Data Value Density Enhancement for Large Language Model Training: A Comprehensive Survey

Yiliu Sun[1*], Yanchao Lu[1*], Jiaxi Cao[1*], Linyang Li[2], Wei Liu[1†], Chen Gong[1†]

[1]Shanghai Jiao Tong University, Shanghai, P. R. China

[2]Shanghai Artificial Intelligence Laboratory, Shanghai, P. R. China

[*]Equal contribution

[†]Corresponding author. E-mails: {weiliucv, chen.gong}@sjtu.edu.cn

Contributing authors: sunyiliu@pjlab.org.cn, cse.luyanchao@gmail.com, caojiaxi@sjtu.edu.cn, lilinyang@pjlab.org.cn

14 April 2026

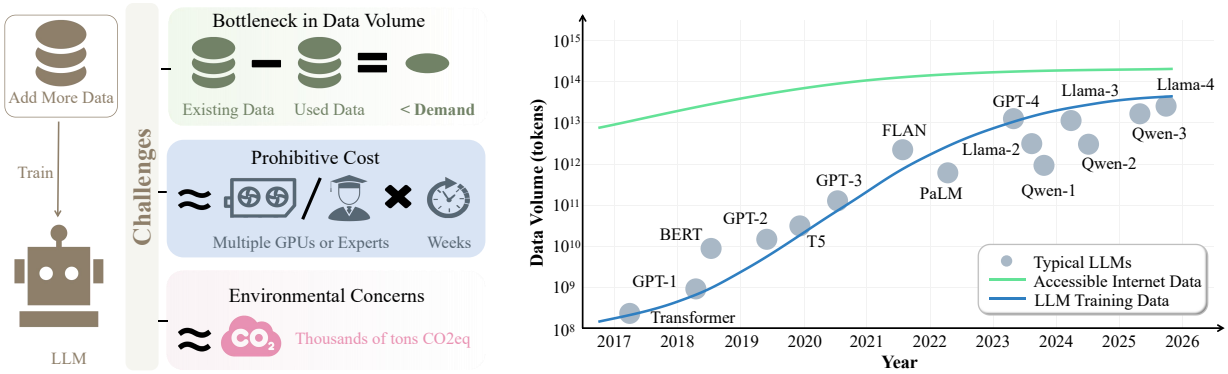
Abstract

Driven by the scaling laws of Large Language Models (LLMs), increasing the volume of training data has been one of the primary strategies for enhancing the capabilities of LLMs over the past few years. However, as the accessible yet unused internet data becomes increasingly scarce, improving model performance by merely increasing data volume is no longer sustainable. In response, researchers have shifted their focus toward improving the performance of LLMs with limited training data and have proposed a variety of modern methods. Despite the progress, this research field lacks a clear definition and a comprehensive review, resulting in unclear research objectives and a fragmented landscape of methodologies. To bridge this gap, this paper provides a comprehensive survey aimed at offering a thorough understanding of the methodologies in this field. Specifically, we introduce the concept of "Data Value Density (DVD) enhancement" for LLM training, which serves as a unified perspective to summarize the main progresses of this research field. Based on the formal definition of DVD enhancement, we categorize existing methods into five primary directions. By employing this unified taxonomy, we provide an extensive review of state-of-the-art DVD enhancement techniques and highlight their strengths and weaknesses. Additionally, we review the representative datasets used for training and evaluating DVD enhancement models. Finally, we identify four major challenges faced by the current methods of DVD enhancement and present promising research directions for future advancements in this field.

Keywords—Large Language Models, Data Value Density Enhancement, Data Bottleneck, Data Selection, Data Evolution

1 Introduction

The rapid advance of Large Language Models (LLMs) has been largely driven by the vast data used during their training pipeline, such as pre-training and post-training stages. Such data provides sufficient and diverse knowledge for LLMs to learn, enabling them to achieve promising performance across a wide range of tasks such as math, coding, question answering, and so on. In the past few years, inspired by the scaling laws of LLMs (Kaplan et al 2020; Hoffmann et al 2022) that reveal the relationship between the training data scale and the LLM performance, expanding the volume of training data has become one of the primary strategies for enhancing the capabilities of LLMs in the pre-training stage (Brown et al 2020; Touvron et al 2023; Achiam et al 2023; Team et al 2023). Following this success, the same paradigm has also been extended



(a) (b)

Figure 1: (a) Three critical challenges faced by continuously scaling up LLM training data, *i.e.*, data bottleneck, prohibitive cost, and environment concern. (b) The volumes of accessible internet data and training data used to train popular LLMs in the past years. The estimates of accessible internet data are derived from the study conducted by Villalobos *et al.* (Villalobos et al 2022), whereas the reported training data scales of popular LLMs are sourced from their respective technical reports.

to the post-training stage, where ever-larger instruction-following or preference datasets are constructed to further improve the performance of pre-trained LLMs (Hendrycks et al 2021b; Guo et al 2025a; Guha et al 2025; Gu et al 2024b). However, as illustrated in Fig. 1a, this expansion-oriented paradigm faces several critical challenges:

- **Data Bottleneck.** As shown in Fig. 1b, due to the severe imbalance between the growth rate of accessible internet data (< 10% per year) and that of the data scale used for LLM training (exponential growth per year), the accessible yet unused data is becoming increasingly scarce (Villalobos et al 2022). Moreover, constructing new datasets is costly, as it often requires extensive data collection, careful curation, and substantial human or computational effort to ensure reliability and task relevance. As a result, relying on expanding training data volume is no longer a sustainable strategy for improving LLM performance.

- **Prohibitive Cost.** Increasing the volume of training data often results in a proportional rise in training cost. For instance, the pre-training of Llama3.1-405B (Grattafiori et al 2024) requires 39.3 million GPU hours of computation on H100-80GB type hardware. The post-training of DeepSeek-R1-Zero (Guo et al 2025a) requires 101 kilo GPU hours of computation on H800-80GB type hardware. Such cost is usually out of reach for most research institutions and organizations.

- **Environment Concern.** Training on massive data requires many GPUs to run continuously for weeks or even months to process massive tokens. This process consumes massive energy, leading to substantial carbon emissions that conflict with global sustainability goals.

Given the above limitations, how to further improve the performance of LLMs with limited training data has become an urgent problem. To address this, many works (Muennighoff et al 2025; Ye et al 2025; Ma et al 2022; Xu et al 2023a; Ma et al 2025; Deb et al 2025; Lin et al 2024b) have been proposed to achieve a better trade-off between training effectiveness and efficiency from a data-centric perspective, which employ various strategies such as data selection (Miranda et al 2023; Yang et al 2024; Kung et al 2023), data scheduling (Chen et al 2023b; Wang et al 2024a; Xia et al 2024a), and data generation (Ye et al 2025; Zhou et al 2024b; Wang et al 2023b). This emerging line of research shifts the focus from expanding data volume to maximizing the training effect of limited data, enabling LLMs to achieve better performance with the same or even smaller training data scale.

1.1 Motivation and Contribution

The research area that aims to maximize the training effect of limited data from a data-centric perspective has become increasingly important for both academia and industry. There are three major reasons behind this trend:

- **Data.** Collecting large amounts of data can be challenging, particularly in specialized domains such as medicine, psychology, and law. In these data-scarce domains, exploiting the training value of existing data is more meaningful than simply expanding data scale, which is well aligned with the core objective of this research area.
- **Training.** The computational resources consumed by training on large-scale data are extremely substantial, which are unaffordable for most research institutions and organizations. Therefore, as a paradigm that can significantly reduce training costs, this area naturally attracts widespread attention.
- **Application.** To maximize the training effect, the data processed by methods in this area is typically of high quality and contains little noise. Training on such data can mitigate the hallucinations of LLMs, making them reliable in scenarios that require high precision, such as medical diagnosis and financial analysis.

To our best knowledge, there indeed exist some relevant surveys. For instance, Wang *et al.* (Wang et al 2024b) and Albalak *et al.* (Albalak et al 2024) focus on two specific strategies of "data augmentation and synthesis" and "data selection", respectively. Luo *et al.* (Luo et al 2025a) and Mo *et al.* (Mo et al 2025) respectively review this research area in two specific stages of LLM training, namely the post-training stage and the mid-training stage. Other surveys (Minaee et al 2024; Shengyu et al 2023) mention this area only as a minor component within their broader taxonomies without providing dedicated reviews. Since existing surveys have only provided partial or coarse-grained overviews, this research area still lacks a formal definition and a systematic review, leading to unclear research objectives and a fragmented landscape of methodologies. To fill this gap, we contribute to this research area as follows:

- We introduce the concept of **Data Value Density (DVD) Enhancement**, which serves as a unified perspective to characterize this emerging research area. Its definition is not only general enough to encompass the existing methods, but also specific enough to clarify what the fundamental goals of this area are and how to achieve them.
- We systematically interpret and organize existing methods of this research area, covering multiple stages of LLM training and various strategies. Specifically, we trace the bibliographies of papers presented at top-tier conferences and journals (*e.g.*, ICLR, ICML, NeurIPS, AAAI, ACL, CVPR, ICCV, JMLR, and IEEE TPAMI), and categorize existing methods of DVD enhancement in **a unified taxonomy**. For each category, we introduce the specific problem definition and the representative methods.
- We introduce the existing representative datasets employed in DVD enhancement for LLM training, covering various tasks, problem types, and data scales. Moreover, we highlight the data characteristics of each task to provide prior knowledge for designing effective DVD enhancement strategies in the corresponding task.
- We identify the current challenges and spark several promising research directions in the area of DVD enhancement, addressing theory, implementation, and application perspectives.

1.2 Organization

The remainder of this survey is organized as follows. Section 2 introduces the concept of DVD enhancement, presenting the fundamental definition of DVD and the target of DVD enhancement. Based on the definition, Section 3 reviews existing methods in the area of DVD enhancement and categorizes them in a unified taxonomy. Section 4 reviews the representative training and evaluation datasets employed in this research area, highlighting the data characteristics of different tasks. Section 5 introduces several closely related research topics, discussing their connections and distinctions with DVD enhancement. Section 6 outlines current challenges and promising future research directions. Finally, Section 7 concludes the survey and summarizes the key insights.

2 Basic Notions

In this section, we introduce the fundamental concepts necessary for understanding DVD enhancement. Firstly, we provide the formal definition of DVD in the context of LLM training, followed by a discussion of its practical implementations under different DVD enhancement strategies. Then, we formulate the optimization objective of DVD enhancement for LLMs, based on which our unified taxonomy is derived.

