finds that computationally expensive techniques like perplexity-based and gradient-based selection are often not compute-optimal. Interestingly, Ivison *et al.* [41] utilized a systematic study to reveal that many existing methods fail to improve or even degrade in performance when applied to large-scale data pools of up to 5.8 million samples.

Xia *et al.* [107] re-evaluated self-scoring data selection methods for supervised fine-tuning at a large scale. It found that, based on replicating methods on two million-scale datasets (Openhermes2.5 and WildChat-1M), nearly all existing self-scoring methods fail to significantly outperform simple random selection. The analysis concludes that data diversity is more critical than data quality for SFT at scale. Finally, the study proposes that filtering data by token length offers a stable and efficient method that yields superior results. Similarly, Shen *et al.* [91] highlighted that, due to the superficial nature of SFT, the key is to select demonstrations that best reflect human-like interactions. It introduces a simple yet effective heuristic of selecting instances with the longest responses, hypothesizing that detailed responses are more helpful.

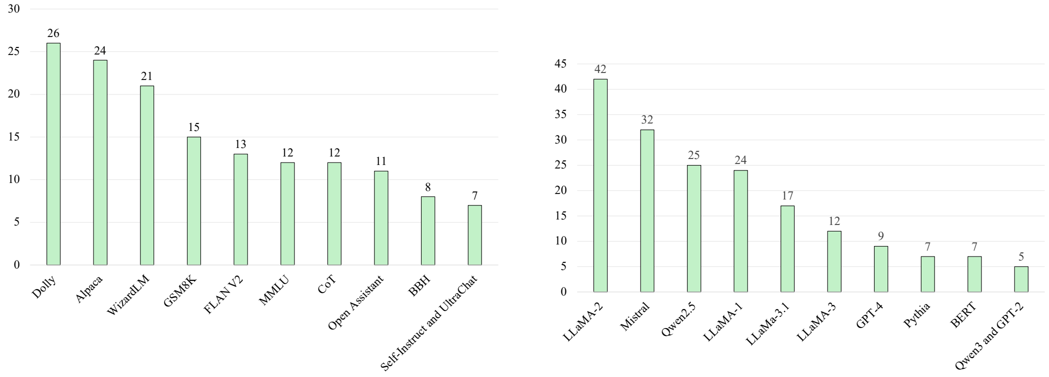
Moreover, Liu *et al.* [66] addressed the absence of a unified framework and the difficulty of systematically comparing diverse experimental settings in the field of data selection for fine-tuning LLMs. They proposed a three-stage taxonomy covering feature extraction, criteria design, and selector evaluation, augmented by a unified methodology that assesses efficiency and feasibility through ratio and ranking metrics. The study demonstrated that methods employing targeted quality measurements, such as IFD, yield higher efficiency in model performance improvement but often at the expense of practical feasibility compared to simpler approaches.

From a security perspective, He *et al.* [36] investigated the phenomenon of safety degradation during benign fine-tuning to identify benign data that shares optimization characteristics with harmful content. The study reveals a critical security flaw whereby fine-tuning safety-aligned LLMs on seemingly harmless datasets, such as lists or mathematical problems, can unintentionally weaken built-in guardrails and enable jailbreaking. Experimental results demonstrate that fine-tuning on a subset of just 100 selected benign data points significantly increases the Attack Success Rate (ASR) to over 70%, compared to less than 20% when using randomly selected data.

Unlike other works that focus on benchmarking selection algorithms or analyzing scale, Jha *et al.* [43] investigated the impact of dataset composition across conflicting evaluation paradigms. They proposed a data composition strategy to resolve the tension between traditional perplexity-based benchmarks and model-based evaluations. Observing that different datasets optimize for different evaluation metrics, such as style and factual knowledge, the authors fine-tune MPT models by combining the high-quality LIMA dataset with random subsets from larger instruction corpora.

The study finds that a mixture of approximately 2,000 to 6,000 samples is sufficient to achieve performance parity with full datasets of over 50,000 examples on traditional benchmarks.

5.2.4 *Data and Model Analysis*



(a) Distribution of top-10 datasets utilized in the surveyed papers.

(b) Distribution of top-10 models utilized in the surveyed papers.

Fig. 3. Statistical overview of datasets and models in the collected papers.

Figure 3a depicts the top-10 frequently used datasets in these 98 fine-tuning data selection works. We can see that the Instruction Tuning task is the most commonly studied, where Dolly, Alpaca, and WizardLM appeared 26, 24, and 21 times, respectively. There is also a notable emphasis on reasoning and logic capabilities, evidenced by the frequent inclusion of GSM8K (15), MMLU (12), and Chain-of-Thought (CoT) (12) datasets. This distribution suggests that while basic instruction alignment remains the primary testbed, researchers are increasingly validating their methods against complex mathematical and multi-step reasoning tasks. Figure 3 shows the top-10 frequently used models. It is clear that the LLaMA family is the most widely used base model. The potential reason is that LLaMA is one of the earliest open-source LLMs, establishing it as the *de facto* standard for fine-tuning research. Beyond this dominant family, the Mistral model (32) and the Qwen series (specifically Qwen2.5 with 25 citations) have established themselves as significant contenders.

6 Discussion

6.1 Pre-training vs. Fine-tuning

Research focus. Among our surveyed 116 works, only 18 papers (15.5%) focus on data selection in the pre-training stage, while the remaining ones target the fine-tuning stage. This disparity indicates that researchers prioritize data selection for instruction tuning and alignment over foundational pre-training. The potential reasons are: 1) *Difficulty.* The pre-training process needs to consider multiple tasks, e.g., QA, programming, and other reasoning tasks. This multi-objective purpose significantly increases the challenge of designing new data selection methods. 2) *Computational cost.* Compared to the fine-tuning datasets, pre-training datasets are much larger in scale, requiring more computation costs. However, the majority of research originates from academia, which lacks enough computational resources. Interestingly, 11 (61.1%) of 18 pre-training works were authored in collaboration with major technology companies or large-scale research institutes. This phenomenon supports the hypothesis that the pre-training stage remains a high-resource domain.

Table 7. Statistics of different data selection studies.