2.1 Data Value Density

Let $\mathcal{D}$ be a training dataset and $\mathcal{C}$ be a training context composed of multiple factors that provide the information necessary for evaluating data value, such as the characteristics of trained LLMs, the properties of target tasks, and so on. The DVD of $\mathcal{D}$ under $\mathcal{C}$ (*i.e.*, $f(\mathcal{D} \mid \mathcal{C})$) is defined as:

$$f(\mathcal{D} \mid \mathcal{C}) = \frac{V(\mathcal{D} \mid \mathcal{C})}{\mu(\mathcal{D})}, \tag{2.1}$$

where $V(\mathcal{D} \mid \mathcal{C})$ denotes a data value function that measures how much $\mathcal{D}$ contributes to model performance under $\mathcal{C}$, and $\mu(\mathcal{D})$ denotes a dataset scale function that quantifies the size of $\mathcal{D}$. The definition in Eq. (2.1) is general enough to encompass various DVD enhancement strategies, as it abstracts the concept of DVD without constraining how $V(\mathcal{D} \mid \mathcal{C})$ or $\mu(\mathcal{D})$ is measured. In practice, both $V(\mathcal{D} \mid \mathcal{C})$ and $\mu(\mathcal{D})$ are calculated differently across methods of DVD enhancement for LLM training. We discuss these variations in detail in the following sections.

2.1.1 Training Context $\mathcal{C}$

The evaluation of DVD is inherently context-dependent. The same data may contribute differently to the performance improvement of LLMs under distinct model characteristics, training objectives, and task settings. We denote the collection of such required information by $\mathcal{C}$. In practice, $\mathcal{C}$ may include the following factors:

- **Trained LLM $\mathcal{M}$.** It describes the properties of the employed LLM (*e.g.*, its capability distribution and training loss), which provide signals for evaluating how the dataset $\mathcal{D}$ contributes to model performance improvement, such as capability enhancement or training loss reduction.
- **Target Task $\mathcal{T}$.** It describes the properties of the target task, enabling the assessment of how relevant the data is to the task and how much task-specific knowledge it contains.
- **Information Pool $\mathcal{P}$.** It specifies the set of information that $\mathcal{D}$ is expected to contain. It is often included as a contextual factor in data distillation and data synthesis, where the goal is to make $\mathcal{D}$ cover as much information in $\mathcal{P}$ as possible.
- **Data-Intrinsic Knowledge $\mathcal{K}$.** It refers to the properties inherent to the data itself that are independent of the trained LLM and the target task. These properties capture the stable characteristics of the data, such as the presence of long reasoning chains and factual correctness, and provide signals for assessing the training value of the data.

Therefore, the training context $\mathcal{C}$ of one DVD enhancement method can be formally expressed as:

$$\mathcal{C} \subseteq \{\mathcal{M}, \mathcal{T}, \mathcal{P}, \mathcal{K}\}. \tag{2.2}$$

2.1.2 Data Value Function $V(\mathcal{D} \mid \mathcal{C})$

It calculates the contribution of the dataset $\mathcal{D}$ to the performance improvement of LLM under the training context $\mathcal{C}$. Its concrete form depends on the optimization objective and the training context $\mathcal{C}$ of the specific DVD enhancement method. For instance, the data value function for duplicate data removal strategies is generally defined as:

$$V(\mathcal{D} \mid \mathcal{C} = \emptyset) = \sum_{d \in \mathcal{D}} v(d) = \sum_{d \in \mathcal{D}} \left(1 - \max_{d' \in \mathcal{D} \setminus \{d\}} \operatorname{sim}\left(d, d'\right)\right), \tag{2.3}$$

where d and d' denote two different data points within the dataset $\mathcal{D}$, $\text{sim}(\cdot,\cdot) \in [0,1]$ denotes a semantic similarity function (*e.g.*, cosine similarity in an embedding space), and $v(\cdot)$ calculates the value of one data point. It is worth noting that the specific value range of $v(\cdot)$ varies depending on the target scenarios and method categories. Differently, the data value function for data selection strategies under different domains is generally defined as:

$$V(\mathcal{D} \mid \mathcal{C} = \{\mathcal{M}, \mathcal{T}\}) = \sum_{d \in \mathcal{D}} v(d \mid \mathcal{M}, \mathcal{T}) = \sum_{d \in \mathcal{D}} \left(\text{Score}\left(d, \text{Gap}(\mathcal{M}, \mathcal{T})\right)\right), \tag{2.4}$$

where $\text{Gap}(\mathcal{M}, \mathcal{T})$ denotes the gap between the current performance of the trained LLM and the performance required by the target task, and $\text{Score}(\cdot,\cdot)$ measures the contribution of one data point to bridge this gap. The detailed forms of $V(\mathcal{D} \mid \mathcal{C})$ in different DVD enhancement strategies are introduced in Section 3.

2.1.3 Dataset Scale Function $\mu(\mathcal{D})$

It quantifies the amount of data utilized during model training. Its formulation aligns with the granularity of the data value evaluation. For instance, if $V(\mathcal{D} \mid \mathcal{C})$ is computed at the instance level (*i.e.*, assigning a value to each data point), $\mu(\mathcal{D})$ is also measured at the instance level. If $V(\mathcal{D} \mid \mathcal{C})$ is computed at the token level (*i.e.*, assigning a value to each token within a sentence), $\mu(\mathcal{D})$ is correspondingly computed at the token level.

2.2 DVD Enhancement for LLM Training

Based on the definition of DVD, the objective of DVD enhancement for LLM training is to construct a new dataset $\mathcal{D}_{new}$ whose DVD exceeds that of the original dataset $\mathcal{D}_{ori}$ under the training context $\mathcal{C}$.

Objective of DVD Enhancement for LLM Training

Construct a new training dataset $\mathcal{D}_{new}$ such that

$$\Delta f = f(\mathcal{D}_{new} \mid \mathcal{C}) - f(\mathcal{D}_{ori} \mid \mathcal{C}) > 0,$$

where Δf denotes the increase in DVD.

For $\mathcal{D}_{new}$, it can be either a subset or a reconstructed version of $\mathcal{D}_{ori}$. Notably, in the framework of DVD enhancement for LLM training, the ordering of data points is treated as an important part of a dataset. In other words, two datasets containing the same set of data points but arranged in different orders are regarded as non-equivalent, since the data ordering may influence the DVD of the dataset.

3 Taxonomy

According to the definition of DVD in Eq. (2.1), we can see the ultimate goal for improving DVD is to increase the value of $f(\mathcal{D} \mid \mathcal{C})$. Therefore, based on how to increase $f(\mathcal{D} \mid \mathcal{C})$, we categorize the existing methods of DVD enhancement into five distinct categories (see Fig. 2). For each category, we present a detailed description and review representative methods that instantiate the corresponding design principles. Fig. 3 presents the overall taxonomy in a tree-structured format, providing a concise yet comprehensive overview of the five categories and their respective methods.

3.1 $V(\mathcal{D} \mid \mathcal{C})$ Increases with $\mu(\mathcal{D})$ Remains Unchanged

The methods in this category (Kim and Lee 2024; Li et al 2025d; Liu et al 2025d; Li et al 2025c; Wang et al 2025g; Liu et al 2024b; Ma et al 2025) aim to achieve better training effectiveness under a fixed training data budget. Based on their underlying design principles, these methods can be further divided into three subcategories, namely **data scheduling**, **data mixing**, and **augmentation via generation**. Among them, data scheduling focuses on exploiting the latent utility of training data points to generate greater training benefits. Data mixing explores the interactions among different data types and seeks optimal mixture ratios

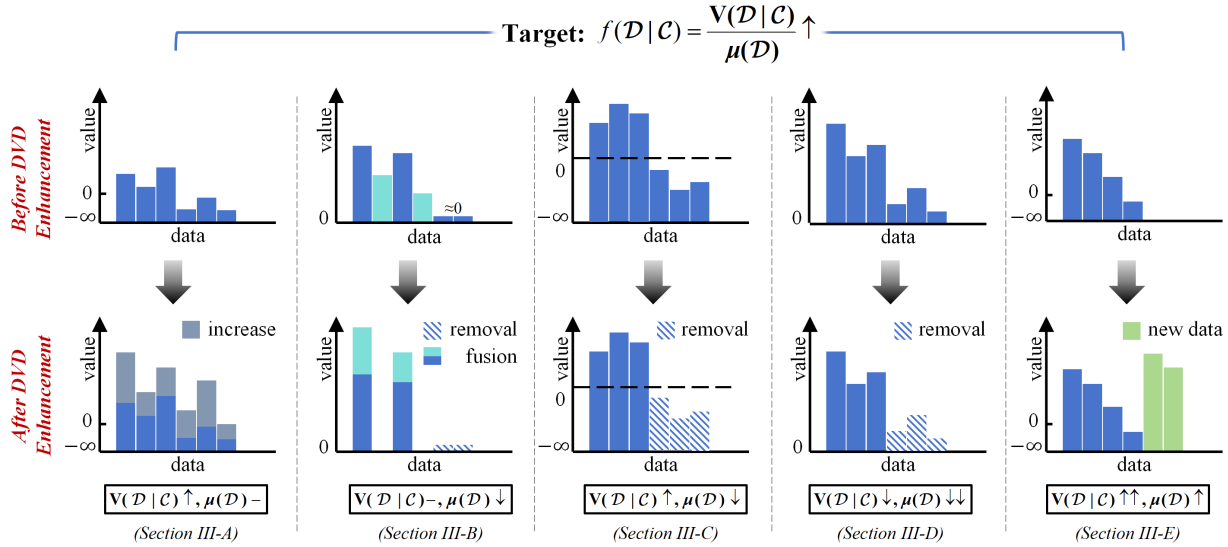


Figure 2: The definition of DVD naturally classifies the existing methods of DVD enhancement for LLM training into five categories.

to improve training effectiveness without increasing data scale. Augmentation via generation enriches the informational content of data points to provide additional knowledge and supervision signals.

3.1.1 Data Scheduling

This strategy is grounded in a key observation: the value of a data point is tied to the state of the model at the time the data point is used for training. For example, in the early stages when the capabilities of the model are still limited, simple data points typically provide high value, as they can be learned quickly and help the model establish a basic understanding of the target task. In contrast, in the later training stages when the model has gained sufficient capability, such easy data points offer diminishing returns. Since the model characteristics evolve throughout the training process, the training value of a data point varies across different training steps, which is formally presented as:

$$v\left(d \mid C=\left\{\mathcal{M}_{t_1}\right\}\right) \neq v\left(d \mid C=\left\{\mathcal{M}_{t_2}\right\}\right), t_1 \neq t_2, \tag{3.1}$$

where $\mathcal{M}_t$ denotes the model characteristics at the t-th training step. From this perspective, the objective of data scheduling can be described as:

The Objective of Data Scheduling

At each training step t, construct $\mathcal{D}^* \subseteq \mathcal{D}$ such that

$$\forall \mathcal{D}' \subseteq \mathcal{D} \text{ and } \mu\left(\mathcal{D}'\right)=\mu\left(\mathcal{D}^*\right),\ V\left(\mathcal{D}^* \mid C=\{\mathcal{M}_t\}\right) \geq V\left(\mathcal{D}' \mid C=\{\mathcal{M}_t\}\right),$$

where $V\left(\mathcal{D}' \mid C=\{\mathcal{M}_t\}\right)=\sum_{d \in \mathcal{D}'} v\left(d \mid C=\{\mathcal{M}_t\}\right)$.

A substantial portion of existing research (Zhang et al 2018; Liu et al 2018; Hu et al 2024; Lai et al 2024; Xu et al 2020; Platanios et al 2019; Wang et al 2021; Elgaar and Amiri 2026) draws inspiration from the curriculum learning paradigm (Bengio et al 2009), which emulates the human learning process by presenting training data in an easy-to-hard sequence. This paradigm enables LLMs to first acquire fundamental capabilities from simple data points before being exposed to more challenging ones. Consequently, its success hinges on how data difficulty is defined and quantified.

One line of research (Nagatsuka et al 2023; Chang et al 2021; Lee et al 2024a; Sun et al 2024a; Qiu et al 2025) quantifies data difficulty based on the intrinsic attributes of the data itself, such as pre-assigned difficulty

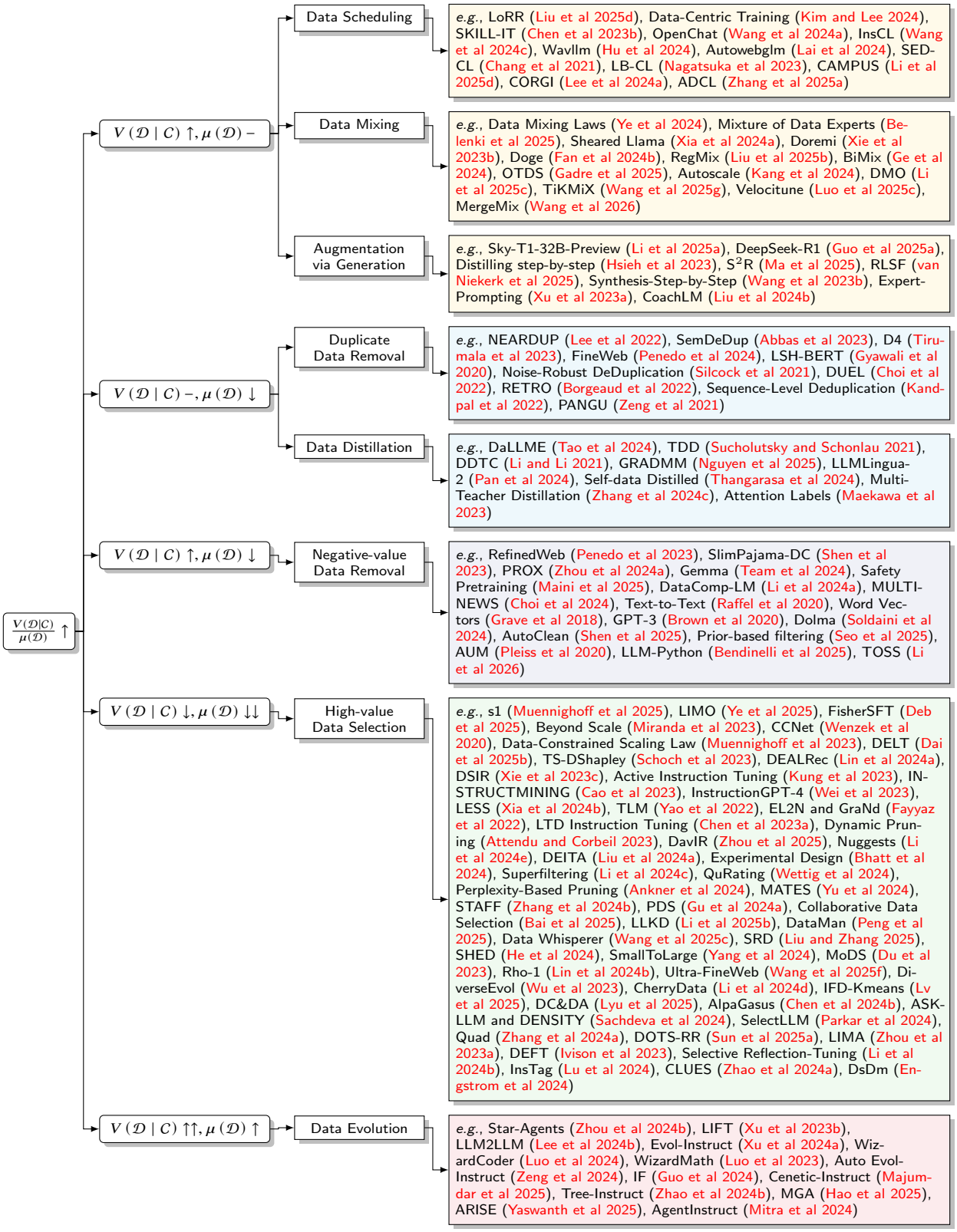


Figure 3: A taxonomy of DVD enhancement for LLM training based on the definition of DVD. Here, "↑", "–", and "↓" mean that a value increases, remains unchanged, and decreases, respectively. Additionally, "↑↑" and "↓↓" denote a more significant increase and decrease, respectively.

labels, expert assessments, or observable data characteristics. For pre-assigned labels, several reasoning datasets (*e.g.*, MATH (Hendrycks et al 2021b), TAL-SCQ5K (Liu et al 2025c), and DeepMath-103K (He et al 2025)) provide explicit difficulty annotations during dataset construction, which can be directly used to organize curriculum sequences. For expert assessments, recent methods typically prompt strong LLMs to evaluate the difficulty of the data (Gao et al 2025; Luo et al 2025b). In addition, difficulty can be inferred from task-specific data characteristics (Jung and Jung 2025; Jia et al 2025). For example, in long chain-of-thought reasoning tasks, the length of the reasoning trace can serve as a proxy for problem complexity, whereas in multi-step deduction tasks, difficulty may be approximated by the number of required reasoning steps.

Alternatively, some methods (Kim and Lee 2024; Liu and Zhang 2025; Dai et al 2025b; Li et al 2025d; Zhang et al 2025a) evaluate data difficulty based on the current capability of the model, which is estimated through real-time feedback signals during training. For instance, Kim and Lee (2024) propose a composite difficulty metric that incorporates sequence length, training loss, and attention scores. Liu and Zhang (2025) introduce a dual-metric design, combining ROUGE-L similarity between model outputs and correct answers with the cross-entropy loss of the model. Dai et al (2025b) further characterize difficulty in terms of quality and learnability, motivated by gradient consistency. Collectively, these works illustrate a trend towards multi-dimensional evaluation.

While the above methods focus on presenting tasks in an easy-to-hard sequence based on task difficulty, Chen et al (2023b) propose an alternative strategy that shifts the focus from modeling task difficulty to modeling the skills required to solve them. They introduce a "skill graph", which represents how learning prerequisite skills reduces the amount of data needed to master more advanced skills. Using this graph, they employ a dynamic sampling strategy that prioritizes data for skills the model has not yet mastered or that are needed for future tasks.

3.1.2 Data Mixing

This strategy optimizes the proportion of data from different task types in the training dataset subject to a fixed data budget. In practice, solving a given task typically requires the coordinated use of multiple capabilities of LLMs, with each capability contributing to a different extent. As a result, the strong performance of LLMs on a target task depends not only on the presence of relevant capabilities, but also on whether the distribution of these capabilities aligns with the capability requirements of that task (Xie et al 2023b; Albalak et al 2023). Data mixing addresses this challenge by adjusting the composition of training data so as to induce a capability distribution that better matches the demands of the target task, thereby improving training effectiveness without increasing the data scale. Accordingly, the central question in data mixing is how to choose the mixture ratios across domains in a principled manner. Early works typically employ heuristically determined or manually specified domain ratios when mixing data from different domains (Gao et al 2020; Raffel et al 2020). However, such hand-crafted ratios are often suboptimal, since they are unable to effectively capture the impact of data from different domains on the distribution of model capabilities (Albalak et al 2023). Therefore, recent research has shifted toward more principled strategies, which can be broadly grouped into two classes, namely offline optimization and online optimization.

Offline Optimization. The methods in this category assume that the relationship between data composition and training effect can be modeled. Under this assumption, offline optimization strategies can be broadly formulated as solving a constrained optimization problem. Specifically, for the mixing ratios $\alpha = \{\alpha_i\}_{i=1}^{n}$ of n domains, these strategies aim to solve:

$$\begin{cases} \arg\min\limits_{\alpha} \mathcal{L}_{\mathcal{T}}(\alpha) \\ \sum_{i=1}^{n} \alpha_i = 1 \\ \alpha_i \geq 0, i = 1, 2, ..., n \end{cases}, \tag{3.2}$$

where $\mathcal{L}_{\mathcal{T}}(\alpha)$ denotes the test loss on the target task $\mathcal{T}$ of the LLM trained on the dataset with the mixing ratios α. For modeling $\mathcal{L}_{\mathcal{T}}(\alpha)$, Ye et al (2024) leverage an exponential function to explicitly correlate test losses of LLMs with mixing ratios of different domains:

$$\mathcal{L}_{\mathcal{T}}(\alpha) = \sum_{j=1}^{K} s_j \mathcal{L}_j(\alpha) = \sum_{j=1}^{K} s_j \left[c_j + k_j \exp\left(\sum_{i=1}^{n} t_{ji} \alpha_i \right) \right], \tag{3.3}$$

where s_j is the weight of the j-th test domain in the target task $\mathcal{T}$ out of K total domains. The $\mathcal{L}_j(\cdot)$ represents the predicted test loss according to a training mixing ratio, while c_j, k_j, and t_{ji} are learnable parameters. By fitting this function, they can predict the task performance of the model trained with different data mixture ratios, thereby allowing us to select the optimal data composition. More directly, many existing methods (Liu et al 2025b; Ge et al 2024; Li et al 2025c) directly train an additional model to capture the relationship between data composition and training effect. Compared with formal mathematical formulations as in Eq. (3.3), this strategy typically yields more accurate predictions, but incurs higher computational cost and has weaker interpretability. To address these issues, Belenki et al (2025) propose a two-layer prediction structure. Firstly, they train multiple small-parameter models to predict the impact of data from each domain on the training effect. Secondly, according to the predictions of these models, they estimate the training effect of mixing data from various domains. This design reduces computational cost and offers improved interpretability while maintaining high prediction accuracy.