Surrogate Type	Selection Criteria Categorization	#Works
Surrogate-Free	Training Dynamics and Influence	12
	Uncertainty	4
	Data Diversity	3
	Intrinsic Data Quality	1
	Composite and Multi-Objective Strategies	9
Surrogate-Based	Training Dynamics and Influence	13
	Uncertainty	2
	Data Diversity	8
	Intrinsic Data Quality	13
	Composite and Multi-Objective Strategies	42

Technique focus. In the pre-training phase, most of the works (17 out of 18) utilize external models or proxy metrics to estimate data utility. This imbalance stems from the massive scale of pre-training corpora. Specifically, the computational overhead of using the model-in-training to evaluate its own data is typically high. Surrogate-based techniques mitigate this by employing efficient, often static, proxies like n-gram overlaps or smaller, frozen language models to filter noise and redundancy. In contrast, the fine-tuning phase exhibits a much higher adoption of surrogate-free methods (28 works). We attribute this trend to the reduced scale and increased specialization of fine-tuning datasets, which typically range from 50,000 to 5 million samples. At this scale, assessing data utility directly with the model-in-training becomes computationally friendly. By using the target model to determine which samples are most informative for instruction-following or alignment, researchers avoid the biases in external scoring models. However, the fine-tuning phase employs surrogate-free techniques much more frequently than the pre-training phase. This trend reflects a technical focus on precision and direct task alignment.

Dataset. Datasets used in pre-training (as shown in Table 3) typically possess massive scale and diversity, primarily consisting of high-volume text corpora and general-purpose Q&A pairs. Researchers frequently use datasets such as C4, The Pile, RedPajama, and The Stack, which provide the multi-terabyte scale necessary for models to learn statistical regularities. More importantly, they cover various domains, including web content, scientific literature, and source code. In this stage, data selection emphasizes knowledge density and broad coverage. In contrast, the 98 fine-tuning papers focus heavily on specialized instruction-following datasets (Fig. 3a). The top three datasets, Dolly (26), Alpaca (24), and WizardLM (21), account for a substantial portion of the research focus. Unlike pre-training corpora, these datasets are relatively small in scale. They consist of (instruction, response) pairs designed to align the model's output with human intent. Furthermore, reasoning-heavy datasets like GSM8K and MMLU are prevalent in the fine-tuning phase. This indicates a strategic shift toward enhancing logical deduction and problem-solving capabilities after the initial knowledge acquisition phase.

Model. In pre-training data selection research, models span various architectural generations, from encoder-only models (e.g., BERT-base, BioBERT) to decoder-only scaling benchmarks (e.g., GPT-2, Pythia, LLaMA-1). This diversity indicates that pre-training research often investigates the fundamental properties of different architectures. It also examines the impact of varied pre-training objectives. In contrast, model usage in fine-tuning data selection is more concentrated. Researchers show a strong preference for state-of-the-art open-weight models. For example, LLaMA-2 (42) and Mistral (32) are the most frequently used.

6.2 Surrogate-based vs. Surrogate-free

Baseline consideration. We found that around 90% of analyzed surrogate-based methods include surrogate-free baselines in their evaluations. Surrogate-based methods introduce the additional computational cost of training or using an auxiliary model. Therefore, they must demonstrate performance gains that justify this overhead against computationally cheaper alternatives. However, some (65%) of the surrogate-based methods consider Random selection as the surrogate-free baseline, which is insufficient for the evaluation. Conversely, around 69% of surrogate-free methods compare against surrogate-based baselines. However, the selection of these baselines is highly specific. Rather than comparing to a broad range of surrogates, surrogate-free methods frequently target baselines, AlpaGasus, and DEITA. The goal is to demonstrate that intrinsic signals (e.g., loss) can achieve selection quality comparable to surrogate pipelines without the associated resource costs.

Selection criteria focus. Table 7 summarizes the distribution of selection criteria used by each paradigm. First, we observe a significant difference in the use of Intrinsic Data Quality metrics. Surrogate-based methods frequently use this metric (13 works), but not for surrogate-free methods (1 work). This trend aligns with the definition of intrinsic quality, which evaluates properties such as fluency, coherence, and correctness. Assessing these characteristics typically requires a powerful auxiliary model. This model must score textual quality independently of the current state of the target model. Moreover, surrogate-based methods (17%) rely less on Training Dynamics and Influence compared to surrogate-free methods (41%). The potential reason is that Training dynamics derive directly from the model's interaction with data during training, using indicators such as sample loss or gradient norms. However, surrogate-based methods often use proxy models to estimate data impact. The loss and gradient norms from surrogates cannot precisely reflect the training behavior of the target model. As a result, Training Dynamics and Influence are less used in a surrogate-based situation.

Furthermore, Composite and Multi-objective Strategies are markedly more prevalent in surrogate-based methods (42 works) than in their surrogate-free counterparts (9 works). This indicates that using a surrogate model facilitates the aggregation of distinct criteria, such as combining diversity with quality. This approach captures a more holistic notion of data value. Surrogate-based frameworks allow for the pre-computation of complex heuristics or diversity metrics (e.g., clustering in a fixed embedding space). This avoids the computational burden of re-evaluating the target model at every step. Conversely, the training loop of the target model constrains surrogate-free methods. Therefore, they are less likely to adopt complex multi-objective formulations that might disrupt training efficiency.

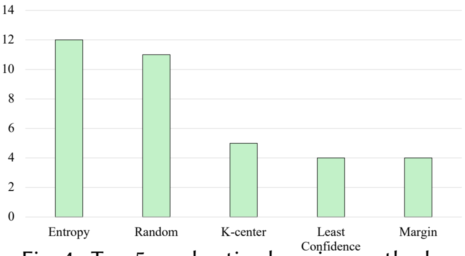


Fig. 4. Top-5 used active learning methods.

6.3 Pitfalls

6.3.1 *Purpose*

Among our surveyed works, 64 works clearly mentioned their purpose for using data-efficient methods. 11 studies aim to reduce labeling cost, seeking to minimize manual annotation by leveraging unlabeled or synthetic data. 53 works target to reduce computational cost.

Pitfall 1: No labeling cost targeted works reported the difficulty of labeling their studied datasets. Currently, multiple automatic data labeling tools [17, 48, 77] have been proposed with promising results. If the work focuses on reducing data labeling cost, it is necessary to highlight why these tools cannot precisely label the datasets. That is, justifying the trade-off between data labeling difficulty and training cost is important.