Online Optimization. The methods in this category dynamically adjust data mixture ratios during the training process of LLMs based on real-time feedback, such as training loss or model performance on the validation data. This adaptive mechanism (Xia et al 2024a; Albalak et al 2023; Fan et al 2024a) relies on continuous monitoring of model performance throughout training. As a result, obtaining such real-time feedback often requires numerous additional training or evaluation on LLMs during the training process, resulting in considerable computational cost and reduced training efficiency. To mitigate this limitation, several works (Xie et al 2023b; Fan et al 2024b) employ small proxy models to approximate the training behavior of target LLMs. By observing the learning performance of the proxy models on data from different domains, they estimate which domains are likely to benefit the target LLMs the most. To ensure that small proxy models can accurately reflect the training behavior of the target LLMs, existing methods align the optimization objectives of the proxies with the ultimate training goals of the target models, such as maintaining a balanced capability distribution or maximizing generalization. For instance, Xie et al (2023b) train a proxy model under group distributionally robust optimization (Oren et al 2019) to emulate the requirement of target models for balanced domain capabilities. Fan et al (2024b) utilize gradient alignment to evaluate whether training in a specific domain effectively reduces the loss on a validation set.

3.1.3 Augmentation via Generation

This strategy aims to enhance the value of existing data points by modifying their content, such as improving reasoning processes and adding domain expert knowledge. When incorporated into training, such augmented data points can provide more training effect, which leads to:

$$\begin{aligned} & \forall d \in \mathcal{D}, v\left(\mathrm{A}\left(d\right) \mid C\right) > v\left(d \mid C\right) \\ \Rightarrow \quad & \sum_{d \in \mathcal{D}} v\left(\mathrm{A}\left(d\right) \mid C\right) > \sum_{d \in \mathcal{D}} v\left(d \mid C\right) \\ \Rightarrow \quad & \frac{\sum_{d \in \mathcal{D}} v\left(\mathrm{A}\left(d\right) \mid C\right)}{\mu\left(\mathcal{D}\right)} > \frac{\sum_{d \in \mathcal{D}} v\left(d \mid C\right)}{\mu\left(\mathcal{D}\right)} \\ \Rightarrow \quad & \Delta f > 0, \end{aligned} \tag{3.4}$$

where $\mathrm{A}\left(d\right)$ denotes the augmented data point obtained by applying the augmentation operation to the original data point.

One of the most effective ways to elevate the value of training data is to externalize the latent reasoning processes underlying the answers. Accordingly, converting a standard instruction–response pair *(problem, answer)* into a triplet *(problem, reasoning process, answer)* has emerged as a predominant paradigm in the field of augmentation via generation. In early studies, the reasoning processes were carefully crafted by human experts (Rajani et al 2019; Ross et al 2017; Hancock et al 2019). As a result, these processes are highly aligned with human cognition, and thus providing reliable optimization signals during LLM training. Unfortunately, producing such carefully curated processes is costly and difficult to scale. To overcome this limitation, recent works (Hsieh et al 2023; Guo et al 2025a) explore leveraging powerful LLMs to generate reasoning processes for training data, such as DeepSeek-R1 (Guo et al 2025a), Qwen3-235B-A22B (Yang et al 2025), and Llama-3.1-405B-Instruct (Grattafiori et al 2024). By using this generated data for training, the target model can learn the knowledge, reasoning strategies, and reflective abilities possessed by these powerful

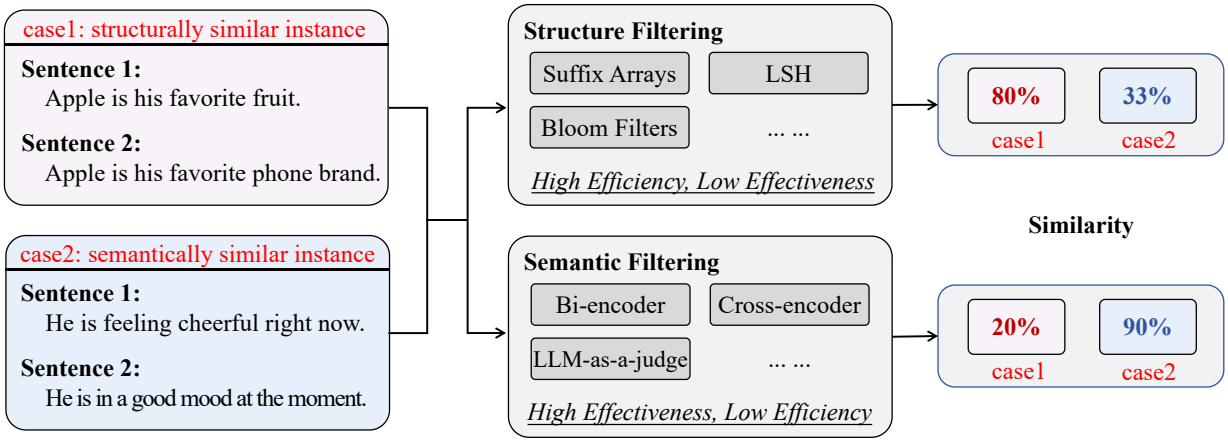


Figure 4: Duplicate data removal can be categorized into two strategies, *i.e.*, structure filtering and semantic filtering. For two texts, the former focuses on the similarity of their structures, while the latter emphasizes the similarity of their semantics.

LLMs, thereby improving its task performance. Despite its good scalability, generating reasoning processes with LLMs also introduces new challenges. Due to the unreliable generation capabilities and hallucination of LLMs, the quality and faithfulness of their generated reasoning processes are difficult to guarantee, which may introduce noise into the training data.

Beyond enriching training data by generating additional reasoning processes, another way to increase data value is to reformulate the structural format of training data. Instead of modifying the content of training data, this strategy alters how the data is presented, such as converting multiple-choice tasks into generative problem-solving tasks (MAA 2023), and organizing reasoning processes into a more structured and explicit format (Liu et al 2024b; Xu et al 2023a). Such reformulation reduces ambiguity and makes intermediate inference processes more explicit, allowing the model to learn more systematic problem-solving strategies.

3.2 $V(\mathcal{D} \mid \mathcal{C})$ Remains Unchanged with $\mu(\mathcal{D})$ Decreased

The methods in this category (Thangarasa et al 2024; Zhang et al 2024c; Tao et al 2024; Penedo et al 2024; Abbas et al 2023; Tirumala et al 2023) are grounded in the hypothesis that large-scale datasets typically contain a large amount of redundant or useless information, and the valuable information they hold is sparse. By removing the valueless information, we can significantly reduce the dataset size $\mu(\mathcal{D})$ while preserving the overall data value $V(\mathcal{D} \mid \mathcal{C})$, thereby yielding a substantial increase in DVD. Existing methods can be grouped into two classes, *i.e.*, **duplicate data removal** and **data distillation**.

3.2.1 Duplicate Data Removal

This strategy focuses on identifying and removing duplicate data points that do not provide additional information. These duplicates do not contribute new insights or knowledge to the model, resulting in zero marginal gain on model performance. In this context, the data value function $V(\mathcal{D} \mid \mathcal{C})$ is defined based on similarity calculation between different data points, as formalized in Eq. (2.3).

In the context of LLM training, data duplication is a prevalent issue, especially in large-scale datasets such as web-crawled corpora (Foundation 2023; Raffel et al 2020; Zellers et al 2019). High rates of duplication can lead to several detrimental effects during LLM training, including model overfitting and memorization tendencies (Lee et al 2022). To mitigate these effects, various deduplication techniques have been developed. These techniques can be categorized into two types: structure filtering, which focuses on eliminating duplicates based on structural similarity; and semantic filtering, which identifies and removes semantically redundant data points. The differences between these two types are shown in Fig. 4.

Structure Filtering. It is one of the most prevalent and stable data deduplication strategies in current construction pipelines for large-scale datasets such as The Pile (Gao et al 2020), C4 (Raffel et al 2020), and

RedPajama (Weber et al 2024), which identifies and removes texts that exhibit high structural similarity. In practice, this strategy is usually implemented through predefined rules, making it applicable to almost all data in the form of natural language, *e.g.*, suffix array, Locality Sensitive Hashing (LSH), and Bloom filters.

A suffix array (Manber and Myers 1993) is a classical data structure for string indexing that sorts all suffixes of a sequence in lexicographical order. Under this ordering, suffixes that share the same prefix are placed next to each other in the array. As a result, duplicated substrings can be detected by comparing adjacent suffixes and computing their longest common prefix. If the length of the common prefix exceeds a predefined threshold, the corresponding substring is considered duplicated. To apply this technique to large text corpora (Lee et al 2022; Penedo et al 2023; Kandpal et al 2022), researchers typically concatenate all documents into a single sequence and construct a global suffix array over it. In this setting, a suffix refers to the token sequence starting from any position and extending to the end of the corpus. This design enables efficient detection of duplicated content across different parts of the dataset. Because it relies on exact string matching, the method is particularly effective for detecting verbatim repetitions, such as boilerplate text, advertisements, and duplicated code snippets.

Unlike the suffix array, which employs exact string comparison, the methods based on LSH perform fuzzy similarity matching. The core idea is to map each document into a low-dimensional hash signature where similar documents produce similar signatures. Duplicate texts can then be identified by comparing these hash signatures instead of the original documents. A common practice (Lee et al 2022; Shen et al 2023; Penedo et al 2023) is to represent each text as a set of n-grams and compute the MinHash signature (Broder 1997), which approximates the Jaccard similarity between texts. Similarly, Gyawali et al (2020) adopt SimHash (Manku et al 2007) to encode document abstracts into fixed-length binary vectors and measure similarity by using Hamming distance (Hamming 1950). Despite their effectiveness, LSH-based methods often exhibit limited efficiency when applied to large-scale corpora, primarily due to the substantial computational overhead required to calculate numerous hash functions for generating document signatures. To address this issue, Zeng et al (2021) design a distributed fuzzy deduplication system built on a big-data processing framework, which significantly accelerates detection through parallelization.

Similar to LSH-based methods, Bloom filters also rely on hashing techniques to project texts into low-dimensional spaces. However, their purpose is fundamentally different. While LSH estimates the similarity between two texts, Bloom filters perform membership queries, which determine whether a given text segment has already appeared in the dataset. By avoiding pairwise similarity comparisons, Bloom filters enable efficient duplicate detection through direct hash lookups. For instance, Soldaini et al (2024) apply Bloom filters at multiple text levels (URL, document, and paragraph) during the construction of the Dolma corpus. To further enhance efficiency, Li et al (2024a) design a unified Bloom filter that simultaneously operates at both paragraph and document levels.

Semantic Filtering. While structure filtering methods are efficient for removing literal repetitions, they fall short in capturing semantic duplicates (*i.e.*, texts that differ in linguistic presentation but convey the same meaning). To address this limitation, semantic filtering methods have been developed. These methods (Chen et al 2024a; Shi et al 2024; Özgür Uğur et al 2026) use pre-trained models to encode each text into a dense vector that captures semantic meanings and then measure redundancy via embedding similarity. To reduce the number of similarity computations and improve scalability, Abbas et al (2023) and Tirumala et al (2023) apply K-means clustering after encoding texts into embeddings, and then calculate text similarities only within each cluster, avoiding exhaustive pairwise comparisons. Despite their efficiency, the strategy that independently compresses each text into a fixed-dimensional vector often shows poor effectiveness when processing complex, long, or noisy texts. Since the expressive capacity of a single fixed-dimensional vector is limited, it cannot preserve all salient semantic details. As a result, when a text contains rich content or complex logic, encoding it into a single vector inevitably discards fine-grained information, which weakens the accuracy of semantic deduplication. To alleviate this limitation, some studies propose hierarchical neural architectures. Silcock et al (2021) first employ a bi-encoder to efficiently retrieve the top-k nearest neighbors for each document as candidate duplicates. Subsequently, these candidates are fed into a cross-encoder alongside the query document. Unlike bi-encoders that compress documents independently, the cross-encoder processes the concatenated text pair simultaneously, allowing direct token-level interactions between the two texts. This design enables the model to capture subtle semantic differences during the encoding process. It is worth mentioning that the above approaches typically employ a static deduplication strategy that performs filtering as a fixed preprocessing step. This rigidity makes them ill-suited for self-supervised learning

pipelines where data redundancy emerges dynamically during training. To address this challenge, Choi et al (2022) introduce a mechanism inspired by human "working memory" that adaptively evaluates the collision probability of incoming data points against previously trained ones, enabling online pruning of redundant data as training progresses.

3.2.2 Data Distillation

For the original dataset $\mathcal{D}_{ori}$, data distillation aims to construct a new and smaller dataset $\mathcal{D}_{new}$ whose information content is comparable with that of the original dataset.

The Objective of Data Distillation

Construct $\mathcal{D}_{new}$ based on $\mathcal{D}_{ori}$ such that

$$\mu\left(\mathcal{D}_{new}\right) < \mu\left(\mathcal{D}_{ori}\right) \text{ and } V\left(\mathcal{D}_{new} \mid \mathcal{C} = \{\mathcal{P}\}\right) \approx V\left(\mathcal{D}_{ori} \mid \mathcal{C} = \{\mathcal{P}\}\right),$$

where $\mathcal{P}$ is commonly instantiated as the information content of $\mathcal{D}_{ori}$, and $V\left(\mathcal{D} \mid \mathcal{C} = \{\mathcal{P}\}\right)$ quantifies the extent to which $\mathcal{D}$ covers the information in $\mathcal{P}$.

Data distillation has attracted increasing attention in recent years. However, most existing studies focus on image data (Wang et al 2025b,d; Sun et al 2024b; Su et al 2024), with comparatively limited work in natural language processing. The main reason lies in the discrete nature of the text. In typical data distillation frameworks, synthetic data are treated as learnable parameters and are optimized by backpropagating the training loss of the model on real data. Such gradient-based optimization is straightforward for continuous inputs, but cannot be directly applied to discrete tokens, which are the universal representation forms for textual data. To address this issue, recent approaches (Li and Li 2021; Nguyen et al 2025; Maekawa et al 2023) map discrete text into continuous embedding spaces, enabling the employment of gradient-based dataset distillation strategies originally developed for image data. Building on this framework, numerous methods have been proposed to improve effectiveness and efficiency, such as co-optimizing attention distributions for richer supervisory signals (Maekawa et al 2023), and restricting gradient matching exclusively to the final layer to alleviate computational overhead (Nguyen et al 2025).

Although gradient-based methods have demonstrated significant efficacy, the continuous embeddings they generate are not human-readable and are difficult to transfer across different LLMs. To address this limitation, the research focus has gradually shifted toward generating synthetic data points that are readable and transferable. Most of the existing methods (Sucholutsky and Schonlau 2021; Sahni and Patel 2023) follow a two-stage framework. They first obtain continuous embeddings using the gradient-based distillation strategy mentioned above. These embeddings are then projected back to discrete text by mapping each vector to the nearest token in the model vocabulary. Moreover, Tao et al (2024) abandon the gradient-based distillation strategy and instead adopt a clustering distillation strategy. They first encode texts into embedding vectors using pretrained LLMs. These embeddings are then clustered to identify representative centroids, which serve as condensed prototypes of the original data distribution. Finally, the centroids are decoded back into natural language by using a vec2text model. This pipeline enables the generation of readable distilled data without relying on complex gradient-based optimization.

3.3 $V\left(\mathcal{D} \mid \mathcal{C}\right)$ Increases while $\mu\left(\mathcal{D}\right)$ Decreases

The methods in this category (Team et al 2024; Li et al 2024a; Choi et al 2024; Soldaini et al 2024; Shen et al 2025; Seo et al 2025; Bendinelli et al 2025) are motivated by the observation that large-scale datasets, particularly web-crawled corpora, often include a non-trivial fraction of low-quality or harmful data that can degrade model performance, such as toxic texts, biased information, and malformed documents. In such special cases, such harmful data points exhibit a negative training utility, meaning their individual data value $v\left(\cdot\right)$ takes a negative value. Identifying these harmful instances allows us to group them into an excluded subset, denoted as $\mathcal{D}_{exc}$. Consequently, by removing such data points in $\mathcal{D}$, we can increase $V\left(\mathcal{D} \mid \mathcal{C}\right)$ while simultaneously reducing $\mu\left(\mathcal{D}\right)$. We refer to these methods as **negative-value data removal**.

Table 1: Examples of negative-value data categorized into three types, *i.e.*, structural, semantic, and cognitive contamination.

Type		Example
Structural Contamination	Malformed Syntax	Results showing improve in model performance.
	Garbled Text	$%#@!! lorem ipsum ### 1234 ???
Semantic Contamination	Toxic Language	People from that race are criminals.
	Biased Information	Men are naturally better leaders than women.
Cognitive Contamination	Factual Inaccuracy	The capital of France is Berlin.
	Logical Inconsistency	All birds can fly. Penguins are birds. Therefore, penguins cannot fly.

The Objective of Negative-value Data Removal

Identify an excluded subset $\mathcal{D}_{exc}$ from $\mathcal{D}_{ori}$ such that

$$\mathcal{D}_{new} = \mathcal{D}_{ori} - \mathcal{D}_{exc},$$

where $\forall d \in \mathcal{D}_{exc},\ v(d \mid C) < 0$.

According to the above definition, the key challenge of negative-value data removal is to accurately identify harmful data within the dataset. As summarized in Table 1, such harmful data can be categorized into three types, *i.e.*, structural, semantic, and cognitive contamination. Structural contamination refers to surface-level formal defects that violate basic linguistic conventions, such as malformed syntax and garbled text. Semantic contamination involves content that is syntactically valid but semantically misleading or harmful, including toxic language and biased information. Cognitive contamination occurs when data contains factual inaccuracies or logical inconsistencies. Based on this categorization, existing methods can be grouped into three corresponding paradigms, *i.e.*, structural contamination removal, semantic contamination removal, and cognitive contamination removal.

Structural Contamination Removal. This category focuses on removing formal defects, which represent data points that violate fundamental structural conventions of natural language. To achieve this goal, a primary strategy is to apply heuristic rules that are manually defined or statistically motivated to detect garbled text, formatting errors, and other structural anomalies. For manually defined heuristic rules, common practices in existing methods include discarding overly short text segments (Grave et al 2018; Rae et al 2021; Xue et al 2021), excluding documents with JavaScript placeholders (Raffel et al 2020; Weber et al 2024), filtering out lines without punctuation (Penedo et al 2023; Park et al 2025), and so on. For statistically motivated heuristic rules, Seo et al (2025) propose a filtering method based on token frequencies. They first estimate the prior probability of each token using a large corpus, and then calculate the mean and standard deviation of these prior probabilities for all tokens within a given document. Since well-formed data points naturally maintain a stable distribution of high-frequency and low-frequency words, the documents that significantly deviate from the normal range in these two metrics are identified as structural anomalies and should be removed. The effectiveness of the above heuristic filtering has been widely demonstrated. Many studies (Penedo et al 2023; Shen et al 2023) show that carefully designed heuristic rules can substantially improve data quality. In many cases, such strategies can significantly reduce the dataset size while meaningfully enhancing the generalization capability of LLMs.

To standardize and systematize these heuristic rules, recent works (Soldaini et al 2024; Li et al 2024a; Chen et al 2024a; Penedo et al 2023; Laurençon et al 2022) have shifted toward unified processing frameworks. For instance, Soldaini et al (2024) propose an open-source data processing toolkit that integrates quality filtering, toxicity detection, and personally identifiable information redaction into a unified and reproducible processing pipeline. This design consolidates various heuristic rules into a coherent and extensible framework.

In parallel, Li et al (2024a) establish a standardized benchmark to quantitatively evaluate the contribution of different heuristic rules to the training value of the dataset. Together, these efforts transform heuristic filtering from loosely assembled rule sets into a more systematic and measurable engineering process.

Semantic Contamination Removal. In contrast to structural contamination removal, this category focuses on identifying and removing the content that is syntactically well-formed yet is semantically harmful to LLMs, such as label–data mismatches in supervised data, toxic language, and biased statements. To detect such semantic contamination, most existing approaches (Grattafiori et al 2024; Penedo et al 2024; Li et al 2024a; Team et al 2024; Korbak et al 2023; Arnett et al 2024; Maini et al 2025) rely on well-trained models to evaluate the semantic information of individual data points and make filtering decisions accordingly. One common strategy (Grattafiori et al 2024; Team et al 2024) is to formulate the problem as a binary classification task, where the semantic information of each data point is directly labeled as "acceptable" or "harmful". Then, the data points predicted as "harmful" are discarded. Alternatively, scoring models are employed to score data points, allowing those with low scores to be filtered out (Korbak et al 2023; Arnett et al 2024; Maini et al 2025).

Despite their effectiveness, the above approaches typically rely on fixed evaluation mechanisms. In these methods, both the assessment model and the decision thresholds remain static once defined, making performance highly dependent on the capacity of the chosen model and the threshold configuration. As a result, their robustness and transferability can be limited when applied to new domains or tasks. To address this limitation, Hu et al (2026) propose a dynamic evaluation framework that scores each data point by calculating the directional alignment between its gradient on the LLM and a reference gradient that preserves the safety alignment of the LLM. Combined with an adaptive thresholding algorithm, this method enables adaptive and accurate identification of detrimental data points.