Pitfall 2: 52 works did not clarify their purpose of using data-efficient methods. This purpose highly affects the evaluation process of their proposed methods, as mentioned in Pitfall 1.

6.3.2 *Baseline Consideration*

Pitfall 3: 16% of works did not consider baselines from different types of methods.

*Pitfall 4:*Active learning is a representative sample-efficient model-training paradigm used in conventional machine learning. Multiple active learning methods have been proposed, such as Dropout-based active learning. These methods can be easily modified and applied to LLMs. However, we found that only 14 papers used active learning methods as baselines. Figure 4 shows the frequency of active learning methods used in these 14 works, where Entropy-based sampling is the predominant method.

6.3.3 ***Surrogate Cost***

Pitfall 5: Integrating surrogate models into data selection pipelines introduces high computational and storage costs. Despite the widespread adoption of this methodology, our review of 78 relevant studies indicates that researchers frequently employ surrogates. However, only 17 of 78 surrogate-based studies discussed the external cost caused by the surrogate model. This oversight is critical, as recent evaluations [107, 116] reveal that the prohibitive computational cost of sophisticated selection often renders it less computationally efficient than simple random baselines or training on larger raw datasets.

6.3.4 ***Measurement***

Pitfall 6: In addition to the pure capability (e.g., accuracy), other metrics, such as robustness and security of LLMs, are also important for evaluation. However, only 44 conducted robustness evaluations. Further analysis of the 44 papers reveals that the most commonly used robustness metrics include: model robustness (18 papers), data robustness (17 papers), cross-domain/task robustness (14 papers), hyperparameter robustness (6 papers), and instruction-following robustness (5 papers), as illustrated in Figure 5.

Model robustness focuses on performance across different model sizes and architectures, ensuring that a method maintains high performance when transferred from smaller surrogate models to larger, more complex ones. This helps verify that the data selection is not overfit to a specific model architecture. *Data robustness* examines performance across varying data distributions and quantities, assessing the model's stability (e.g., F1 score) when faced with duplicate data, varying proportions, or outliers. A robust method should consistently outperform baselines under different data conditions. *Cross-domain/task robustness* emphasizes the model's ability to generalize across different domains and tasks, ensuring performance is maintained not only within the original domain but also in near-domain and out-of-domain scenarios. *Hyperparameter robustness* assesses whether the method remains effective despite changes in hyperparameters, such as learning rates or model scaling, indicating broader applicability. *Instruction-following robustness* evaluates how well the model can adhere to instructions under noisy, complex, or sparse conditions, reflecting the reliability of the method in dynamic environments.

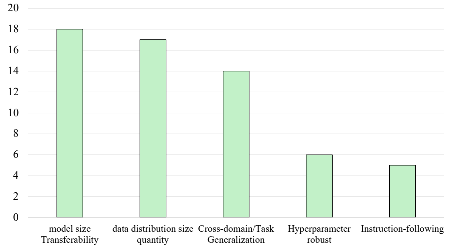


Fig. 5. Top-5 most frequently used robustness metrics.

6.4 Threats to Validity

Construct Threat. A primary threat lies in the potential miss of relevant studies due to the limitations of keyword-based retrieval. As the terminology for data efficiency in LLMs is not yet fully standardized, distinct communities may use varying nomenclatures (e.g., data selection versus data pruning). To mitigate this, we constructed a comprehensive set of search strings combining model-related terms with efficiency-related keywords. Furthermore, we applied bidirectional snowballing through forward and backward citation searches, identifying 13 additional high-value papers. We commit to maintaining our survey website to continuously gather future literature that is relevant to our research. **Internal Threat.** The classification of data selection methods into Surrogate-Based and Surrogate-Free paradigms, as well as their sub-categorizations, involves a degree of subjective judgment. Ambiguity may arise when a method exhibits hybrid characteristics, potentially leading to classification errors. To mitigate this threat, we established a unified problem definition and strict inclusion and exclusion criteria prior to data extraction. The paper selection followed a multi-stage process, screening titles, venues, and abstracts before reviewing full texts to ensure alignment with our scope. Furthermore, each author reviewed the papers involving classification to ensure the reliability of the classification.

7 Opportunities

As the model parameter and data scale increase, data-efficient techniques become increasingly important. There are multiple opportunities to facilitate this domain.

7.1 New Scenarios

Existing data selection research mainly focuses on a single LLM and QA tasks, other trending scenarios should be considered. First, LLMs increasingly function as agents integrated with external tools and long-horizon trajectories. Zhang *et al.* [130]introduced guideline effectiveness to quantify the utility of data for agentic tasks. However, scalable selection methods for complex, multi-step reasoning tasks remain under-explored. Future work should evaluate the quality of interaction trajectories rather than isolated response pairs. Second, specific domains like code generation require distinct evaluation strategies. The efficacy of data selection strategies varies significantly across different data types. Hu *et al.* [40]showed that strategies effective for classification tasks often fail in generative contexts. This demands domain-specific metrics.

7.2 Distribution Aware

Existing works largely treat data selection as a static matching problem, typically aiming to align the training subset solely with a target scenario. We argue that this single-perspective approach is limited. It ignores the complex distributional interplay between the original pre-trained model, the target task, and the evolving training dynamics. Future frameworks must bridge source and target distributions. Approaches that only match the target distribution often fail to optimize the transition from the base model. Kang *et al.* [46] showed that such methods are suboptimal because they ignore the pre-training distribution. Similarly, Zhang *et al.* [124] showed that data aligned with the base model's probability distribution outperforms generic high-quality data. Therefore, future works must consider source and target distributions simultaneously. They should select samples that maximize alignment while minimizing distributional shock to the base model.

Furthermore, methods must anticipate dynamic distributional shifts. Current methods often overlook the dynamic nature of optimal distributions. Data optimal at the onset of fine-tuning may become redundant as the model's internal distribution shifts during training. Yu *et al.* [120] showed that static selection fails to capture the shifting preferences of LLMs. Yang *et al.* [115] corroborated this by introducing pace-adaptive learning, in which data difficulty evolves in real time. We argue that a distribution-aware method must move beyond static curation. It requires a continuous, feedback-driven process that anticipates shifts. This prevents the model from overfitting to a fixed target distribution.

7.3 Multi-Objective Selection

First, only a few works consider integrating multiple evaluation metrics in their method design. In addition to the pure capability of LLMs, other metrics, such as security and privacy, should also be considered during the model optimization. That is, designing multi-objective selection methods is a promising future research direction. Moreover, the construction of current data selection strategies predominantly relies on isolated metrics, such as perplexity and embedding distance. Sometimes, they combine them through rudimentary heuristics. This ad-hoc approach fails to address the complexity of LLM training. Robust data selection requires a formal multi-objective optimization problem and explicitly navigating the Pareto frontiers between conflicting criteria.For example, Fan *et al.* [26] identified that aggressive filtering for high-quality, domain-specific text tends to homogenize the dataset. It collapses feature diversity and degrades general capabilities. Conversely, prioritizing pure diversity often introduces noise. Therefore, future frameworks must advance beyond linear score combinations and identify optimal trade-offs that maximize coverage without sacrificing the signal-to-noise ratio.

Current research rarely treats the computational cost of selection as a primary objective. Yin *et al.* [116] formally modeled the trade-off between selection cost and training efficiency. They showed that sophisticated methods, such as those based on perplexity or gradients, often prove suboptimal compared to simple lexical filters when the total budget is constrained. Future methods should incorporate selection cost as a negative objective. This prioritizes lightweight, scalable filters over expensive model-based scoring for massive datasets.

7.4 Theoretical Guarantee

Although existing methods often yield practical gains, they lack rigorous theoretical justification. Yang *et al.* [112] and Zhou *et al.* [138] pioneered this direction. They established convergence guarantees for specific selection algorithms, such as CRest and HYPERINF. Following these works, future research is encouraged to build rigorous theoretical foundations for data subset selection. It needs to establish certifiable bounds on permissible data reduction without exceeding defined error thresholds. Establishing these guarantees is challenging because the optimization landscape of LLMs is non-convex. Standard CoreSet theories typically assume convex loss functions and do not hold here. Therefore, focusing on approximating data point influence in non-convex settings is a promising research direction.

8 Conclusion

We comprehensively surveyed 116 works on data-efficient LLM training. This survey unified techniques across pre-training and fine-tuning phases into surrogate-based and surrogate-free approaches. We analyzed how these methods identify high-utility data to maximize model performance while minimizing computational costs. Furthermore, we outlined critical research opportunities for theoretical metrics and standardized evaluation. This survey aims to assist researchers in navigating the trade-offs between data quantity and utility for LLM construction. The entire project is available at https://github.com/Data-centric-LLM/Data-Efficient-LLM-Construction.

References

[1] Ishika Agarwal and Dilek Hakkani-Tür. 2025. Data Valuation using Neural Networks for Efficient Instruction Fine-Tuning. *arXiv preprint arXiv:2502.09969* (2025).

[2] Ishika Agarwal, Krishnateja Killamsetty, Lucian Popa, and Marina Danilevksy. 2024. Delift: Data efficient language model instruction fine tuning. *arXiv preprint arXiv:2411.04425* (2024).

[3] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A Survey on Data Selection for Language Models. arXiv:2402.16827 [cs.CL] https://arxiv.org/abs/2402.16827

[4] Moses Ananta, Muhammad Farid Adilazuarda, Zayd Muhammad Kawakibi Zuhri, Ayu Purwarianti, and Alham Fikri Aji. 2025. QLESS: A Quantized Approach for Data Valuation and Selection in Large Language Model Fine-Tuning. *arXiv preprint arXiv:2502.01703* (2025).

[5] Abhinav Arabelly, Jagrut Nemade, Robert D Nowak, and Jifan Zhang. 2025. Improving Task Diversity in Label Efficient Supervised Finetuning of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 21569–21581.

[6] Guirong Bai, Shizhu He, Kang Liu, Jun Zhao, and Zaiqing Nie. 2020. Pre-trained language model based active learning for sentence matching. *arXiv preprint arXiv:2010.05522* (2020).

[7] Md Abul Bashar and Richi Nayak. 2021. Active learning for effectively fine-tuning transfer learning to downstream task. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 2 (2021), 1–24.

[8] David Brandfonbrener, Hanlin Zhang, Andreas Kirsch, Jonathan Richard Schwarz, and Sham Kakade. 2024. Color-filter: Conditional loss reduction filtering for targeted language model pre-training. *Advances in Neural Information Processing Systems* 37 (2024), 97618–97649.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[10] Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. 2024. Data diversity matters for robust instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 3411–3425.

[11] Shuo Cai, Su Lu, Qi Zhou, Kejing Yang, Zhijie Sang, Congkai Xie, and Hongxia Yang. 2025. InfiAlign: A Scalable and Sample-Efficient Framework for Aligning LLMs to Enhance Reasoning Capabilities. *arXiv preprint arXiv:2508.05496* (2025).

[12] Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2023. Instruction mining: Instruction data selection for tuning large language models. *arXiv preprint arXiv:2307.06290* (2023).

[13] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

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

[15] Yicheng Chen, Yining Li, Kai Hu, Ma Zerun, HaochenYe HaochenYe, and Kai Chen. 2025. Mig: Automatic data selection for instruction tuning by maximizing information gain in semantic space. In *Findings of the Association for Computational Linguistics: ACL 2025*. 9902–9915.

[16] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. 2019. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829* (2019).

[17] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. (2023).

[18] Qirun Dai, Dylan Zhang, Jiaqi W Ma, and Hao Peng. 2025. Improving Influence-based Instruction Tuning Data Selection for Balanced Learning of Diverse Capabilities. *arXiv preprint arXiv:2501.12147* (2025).

[19] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS* 35 (2022), 16344–16359.

[20] Devleena Das and Vivek Khetan. 2024. DEFT-UCS: Data Efficient Fine-Tuning for Pre-Trained Language Models via Unsupervised Core-Set Selection for Text-Editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 20296–20312.

[21] Zhiwei Deng, Tao Li, and Yang Li. 2024. Influential Language Data Selection via Gradient Trajectory Pursuit. *arXiv preprint arXiv:2410.16710* (2024).

[22] Hongxin Ding, Yue Fang, Runchuan Zhu, Xinke Jiang, Jinyang Zhang, Yongxin Xu, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. 3DS: Decomposed Difficulty Data Selection's Case Study on LLM Medical Domain Adaptation. *arXiv preprint arXiv:2410.10901* (2024).