Cognitive Contamination Removal. This category primarily targets cognitive errors, including factual inaccuracies and logical inconsistencies. Identifying these errors requires powerful reasoning capabilities and extensive world knowledge. Therefore, the methods in this category (Feng et al 2025; Nie et al 2021; Chen et al 2019; Bai et al 2026) typically involve human experts or powerful LLMs to detect and remove cognitive contamination in the training data. Specifically, human expert review provides high-fidelity judgments, but it is slow and expensive to scale to large-scale datasets. In contrast, using powerful LLMs enables much more efficient detection, yet the resulting decisions can be less reliable. Due to hallucinations and biases, LLMs may miss subtle logical errors and may also incorrectly label high-quality data as contamination.

Notably, the capabilities of powerful LLMs extend beyond detecting cognitive contamination. Owing to their broad linguistic competence, these models can also effectively identify structural and semantic contamination. As a result, several recent approaches (Zhou et al 2024a; Shen et al 2025; Bendinelli et al 2025) integrate multi-level contamination removal into a single automated pipeline, achieving unified removal of structural, semantic, and cognitive contamination. By leveraging the reasoning capabilities of powerful LLMs, this framework enables adaptive detection of negative-value data without relying on manually engineered heuristics. LLMs are used to infer potential contamination in the training data, propose appropriate cleaning strategies, and generate executable procedures to apply these strategies. As a result, this design can achieve robust contamination removal in diverse domains.

3.4 $V(\mathcal{D} \mid \mathcal{C})$ Decreases with $\mu(\mathcal{D})$ Decreased More

This category of approaches is based on the premise that data points within the original dataset $\mathcal{D}_{ori}$ exhibit heterogeneous training utility. While most of these data points contribute positively to model training, their individual contribution varies significantly. Therefore, the methods in this category prioritize the data points with higher training value and construct a subset to serve as the final training dataset $\mathcal{D}_{new}$. Specifically, given a value threshold $\delta > 0$, the original dataset $\mathcal{D}_{ori} = \{d \mid v(d \mid \mathcal{C}) > 0\}$ is partitioned into a selected subset $\mathcal{D}_{new} = \{d \in \mathcal{D}_{ori} \mid v(d \mid \mathcal{C}) > \delta\}$ and an excluded subset $\mathcal{D}_{exc} = \{d \in \mathcal{D}_{ori} \mid v(d \mid \mathcal{C}) \leq \delta\}$. Based on this construction, it naturally follows that the DVD of the selected subset is greater than that of the excluded subset, *i.e.*, $f(\mathcal{D}_{new} \mid \mathcal{C}) > f(\mathcal{D}_{exc} \mid \mathcal{C})$. We refer to these methods as **high-value data selection**. Although this strategy may theoretically lead to a marginal decrease in dataset value (*i.e.*, $V(\mathcal{D}_{new} \mid \mathcal{C}) < V(\mathcal{D}_{ori} \mid \mathcal{C})$), the DVD of $\mathcal{D}_{new}$ can increase significantly. The detailed derivations are below:

$$f(\mathcal{D}_{ori} \mid \mathcal{C}) = \frac{V(\mathcal{D}_{ori} \mid \mathcal{C})}{\mu(\mathcal{D}_{ori})}$$

$$
\begin{aligned}
&= \frac{V(\mathcal{D}_{new} \mid C) + V(\mathcal{D}_{exc} \mid C)}{\mu(\mathcal{D}_{ori})} \\
&= \frac{f(\mathcal{D}_{new} \mid C)\,\mu(\mathcal{D}_{new}) + f(\mathcal{D}_{exc} \mid C)\,\mu(\mathcal{D}_{exc})}{\mu(\mathcal{D}_{new}) + \mu(\mathcal{D}_{exc})} \\
&\overset{1}{<} f(\mathcal{D}_{new} \mid C) \cdot \frac{\mu(\mathcal{D}_{new}) + \mu(\mathcal{D}_{exc})}{\mu(\mathcal{D}_{new}) + \mu(\mathcal{D}_{exc})} \\
&= f(\mathcal{D}_{new} \mid C),
\end{aligned} \tag{3.5}
$$

where the 1st inequality uses the fact that the DVD of the selected subset is strictly greater than that of the excluded subset (*i.e.*, $f(\mathcal{D}_{new} \mid C) > f(\mathcal{D}_{exc} \mid C)$). It is worth mentioning that high-value data selection is conceptually distinct from negative-value data removal discussed in Section 3.3. Specifically, negative-value data removal aims to eliminate data points that adversely affect training while retaining those with positive contributions, leading to a considerable improvement in training effectiveness. In contrast, high-value data selection ranks positive-value data points according to their training utility and retains only those with high marginal benefits. This strategy reduces data scale and computational cost while maintaining competitive performance. In practice, high-value data selection can be applied after negative-value data removal to further increase the DVD of the dataset. For instance, Penedo et al (2024) construct the FineWeb dataset by removing the negative-value data points in the Common Crawl corpus (Foundation 2023). Based on the FineWeb dataset, Wang et al (2025f) select a portion of high-value data to construct the Ultra-FineWeb dataset.

High-value data selection necessarily involves a trade-off between efficiency and data coverage, because it discards data points that contribute positively to training but have relatively low utility. If the data value is estimated inaccurately, informative data points may be removed, while less useful ones may be kept, which can substantially degrade training performance. Therefore, the key challenge of high-value data selection is to assess data value accurately and consistently. To address this issue, existing methods (Muennighoff et al 2025; Ye et al 2025; Deb et al 2025; Miranda et al 2023; Muennighoff et al 2023; Dai et al 2025b; Schoch et al 2023; Lin et al 2024a; Xie et al 2023c; Kung et al 2023; Cao et al 2023; Wei et al 2023; Xia et al 2024b; Yao et al 2022; Fayyaz et al 2022; Chen et al 2023a; Attendu and Corbeil 2023; Busa-Fekete et al 2026; Zhang et al 2026) propose a variety of evaluation strategies to estimate the contribution of each data point from different perspectives. In summary, they can be categorized into two classes, namely intrinsic value evaluation and interactive value evaluation.

Intrinsic Value Evaluation. The approaches in this category assess data value based on the intrinsic properties of data itself, without relying on direct interaction with the target model $\mathcal{M}$ or downstream task $\mathcal{T}$. The underlying assumption is that high-value data points exhibit certain desirable characteristics, such as higher quality, stronger representativeness, richer information, or broader semantic coverage. By constructing standalone scoring mechanisms or geometric selection criteria, these methods can identify high-value data points before model training.

One major direction uses auxiliary evaluators to explicitly assess data quality. A natural choice is to employ powerful teacher models that can judge complex attributes such as logical coherence, instruction quality, and task relevance (Chen et al 2024b; Parkar et al 2024). Because repeated calls to large teacher models are costly, many studies distill such evaluative ability into lightweight scoring models. These approaches typically use teacher-generated labels based on predefined criteria, such as writing style or educational value, to train smaller evaluators for efficient large-scale filtering (Wettig et al 2024; Liu et al 2024a). More recent works further systematize this paradigm by automatically inducing evaluation dimensions and training universal data managers capable of multi-dimensional quality assessment and domain recognition (Peng et al 2025). In parallel, several studies show that small proxy models can provide useful approximations to the judgments made by larger models, enabling efficient value estimation at a lower computational cost (Li et al 2024c; Yang et al 2024; Ankner et al 2024).

The second direction explores training-free intrinsic signals. Instead of fine-tuning evaluators, these methods estimate value directly from the structural or behavioral properties of the data at inference time. One representative idea is the demonstration effect, which suggests high-value data should serve as strong in-context exemplars and improve model behavior when used as demonstrations (Li et al 2024e). Other work analyzes internal model mechanisms such as attention patterns to determine whether a data point provides meaningful semantic support for solving the target task (Wang et al 2025c). Compared with evaluator-based

approaches, these methods avoid additional training and can be appealing when efficiency is critical, although they often depend strongly on the choice of proxy signal.

The third direction assesses value through collective coverage in feature space. Rather than evaluating each data point independently, these methods emphasize the geometric diversity and representativeness of the selected subset. Motivated by theoretical and empirical findings that broader semantic coverage is beneficial for generalization (Miranda et al 2023), prior works use clustering, greedy sampling, or distance-based subset construction to retain the data points that best preserve the global structure of the original dataset (Wu et al 2023; Chen et al 2023a). Some methods further refine this idea by introducing tag-based or concept-level annotations, and then selecting the data points that maximize both coverage and novelty in semantic space (Lu et al 2024). This line of work is especially useful when the goal is to maintain diversity and avoid redundancy.

Overall, intrinsic value evaluation can be performed prior to training and does not require repeated interaction with the target model. This makes it computationally efficient, reusable across settings, and suitable for large-scale preprocessing. However, because the value is inferred from proxy signals or structural properties, these methods may fail to capture task-specific utility. Their effectiveness therefore depends heavily on the choice of evaluator, signal, or feature representation. In addition, different intrinsic criteria, such as quality, difficulty, and diversity, may favor different subsets, making it nontrivial to define a universally reliable notion of value.

Interactive Value Evaluation. The methods in this category assess data value through interaction with a specific training context. Rather than treating value as an intrinsic property of the data, they assess it based on how an example affects the target model, validation objective, or downstream task. The common assumption is that a valuable data point is the one that provides stronger learning signals under the current training setting, for example by producing larger performance gains, higher information gain, or better alignment with the target task.

One important direction evaluates data value from training feedback. These methods use signals generated during optimization, such as loss, gradients, or prediction uncertainty, to estimate the current utility of each data point. A representative line of work ranks data points according to gradient magnitude or prediction error, based on the intuition that the data points inducing larger parameter updates are often more informative for learning (Fayyaz et al 2022; Attendu and Corbeil 2023). Some methods further extend this idea dynamically by updating scores throughout training, so that already-mastered data points and persistently unlearnable noisy data points can be gradually deprioritized. Since these signals are already produced during training, such methods are often computationally efficient and are naturally adaptive to the evolving state of the model.

However, raw training signals are often sensitive to superficial factors such as text length, token frequency, and domain-specific style. To improve robustness, another line of work introduces reference-model-based relative metrics (Lin et al 2024b; Zhou et al 2025). Instead of relying on the absolute loss of the current model, these methods compare the target model with a reference model and measure relative discrepancy, for example, through excess loss. This design helps to distinguish genuinely informative data points from those that are universally trivial or universally difficult. Related approaches also incorporate information gain or uncertainty-aware criteria to reduce selection bias and better capture data points that are challenging yet still learnable (Deb et al 2025; Kung et al 2023; Li et al 2025b; Liu and Zhang 2025). In addition, some methods argue that valuable data should exhibit strong input-output dependency, and therefore measuring how much the model relies on the input context during generation (Li et al 2024d; Lv et al 2025). This perspective is particularly useful in instruction tuning and code generation, where faithful conditional generation is more important than raw linguistic complexity.