[23] Dai Do, Manh Nguyen, Svetha Venkatesh, and Hung Le. 2025. SPaRFT: Self-Paced Reinforcement Fine-Tuning for Large Language Models. *arXiv preprint arXiv:2508.05015* (2025).

[24] Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653* (2023).

[25] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[26] Ziqing Fan, Siyuan Du, Shengchao Hu, Pingjie Wang, Li Shen, Ya Zhang, Dacheng Tao, and Yanfeng Wang. 2025. Combatting dimensional collapse in LLM pre-training data via submodular file selection. In *The Thirteenth International Conference on Learning Representations*.

[27] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[28] Lan Feng, Fan Nie, Yuejiang Liu, and Alexandre Alahi. 2024. Tarot: Targeted data selection via optimal transport. *arXiv preprint arXiv:2412.00420* (2024).

[29] Yanjun Fu, Faisal Hamman, and Sanghamitra Dutta. 2025. T-SHIRT: Token-Selective Hierarchical Data Selection for Instruction Tuning. *arXiv preprint arXiv:2506.01317* (2025).

[30] Shuzheng Gao, Wenxin Mao, Cuiyun Gao, Li Li, Xing Hu, Xin Xia, and Michael R Lyu. 2024. Learning in the wild: Towards leveraging unlabeled data for effectively tuning pre-trained code models. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.

[31] Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Mahong Xia, Zhang Li, Boxing Chen, Hao Yang, et al. 2024. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 464–478.

[32] Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. 2024. Data selection via optimal control for language models. *arXiv preprint arXiv:2410.07064* (2024).

[33] Yuxian Gu, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Miniplm: Knowledge distillation for pre-training language models. *arXiv preprint arXiv:2410.17215* (2024).

[34] Wenya Guo, Zhengkun Zhang, Xumeng Liu, Ying Zhang, Ziyu Lu, Haoze Zhu, Xubo Liu, and Ruxue Yan. 2025. ProDS: Preference-oriented Data Selection for Instruction Tuning. *arXiv preprint arXiv:2505.12754* (2025).

[35] Bowei He, Lihao Yin, Hui-Ling Zhen, Xiaokun Zhang, Mingxuan Yuan, and Chen Ma. 2025. PASER: Post-Training Data Selection for Efficient Pruned Large Language Model Recovery. *arXiv preprint arXiv:2502.12594* (2025).

[36] Luxi He, Mengzhou Xia, and Peter Henderson. 2024. What is in your safe data? identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099* (2024).

[37] Xixiang He, Hao Yu, Qiyao Sun, Ao Cheng, Tailai Zhang, Cong Liu, and Shuxuan Guo. 2025. TACOS: Open Tagging and Comparative Scoring for Instruction Fine-Tuning Data Selection. *arXiv preprint arXiv:2507.03673* (2025).

[38] Yexiao He, Ziyao Wang, Zheyu Shen, Guoheng Sun, Yucong Dai, Yongkai Wu, Hongyi Wang, and Ang Li. 2024. Shed: Shapley-based automated dataset refinement for instruction fine-tuning. *Advances in Neural Information Processing Systems* 37 (2024), 99382–99403.

[39] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2024. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology* 33, 8 (2024), 1–79.

[40] Qiang Hu, Yuejun Guo, Xiaofei Xie, Maxime Cordy, Lei Ma, Mike Papadakis, and Yves Le Traon. 2024. Active code learning: Benchmarking sample-efficient training of code models. *IEEE Transactions on Software Engineering* 50, 5 (2024), 1080–1095.

[41] Hamish Ivison, Muru Zhang, Faeze Brahman, Pang Wei Koh, and Pradeep Dasigi. 2025. Large-Scale Data Selection for Instruction Tuning. *CoRR* (2025).

[42] Pieter Floris Jacobs, Gideon Maillette de Buy Wenniger, Marco Wiering, and Lambert Schomaker. 2021. Active learning for reducing labeling effort in text classification tasks. In *Benelux Conference on Artificial Intelligence*. Springer, 3–29.

[43] Aditi Jha, Sam Havens, Jeremy Dohmann, Alex Trott, and Jacob Portes. 2023. Limit: Less is more for instruction tuning across evaluation paradigms. *arXiv preprint arXiv:2311.13133* (2023).

[44] Qi Jia, Siyu Ren, Ziheng Qin, Fuzhao Xue, Jinjie Ni, and Yang You. 2024. Boosting LLM via Learning from Data Iteratively and Selectively. *arXiv preprint arXiv:2412.17365* (2024).

[45] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169* (2023).

[46] Feiyang Kang, Hoang Anh Just, Yifan Sun, Himanshu Jahagirdar, Yuanzhi Zhang, Rongxing Du, Anit Kumar Sahu, and Ruoxi Jia. 2024. Get more for less: Principled data selection for warming up fine-tuning in llms. *arXiv preprint arXiv:2405.02774* (2024).

[47] 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:2001.08361 [cs.LG] https://arxiv.org/abs/2001.08361

[48] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *ICLR*.

[49] Barbara Kitchenham, O Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology* 51, 1 (2009), 7–15.

[50] Jiazheng Li, Lu Yu, Qing Cui, Zhiqiang Zhang, Jun Zhou, Yanfang Ye, and Chuxu Zhang. 2025. MASS: Mathematical Data Selection via Skill Graphs for Pretraining Large Language Models. *arXiv preprint arXiv:2503.14917* (2025).

[51] Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. In *ACL 2024*.

[52] Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *arXiv preprint arXiv:2402.00530* (2024).

[53] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 7602–7635.

[54] Xiaomin Li, Mingye Gao, Zhiwei Zhang, Chang Yue, and Hong Hu. 2024. Rule-based data selection for large language models. *arXiv preprint arXiv:2410.04715* (2024).

[55] Xianming Li and Jing Li. 2024. Generative Deduplication For Socia Media Data Selection. *arXiv preprint arXiv:2401.05883* (2024).

[56] Yisen Li, Lingfeng Yang, Wenxuan Shen, Pan Zhou, Yao Wan, Weiwei Lin, and Dongping Chen. 2025. CrowdSelect: Synthetic Instruction Data Selection with Multi-LLM Wisdom. *arXiv preprint arXiv:2503.01836* (2025).