The second interactive direction estimates data value through validation feedback. The central idea is that a data point is valuable if training on it yields larger improvements on a validation set. Classical approaches formulate this as a marginal contribution estimation problem. For example, Shapley-value-based methods measure the average contribution of a data point across different subsets of the training data (Shapley 1953). Although theoretically attractive, exact computation is prohibitively expensive for large models, so practical methods rely on approximations such as proxy models or clustering-based estimation (Schoch et al 2023; He et al 2024). Gradient-based alternatives estimate the value by analyzing how a data point would affect validation performance through its gradient update (Xia et al 2024b), while more recent works introduce dynamic evaluation mechanisms that periodically probe the model and update value estimates as training progresses (Yu et al 2024; Gu et al 2024a). Since selecting only the highest-scoring data points can

reduce diversity and produce overly homogeneous datasets, some methods further incorporate exploration, complementarity, or diversity-aware metrics into the selection process (Zhang et al 2024a; Bai et al 2025).

The third direction defines data value through task alignment. These methods assume that high-value data should be well matched to the statistical properties of a target task or domain. A common strategy is to use small sets of task-specific data as anchors, and then retrieve training data points that are closest to them in feature space (Yao et al 2022; Ivison et al 2023). Other methods estimate density ratios between general data and target-task data, and use these ratios as sampling weights to construct subsets that are better aligned with the target distribution (Xie et al 2023c). Compared with general-purpose data selection, such methods are especially effective when strong domain adaptation is required.

Overall, interactive value evaluation measures data value with explicit respect to a model, a task, or a validation objective, giving it strong target relevance. It can adapt to different training stages and often produces the data subsets that are closely aligned with downstream performance. However, these methods also have clear limitations. First, the estimated value is highly context-dependent and may not transfer across models, tasks, or training stages. Second, frequent interaction with the target model, validation set, or auxiliary reference models can introduce substantial computational overhead. Third, optimizing only for immediate feedback may bias selection toward short-term gains and may reduce diversity. These trade-offs make efficiency, robustness, and diversity preservation the central challenges in this line of work.

3.5 $V(\mathcal{D} \mid C)$ Increases with $\mu(\mathcal{D})$ Increases Less

This category of approaches (Zhou et al 2024b; Xu et al 2024a; Majumdar et al 2025; Hao et al 2025; Yaswanth et al 2025; Luo et al 2024) first derives an evolved dataset $\mathcal{D}_{evol}$ from the original one $\mathcal{D}_{ori}$ and then integrates the two to produce the new training dataset $\mathcal{D}_{new}$. We refer to these methods as **data evolution**. It is worth mentioning that $\mathcal{D}_{evol}$ generally has a higher DVD when compared with $\mathcal{D}_{ori}$. Therefore, after merging the two, $\mathcal{D}_{new}$ also attains a higher DVD than $\mathcal{D}_{ori}$. This rationale is formally described as:

$$
\begin{aligned}
f(\mathcal{D}_{new} \mid C) &= \frac{V(\mathcal{D}_{new} \mid C)}{\mu(\mathcal{D}_{new})} \\
&= \frac{V(\mathcal{D}_{ori} \mid C) + V(\mathcal{D}_{evol} \mid C)}{\mu(\mathcal{D}_{new})} \\
&= \frac{f(\mathcal{D}_{ori} \mid C)\,\mu(\mathcal{D}_{ori}) + f(\mathcal{D}_{evol} \mid C)\,\mu(\mathcal{D}_{evol})}{\mu(\mathcal{D}_{ori}) + \mu(\mathcal{D}_{evol})} \\
&\overset{1}{>} f(\mathcal{D}_{ori} \mid C) \cdot \frac{\mu(\mathcal{D}_{ori}) + \mu(\mathcal{D}_{evol})}{\mu(\mathcal{D}_{ori}) + \mu(\mathcal{D}_{evol})} \\
&= f(\mathcal{D}_{ori} \mid C)\,.
\end{aligned}
\tag{3.6}
$$

In above derivations, the 1st inequality uses the fact that the DVD of the evolved dataset is strictly greater than that of the original dataset (*i.e.*, $f(\mathcal{D}_{evol} \mid C) > f(\mathcal{D}_{ori} \mid C)$). According to the above definition, the core research question in data evolution is how to construct $\mathcal{D}_{evol}$ based on $\mathcal{D}_{ori}$. Compared with $\mathcal{D}_{ori}$, $\mathcal{D}_{evol}$ should not only exhibit higher quality but also possess sufficient heterogeneity, so that the merged dataset $\mathcal{D}_{new}$ contains more diverse and richer information. Table 2 shows some strategies for constructing new data based on the original data in data evolution.

A major direction in this line of work focuses on complexity-oriented evolution. The basic idea is to rewrite simple data points into more difficult ones, so that the evolved data can better challenge the capabilities of LLMs. Early methods, represented by Evol-Instruct (Xu et al 2024a), rely on predefined heuristic rules to increase difficulty, such as adding constraints, increasing reasoning steps, or generating related variants. This paradigm was later extended to domain-specific settings such as coding and mathematics, where the evolution process is designed according to the characteristics of the target domain. For example, in the coding domain, Luo et al (2024) adapt evolution rules to programming-specific settings, such as imposing time and space complexity constraints or introducing debugging-oriented tasks. In the mathematical domain, Luo et al (2023) enrich the dataset not only by increasing problem complexity through additional reasoning steps, but also by generating simpler prerequisite questions to strengthen model understanding of fundamental concepts. Collectively, these methods show that carefully designed transformations can substantially improve the training utility of the resulting data.

Table 2: Examples of data evolution. Starting from the original data, an evolved sample can be constructed through rewriting, augmentation, fusion, or transformation, resulting in data with higher diversity or difficulty.

Original data	**Evolved data**
Write a Python function to determine whether a number is prime.	Write a Python function to determine whether a number is prime, analyze its time complexity, and then provide an optimized version for large numbers with test cases.
Solve the equation $2x + 3 = 11$.	Solve the equation $2x + 3 = 11$, explain the rationale behind each transformation step, and summarize the general procedure for solving linear equations.
Summarize the main idea of this article.	Summarize the article from the perspectives of policy impact, economic impact, and social impact, and identify possible bias of the author.
Recommend three introductory books on machine learning.	Recommend three introductory books on machine learning for beginners, and describe the target readers, prerequisites, strengths, and limitations of each book.
Translate the following sentence into English.	Translate the following sentence into English, and provide both a formal version and a colloquial version.
Write an email to ask for leave.	Write a formal leave request email to a supervisor, including the reason, duration, make-up plan, and a polite tone.
Instruction 1: Explain Newton's first law. Instruction 2: Give a real-life example of inertia.	Explain Newton's first law, provide two real-life examples of inertia, and clarify a common misunderstanding about the concept.
Instruction 1: What is recursion? Instruction 2: Write Python code for Fibonacci numbers.	First explain recursion, then implement both recursive and iterative Fibonacci algorithms in Python, and compare their efficiency.
Write a product description.	Rewrite the product description into three versions: an e-commerce listing, a social media post, and a formal product manual entry.
Judge whether the following argument is valid.	Judge whether the following argument is valid, identify possible logical fallacies, and provide a revised version with stronger reasoning.
Answer this multiple-choice question.	This question is a hard variant targeting a model weakness: answer it and explain why each incorrect option is wrong.

However, unconstrained complexity expansion often introduces new problems. Repeated rewriting may cause the evolved instruction to drift away from the original intent, while excessive or poorly verified constraints can make the generated task unnatural, inconsistent, or even unsolvable. To address these issues, later work explores constraint-aware evolution strategies that explicitly preserve semantic coherence during rewriting (Zhao et al 2024b). Instead of allowing free-form expansion, these methods control how complexity is introduced, so that evolved data points become harder while remaining faithful to the original semantics. A related direction uses instruction fusion rather than repeated deepening, which proposes to generate more complex data by combining multiple seed instructions into a single coherent task (Guo et al 2024). Other works further improve the flexibility of fusing $\mathcal{D}_{ori}$ and $\mathcal{D}_{evol}$ by allowing LLMs to design or refine evolution strategies themselves, reducing dependence on manually specified heuristics (Zeng et al 2024).

Another important trend is to enhance data evolution through multi-agent collaboration. A common motivation is that evolution performed by a single model is more prone to hallucination, mode collapse, and

limited diversity. To improve both validity and diversity, recent approaches decompose evolution into multiple roles or stages, allowing different agents to handle instruction rewriting, response generation, verification, or policy updating (Zhou et al 2024b; Mitra et al 2024; Majumdar et al 2025). Compared with single-model generation, such collaborative frameworks provide stronger quality control and enable a more diverse exploration of the data space.

Beyond instruction tuning, data evolution has also been extended to broader scenarios. In pre-training corpus construction, some methods improve diversity by rewriting the same document into multiple variants with different genres, styles, or target audiences (Hao et al 2025). In another direction, evolution can be guided by model weaknesses, which suggests difficult data points are first identified, and then expanded into new data that specifically target the current deficiencies of the model (Lee et al 2024b). These extensions suggest that data evolution is not only limited to instruction complexity, but can also serve as a general mechanism for diversity enhancement and targeted capability improvement.

Overall, data evolution is attractive because it actively creates new supervision signals rather than relying only on existing data. It can improve data quality, increase diversity, and generate training data points that are better aligned with the target capabilities. However, its effectiveness strongly depends on the reliability of the evolution process itself. Uncontrolled rewriting may introduce semantic drift, factual errors, or artificial difficulty, while more advanced solutions such as multi-agent collaboration or iterative verification often come with higher computational cost and pipeline complexity. These trade-offs make controllability, validity, and diversity preservation the central design challenges in this line of work.

4 Tasks and Datasets

As a data-centric paradigm, the continuous evolution of DVD enhancement for LLM training has inevitably led to the rapid development of the datasets employed in this research field. These datasets have become increasingly diverse and heterogeneous, varying in task type, problem format, and difficulty. Since existing methods in the field of DVD enhancement for LLM training typically take into account the characteristics of the target tasks and the data used, a clear understanding of the tasks and datasets helps to identify the strengths and limitations of current techniques, as well as potential areas for further improvement. As a result, this section reviews mainstream tasks, introduces their typical data structures and representative datasets, which were originally constructed for other specific capability evaluations but are now usually repurposed to study and test DVD enhancement methods. The overview of the datasets discussed in this section is presented in Table 3.