[57] Zhuang Li, Yuncheng Hua, Thuy Vu, Haolan Zhan, Lizhen Qu, and Gholamreza Haffari. 2025. SCAR: Data Selection via Style Consistency-Aware Response Ranking for Efficient Instruction-Tuning of Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12756–12790.

[58] Hao Liang, Linzhuang Sun, Jingxuan Wei, Xijie Huang, Linkun Sun, Bihui Yu, Conghui He, and Wentao Zhang. 2024. Synth-empathy: Towards high-quality synthetic empathy data. *arXiv preprint arXiv:2407.21669* (2024).

[59] Xu Liangyu, Xuemiao Zhang, Feiyu Duan, Sirui Wang, Rongxiang Weng, Jingang Wang, and Xunliang Cai. 2025. FIRE: Flexible Integration of Data Quality Ratings for Effective Pretraining. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 14532–14552.

[60] Xiaotian Lin, Yanlin Qi, Yizhang Zhu, Themis Palpanas, Chengliang Chai, Nan Tang, and Yuyu Luo. 2025. Lead: Iterative data selection for efficient llm instruction tuning. *arXiv preprint arXiv:2505.07437* (2025).

[61] Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2024. Data-efficient Fine-tuning for LLM-based Recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*. 365–374.

[62] Zhenqing Ling, Daoyuan Chen, Liuyi Yao, Qianli Shen, Yaliang Li, and Ying Shen. 2025. Diversity as a reward: Fine-tuning llms on a mixture of domain-undetermined data. *arXiv preprint arXiv:2502.04380* (2025).

[63] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434* (2024).

[64] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685* (2023).

[65] Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. 2024. Tsds: Data selection for task-specific model finetuning. *Advances in Neural Information Processing Systems* 37 (2024), 10117–10147.

[66] Ziche Liu, Rui Ke, Yajiao Liu, Feng Jiang, and Haizhou Li. 2025. Take the essence and discard the dross: A rethinking on data selection for fine-tuning large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6595–6611.

[67] Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. [n. d.]. # InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

[68] Junyu Luo, Bohan Wu, Xiao Luo, Zhiping Xiao, Yiqiao Jin, Rong-Cheng Tu, Nan Yin, Yifan Wang, Jingyang Yuan, Wei Ju, and Ming Zhang. 2025. A Survey on Efficient Large Language Model Training: From Data-centric Perspectives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 30904–30920. https://doi.org/10.18653/v1/2025.acl-long.1493

[69] Weijie Lv, Xuan Xia, and Sheng-Jun Huang. 2024. Codeact: Code adaptive compute-efficient tuning framework for code llms. *arXiv preprint arXiv:2408.02193* (2024).

[70] Michael R. Lyu, Baishakhi Ray, Abhik Roychoudhury, Shin Hwei Tan, and Patanamon Thongtanunam. 2024. Automatic Programming: Large Language Models and Beyond. arXiv:2405.02213 [cs.SE] https://arxiv.org/abs/2405.02213

[71] Da Ma, Gonghu Shang, Zhi Chen, Libo Qin, Yijie Luo, Lei Pan, Shuai Fan, Lu Chen, and Kai Yu. 2025. Neuronal Activation States as Sample Embeddings for Data Selection in Task-Specific Instruction Tuning. *arXiv e-prints* (2025), arXiv–2503.

[72] Da Ma, Gonghu Shang, Zhi Chen, Libo Qin, Yijie Luo, Lei Pan, Shuai Fan, Lu Chen, and Kai Yu. 2025. Task-Specific Data Selection for Instruction Tuning via Monosemantic Neuronal Activations. *arXiv preprint arXiv:2503.15573* (2025).

[73] 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* (2023).

[74] Tiehua Mei, Hengrui Chen, Peng Yu, Jiaqing Liang, and Deqing Yang. 2025. GORACS: Group-level Optimal Transport-guided Coreset Selection for LLM-based Recommender Systems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 2126–2137.

[75] Paramita Mirza, Lucas Weber, and Fabian Küch. 2025. Stratified Selective Sampling for Instruction Tuning with Dedicated Scoring Strategy. *arXiv preprint arXiv:2505.22157* (2025).

[76] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. 2020. Coresets for data-efficient training of machine learning models. In *ICML*. PMLR, 6950–6960.

[77] Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *ACL 2024*. 12585–12611.

[78] Xinzhe Ni, Yeyun Gong, Zhibin Gou, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. Exploring the mystery of influential data for mathematical reasoning. *arXiv preprint arXiv:2404.01067* (2024).

[79] Mahdi Nikdan, Vincent Cohen-Addad, Dan Alistarh, and Vahab Mirrokni. 2025. Efficient Data Selection at Scale via Influence Distillation. *arXiv preprint arXiv:2505.19051* (2025).

[80] Ian Osband, Seyed Mohammad Asghari, Benjamin Van Roy, Nat McAleese, John Aslanides, and Geoffrey Irving. 2022. Fine-tuning language models via epistemic neural networks. *arXiv preprint arXiv:2211.01568* (2022).

[81] Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. 2024. G-dig: Towards gradient-based diverse and high-quality instruction data selection for machine translation. *arXiv preprint arXiv:2405.12915* (2024).

[82] Jinlong Pang, Na Di, Zhaowei Zhu, Jiaheng Wei, Hao Cheng, Chen Qian, and Yang Liu. 2025. Token cleaning: Fine-grained data selection for llm supervised fine-tuning. *arXiv preprint arXiv:2502.01968* (2025).

[83] Jinlong Pang, Jiaheng Wei, Ankit Parag Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, Yujia Bao, and Wei Wei. 2024. Improving data efficiency via curating llm-driven rating systems. *arXiv preprint arXiv:2410.10877* (2024).

[84] Lv Qingsong, Yangning Li, Zihua Lan, Zishan Xu, Jiwei Tang, Yinghui Li, Wenhao Jiang, Hai-Tao Zheng, and Philip S Yu. 2025. RAISE: Reinforenced Adaptive Instruction Selection For Large Language Models. *arXiv preprint arXiv:2504.07282* (2025).

[85] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–16.

[86] HSVNS Kowndinya Renduchintala, Krishnateja Killamsetty, Sumit Bhatia, Milan Aggarwal, Ganesh Ramakrishnan, Rishabh Iyer, and Balaji Krishnamurthy. 2023. INGENIOUS: using informative data subsets for efficient pre-training of language models. *arXiv preprint arXiv:2305.06677* (2023).

[87] Dongyu Ru, Jiangtao Feng, Lin Qiu, Hao Zhou, Mingxuan Wang, Weinan Zhang, Yong Yu, and Lei Li. 2020. Active Sentence Learning by Adversarial Uncertainty Sampling in Discrete Space. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 4908–4917.

[88] Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668* (2024).

[89] Burr Settles. 2009. Active learning literature survey. (2009).

[90] Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. 2024. Seal: Safety-enhanced aligned llm fine-tuning via bilevel data selection. *arXiv preprint arXiv:2410.07471* (2024).

[91] Ming Shen. 2024. Rethinking data selection for supervised fine-tuning. *arXiv preprint arXiv:2402.06094* (2024).

[92] Zhenning Shi, Yijia Zhu, Yi Xie, Junhan Shi, Guorui Xie, Haotian Zhang, Yong Jiang, Congcong Miao, and Qing Li. 2025. Reasoning under Uncertainty: Efficient LLM Inference via Unsupervised Confidence Dilution and Convergent Adaptive Sampling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 32192–32206.

[93] Kashun Shum, Yuzhen Huang, Hongjian Zou, Qi Ding, Yixuan Liao, Xiaoxin Chen, Qian Liu, and Junxian He. 2025. Predictive data selection: The data that predicts is the data that teaches. *arXiv preprint arXiv:2503.00808* (2025).

[94] Shuzheng Si, Haozhe Zhao, Gang Chen, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Kaikai An, Kangyang Luo, Chen Qian, Fanchao Qi, et al. 2025. Aligning Large Language Models to Follow Instructions and Hallucinate Less via Effective Data Filtering. *CoRR* (2025).

[95] Jielin Song, Siyu Liu, Bin Zhu, and Yanghui Rao. 2024. Iterselecttune: An iterative training framework for efficient instruction-tuning data selection. *arXiv preprint arXiv:2410.13464* (2024).

[96] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159* (2018).

[97] Zeao Tu, Xiangdi Meng, Yu He, Zihan Yao, Tianyu Qi, Jun Liu, and Ming Li. 2025. ResoFilter: Fine-grained Synthetic Data Filtering for Large Language Models through Data-Parameter Resonance Analysis. In *NAACL 2025*.

[98] Peiqi Wang, Yikang Shen, Zhen Guo, Matthew Stallone, Yoon Kim, Polina Golland, and Rameswar Panda. 2024. Diversity measurement and subset selection for instruction tuning datasets. *arXiv preprint arXiv:2402.02318* (2024).

[99] Qiang Wang, Dawei Feng, Xu Zhang, Ao Shen, Yang Xu, Bo Ding, and Huaimin Wang. 2025. Pay More Attention to the Robustness of Prompt for Instruction Data Mining. *arXiv preprint arXiv:2503.24028* (2025).

[100] Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, et al. 2025. Data whisperer: Efficient data selection for task-specific llm fine-tuning via few-shot in-context learning. *arXiv preprint arXiv:2505.12212* (2025).

[101] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939* (2025).

[102] Zige Wang, Qi Zhu, Fei Mi, Minghui Xu, Ruochun Jin, and Wenjing Yang. 2025. ClusterUCB: Efficient Gradient-Based Data Selection for Targeted Fine-Tuning of LLMs. *arXiv preprint arXiv:2506.10288* (2025).

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

[104] Minghao Wu, Thuy-Trang Vu, Lizhen Qu, and Gholamreza Haffari. 2024. The best of both worlds: Bridging quality and diversity in data selection with bipartite graph. *arXiv preprint arXiv:2410.12458* (2024).

[105] Yang Wu, Huayi Zhang, Yizheng Jiao, Lin Ma, Xiaozhong Liu, Jinhong Yu, Dongyu Zhang, Dezhi Yu, and Wei Xu. 2024. Rose: A reward-oriented data selection framework for llm task-specific instruction tuning. *arXiv preprint arXiv:2412.00631* (2024).

[106] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*. 54104–54132.

[107] Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. 2024. Rethinking data selection at scale: Random selection is almost all you need. *arXiv preprint arXiv:2410.09335* (2024).

[108] Xinnuo Xu, Minyoung Kim, Royson Lee, Brais Martinez, and Timothy Hospedales. 2024. A Bayesian approach to data point selection. *Advances in Neural Information Processing Systems* 37 (2024), 38171–38198.

[109] Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Xiaojun Wu, Honghao Liu, Hui Xiong, and Jian Guo. 2025. Select2Reason: Efficient Instruction-Tuning Data Selection for Long-CoT Reasoning. *arXiv preprint arXiv:2505.17266* (2025).

[110] Xianjun Yang, Shaoliang Nie, Lijuan Liu, Suchin Gururangan, Ujjwal Karn, Rui Hou, Madian Khabsa, and Yuning Mao. 2025. Diversity-driven data selection for language model tuning through sparse autoencoder. *arXiv preprint arXiv:2502.14050* (2025).

[111] Yixin Yang, Qingxiu Dong, Linli Yao, Fangwei Zhu, and Zhifang Sui. 2025. RICo: Refined In-Context Contribution for Automatic Instruction-Tuning Data Selection. *arXiv preprint arXiv:2505.05327* (2025).

[112] Yu Yang, Hao Kang, and Baharan Mirzasoleiman. 2023. Towards sustainable learning: Coresets for data-efficient deep learning. In *International Conference on Machine Learning*. PMLR, 39314–39330.

[113] Yu Yang, Siddhartha Mishra, Jeffrey Chiang, and Baharan Mirzasoleiman. 2024. Smalltolarge (s2l): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. *Advances in Neural Information Processing Systems* 37 (2024), 83465–83496.

[114] Yu Yang, Aaditya K Singh, Mostafa Elhoushi, Anas Mahmoud, Kushal Tirumala, Fabian Gloeckle, Baptiste Rozière, Carole-Jean Wu, Ari S Morcos, and Newsha Ardalani. 2023. Decoding data quality via synthetic corruptions: Embedding-guided pruning of code data. *arXiv preprint arXiv:2312.02418* (2023).

[115] Yingxuan Yang, Huayi Wang, Muning Wen, Xiaoyun Mo, Qiuying Peng, Jun Wang, and Weinan Zhang. 2024. P3: A policy-driven, pace-adaptive, and diversity-promoted framework for data pruning in llm training. *arXiv preprint arXiv:2408.05541* (2024).

[116] Junjie Oscar Yin and Alexander M Rush. 2024. Compute-constrained data selection. *arXiv preprint arXiv:2410.16208* (2024).

[117] Simon Yu, Liangyu Chen, Sara Ahmadian, and Marzieh Fadaee. 2024. Diversify and conquer: Diversity-centric data selection with iterative refinement. *arXiv preprint arXiv:2409.11378* (2024).

[118] Yu Yu, Shahram Khadivi, and Jia Xu. 2022. Can data diversity enhance learning generalization?. In *Proceedings of the 29th international conference on computational linguistics*. 4933–4945.

[119] Yue Yu, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen, and Chao Zhang. 2023. Cold-start data selection for better few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*. 2499–2521.

[120] Zichun Yu, Spandan Das, and Chenyan Xiong. 2024. Mates: Model-aware data selection for efficient pretraining with data influence models. *Advances in Neural Information Processing Systems* 37 (2024), 108735–108759.

[121] 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:2303.10158 [cs.LG] https://arxiv.org/abs/2303.10158

[122] Chi Zhang, Huaping Zhong, Hongtao Li, Chengliang Chai, Jiawei Hong, Yuhao Deng, Jiacheng Wang, Tian Tan, Yizhou Yan, Jiantao Qiu, et al. 2025. Not All Documents Are What You Need for Extracting Instruction Tuning Data. *arXiv preprint arXiv:2505.12250* (2025).

[123] Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Jiantao Qiu, Lei Cao, Ju Fan, et al. 2024. Harnessing diversity for important data selection in pretraining large language models. *arXiv preprint arXiv:2409.16986* (2024).

[124] Dylan Zhang, Qirun Dai, and Hao Peng. 2025. The best instruction-tuning data are those that fit. *arXiv preprint arXiv:2502.04194* (2025).

[125] Jifan Zhang, Ziyue Luo, Jia Liu, Ness Shroff, and Robert Nowak. 2024. GPT-4o as the Gold Standard: A Scalable and General Purpose Approach to Filter Language Model Pretraining Data. *arXiv preprint arXiv:2410.02755* (2024).

[126] Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan, and Tong Zhang. 2025. Tagcos: Task-agnostic gradient clustered coreset selection for instruction tuning data. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 4671–4686.

[127] Jia Zhang, Chen-Xi Zhang, Yao Liu, Yi-Xuan Jin, Xiao-Wen Yang, Bo Zheng, Yi Liu, and Lan-Zhe Guo. 2025. D3: Diversity, Difficulty, and Dependability-Aware Data Selection for Sample-Efficient LLM Instruction Tuning. *arXiv preprint arXiv:2503.11441* (2025).

[128] Linhai Zhang, Jialong Wu, Deyu Zhou, and Guoqiang Xu. 2024. Star: Constraint lora with dynamic active learning for data-efficient fine-tuning of large language models. *arXiv preprint arXiv:2403.01165* (2024).

[129] Xiaoyu Zhang, Juan Zhai, Shiqing Ma, Chao Shen, Tianlin Li, Weipeng Jiang, and Yang Liu. 2024. Speculative Coreset Selection for Task-Specific Fine-tuning. *CoRR* (2024).

[130] Yunxiao Zhang, Guanming Xiong, Haochen Li, and Wen Zhao. 2025. EDGE: Efficient Data Selection for LLM Agents via Guideline Effectiveness. *arXiv preprint arXiv:2502.12494* (2025).

[131] Wanru Zhao, Hongxiang Fan, Shell Xu Hu, Wangchunshu Zhou, Bofan Chen, and Nicholas D Lane. 2025. CLUES: Collaborative High-Quality Data Selection for LLMs via Training Dynamics. *arXiv preprint arXiv:2507.03004* (2025).

[132] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* 1, 2 (2023).

[133] Yang Zhao, Li Du, Xiao Ding, Yangou Ouyang, Hepeng Wang, Kai Xiong, Jinglong Gao, Zhouhao Sun, Dongliang Xu, Qing Yang, et al. 2025. Beyond similarity: A gradient-based graph method for instruction tuning data selection. In *ACL 2025*.

[134] Yang Zhao, Kai Xiong, Xiao Ding, Li Du, Zhouhao Sun, Jiannan Guan, Wenbin Zhang, Bin Liu, Dong Hu, Bing Qin, et al. 2025. UFO-RL: Uncertainty-Focused Optimization for Efficient Reinforcement Learning Data Selection. *arXiv preprint arXiv:2505.12457* (2025).

[135] Qihuang Zhong, Liang Ding, Fei Liao, Juhua Liu, Bo Du, and Dacheng Tao. 2025. Resolving Knowledge Conflicts in Domain-specific Data Selection: A Case Study on Medical Instruction-tuning. *arXiv preprint arXiv:2505.21958* (2025).

[136] Haotian Zhou, Tingkai Liu, Qianli Ma, Yufeng Zhang, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2025. Davir: Data selection via implicit reward for large language models. In *ACL 2025*.

[137] Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. 2024. A Survey on Generative AI and LLM for Video Generation, Understanding, and Streaming. arXiv:2404.16038 [cs.CV] https://arxiv.org/abs/2404.16038

[138] Xinyu Zhou, Simin Fan, and Martin Jaggi. 2024. HyperINF: Unleashing the HyperPower of the Schulz's Method for Data Influence Estimation. *arXiv preprint arXiv:2410.05090* (2024).

[139] Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, et al. 2024. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *arXiv preprint arXiv:2406.11931* (2024).

[140] Xiaoxuan Zhu, Zhouhong Gu, Baiqian Wu, Suhang Zheng, Tao Wang, Tianyu Li, Hongwei Feng, and Yanghua Xiao. 2025. ToReMi: Topic-Aware Data Reweighting for Dynamic Pre-Training Data Selection. *arXiv preprint arXiv:2504.00695* (2025).

[141] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Large Language Models for Information Retrieval: A Survey. *ACM Transactions on Information Systems* 44, 1 (Nov. 2025), 1–54. https://doi.org/10.1145/3748304

[142] Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. 2025. Meta-rater: A multi-dimensional data selection method for pre-training language models. In *ACL 2025*. 10856–10896.

[143] Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. 2024. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877* (2024).