4.1 Text Understanding Task

Text understanding requires LLMs to comprehend and represent textual content, capturing not only literal meaning but also deeper aspects such as emotion, intention, and contextual relevance. For LLMs, it serves as a fundamental capability that underpins other competencies. Enhancing this capability enables accurate reasoning and robust knowledge utilization, thereby improving the performance of LLMs across a wide range of downstream tasks such as question answering, summarization, and dialogue. Consequently, extensive works (Press et al 2021; Chen et al 2023c; Ding et al 2023; Wu et al 2022; Sun et al 2023) have been proposed to strengthen the text understanding capability of LLMs, leading to the emergence of numerous datasets (Shaham et al 2022, 2023; An et al 2024; Kočiskỳ et al 2018; Dasigi et al 2021) in this area. Depending on the availability of data labels, existing text understanding datasets can be categorized into *unlabeled* datasets and *labeled* datasets. Table 4 summarizes the key characteristics of these two types of datasets across multiple dimensions, which are discussed in detail in the following subsections.

Unlabeled datasets of the text understanding task are primarily employed during the pre-training stage of LLMs. They are designed to enable LLMs to acquire a general understanding of human language, such as grammar, semantics, and discourse structure, through self-supervised learning objectives. To achieve this goal, such datasets (Foundation 2023; Weber et al 2024; Suarez et al 2020; Raffel et al 2020) are typically constructed at an extremely large scale by collecting massive and diverse textual corpora from web sources and public databases (Ide and Suderman 2004; Leech 1992; Sebastian Nagel 2016) to ensure broad linguistic and domain coverage.

Table 3: Popular datasets employed in DVD enhancement for LLM training. For dataset volume, "K", "M", "B", and "T" denote kilo, million, billion, and trillion, respectively. The symbol "-" indicates the datasets that are continuously updated or have undisclosed sizes.

Dataset	Volume	Description	Variant / Series
Text Understanding			
The Pile (Gao et al 2020)	-	English text corpus covering 22 diverse high-quality subtasks	
ANC (Ide and Suderman 2004)	11.0M	The American National Corpus	
BNC (Leech 1992)	100.0M	The British National Corpus	
News-Crawl (Sebastian Nagel 2016)	-	Text data of news in multiple languages	
BookCorpus (Zhu et al 2015)	-	A single-domain corpora of books	
Common Crawl (Foundation 2023)	410.0B	Petabytes of data regularly collected since 2008	OSCAR (Suarez et al 2020)
C4 (Raffel et al 2020)	156.0B	A colossal and cleaned version of Common Crawl	mC4 (Raffel et al 2020)
RedPajama-V1 (Weber et al 2024)	1.2T	Multilingual pre-training dataset	RedPajama-V2 (Weber et al 2024)
DocRed (Yao et al 2019)	106.9K	The relation extraction dataset constructed based on Wikipedia and Wikidata	Re-DocRed (Tan et al 2022)
LongBench (Bai et al 2023)	4.7K	21 datasets across 6 long-context understanding subtasks in both English and Chinese	LongBench v2 (Bai et al 2024)
GLUE (Wang et al 2018)	1.4M	A collection of natural language understanding tasks including question answering, sentiment analysis, and textual entailment	SuperGLUE (Wang et al 2019)
TACRED (Zhang et al 2017)	106.3K	Relation extraction problems covering 41 different relation types	TACREV (Alt et al 2020)
Reasoning			
MATH (Hendrycks et al 2021b)	12.5K	Competition mathematics problems with step-by-step solutions	
CommonsenseQA (Talmor et al 2019)	12.2K	Commonsense question answering	CommonsenseQA 2.0 (Talmor et al 2021)
OpenThoughts (Guha et al 2025)	114.0K	Reasoning problems covering math, science, code, and puzzles	OpenThoughts 2,3 (Guha et al 2025)
AMC'23 (MAA 2023)	83	Mathematics problems from AMC12 2022 and AMC12 2023	
SWAG (Zellers et al 2018)	113.0K	Grounded commonsense inference	
CoT Collection (Kim et al 2023)	1.8M	Reasoning problems covering 1,060 tasks	
GSM8K (Cobbe et al 2021)	8.5K	Grade school math problems	GSM-IC (Shi et al 2023)
AIME'24 (MAA 2024)	30	Problems collected from the 2024 American Invitational Mathematics Examination	
GPQA (Rein et al 2023)	448	Multiple-choice questions written by domain experts in biology, physics, and chemistry	GPQA Diamond (Rein et al 2023)
BIG-Bench (Srivastava et al 2023)	204	Problems from linguistics, math, common-sense reasoning, biology, physics, and beyond	BIG-bench-Hard (Suzgun et al 2023)
KnowLogic (Zhan et al 2025)	3.0K	Commonsense knowledge, plausible scenarios, and various types of logical reasoning	
Vertical Domain			
CBLUE (Zhang et al 2022)	195.9K	Biomedical language understanding problems in Chinese	PromptCBLUE (Zhu et al 2023)
LegalBench (Guha et al 2023)	91.7k	Legal problems in the United States legal system	
FinanceBench (Islam et al 2023)	10.2K	Financial question answering about publicly traded companies	
Xiezhi (Gu et al 2024b)	249.6K	Multiple-choice questions across 516 diverse disciplines ranging from 13 different subjects	
LawBench (Fei et al 2024)	10.0K	Legal problems divided into three levels: legal knowledge memorization, legal knowledge understanding, and legal knowledge applying	

Table 4: Comparison between unlabeled and labeled datasets in the text understanding task.

	Unlabeled Datasets	Labeled Datasets
Format	{"text": " "}	{"text": " ", "label": " "}
Scale	B (billion), T (trillion)	K (kilo)
Source	Website, Database	Human & Machine
Quality	Low	High
Application	Pre-training	Post-training, Evaluation

However, due to the vast volumes and heterogeneous origins of these corpora, fine-grained DVD enhancement strategies are difficult to implement. Most existing methods for these unlabeled datasets (Wang et al 2025f; Penedo et al 2023; Wenzek et al 2020) improve DVD at a coarse-grained level, such as garbled text removal and data deduplication. Consequently, these corpora often remain suboptimal, containing a large amount of noise such as factual errors, grammatical errors, and spam.

Labeled datasets of the text understanding task are mainly used during the post-training and evaluation stages of LLMs. After LLMs acquire the text understanding capability from vast amounts of unlabeled data during the pre-training stage, they are designed to enhance or assess specific aspects of this capability, such as long-content understanding (Shaham et al 2022, 2023; Ma et al 2024), information extraction (Zhang et al 2017; Pontiki et al 2015; Yao et al 2019), and sentiment analysis (Wang et al 2018, 2019; Maas et al 2011). These datasets are usually small in scale (ranging from 1K to 100K data points) and feature high-quality data with minimal noise, as each data point is carefully designed and annotated. As a result, these well-constructed datasets can provide robust support for both post-training and performance evaluation of LLMs.

However, the high quality of these labeled datasets relies heavily on substantial resources invested in their construction. For instance, data annotation typically requires the involvement of domain experts to ensure the accuracy and consistency of annotation results. To reduce this burden, recent studies (Liu et al 2025a; Muennighoff et al 2025; Xu et al 2024b) have explored using LLMs for automatic data annotation. Unfortunately, the inherent limitations of current LLMs, such as hallucination and cognitive bias, make it difficult to guarantee annotation reliability, potentially reducing the overall quality of datasets.

4.2 Reasoning Task

Among the diverse abilities of LLMs, reasoning stands out as a core yet challenging competency that supports complex decision making, problem solving, and multi-step inference (Yu et al 2020; Qin et al 2025; Shi et al 2023; Wang et al 2025a; Lin et al 2025). Recent studies have explored a wide range of strategies to enhance this capability, such as prompt engineering (Kojima et al 2022; Wei et al 2022; Zhou et al 2023b; Wang et al 2023a; Sun et al 2025b; Zhang et al 2025b), automated reasoning processes (Saha et al 2024; Lei et al 2023a; Chen et al 2024c; Besta et al 2024; Sun et al 2025c), supervised fine-tuning (Brown et al 2020; Radford et al 2021; Wei et al 2021), and reinforcement learning (Yu et al 2025; Rafailov et al 2023; Shao et al 2024).

In contrast, DVD enhancement acquires reasoning improvement from a data-centric perspective. It seeks to substantially boost the reasoning capability of LLMs by constructing training data that are structurally rich, logically coherent, and complete in intermediate inference steps. As a natural consequence of this data-focused paradigm, the rapid development of DVD enhancement methods is tightly coupled with the emergence of new reasoning datasets. Each newly proposed strategy often gives rise to training or evaluation datasets (Hendrycks et al 2021b; MAA 2024; Balunović et al 2025; MAA 2023; Rein et al 2023; Dua et al 2019; Ling et al 2017; Cobbe et al 2021; Hendrycks et al 2021a) that reflect its underlying design principles. In summary, existing reasoning datasets typically share a unified data structure:

```
{
    "problem": "",
    "ground_truth": {
        "reasoning": "",
        "result": ""
```

```
        }
    },
```

where *problem*, *reasoning*, and *result* denote the question to be solved, the step-by-step reasoning process, and the right answer, respectively. Under the unified structure, existing reasoning datasets vary in discipline (*e.g.*, mathematics, physics, and chemistry), difficulty level (*e.g.*, primary school, high school, and competition), and task format (*e.g.*, question answering, multiple choice, and text evaluation). In the following part, we introduce several representative subcategories of the reasoning task, along with the popular datasets in each subcategory.

4.2.1 Logical Reasoning

Logical reasoning involves drawing valid conclusions from a set of premises or conditions according to formal logic principles such as deduction, induction, and implication. In the context of LLMs, it assesses whether the model can understand logical relations, maintain logical consistency, and apply rules to reach correct conclusions rather than generating answers based on shallow lexical associations or semantic similarity. To improve and evaluate the logical reasoning ability of LLMs, a large number of datasets have been constructed. For instance, BIG-bench-Hard (Suzgun et al 2023), S59K (Muennighoff et al 2025), LLaVA-CoT (Xu et al 2024b), and MMLU-pro (Wang et al 2024d) are constructed through the strategy of "Augmentation via Generation", while S1K (Muennighoff et al 2025), GPQA Diamond (Rein et al 2023), and MATH 500 (Hendrycks et al 2021a) are constructed mainly through the strategy of "High-value Data Selection" (Section 3.4). These datasets typically comprise various types of logical reasoning problems, such as math, code, Boolean expressions, and logical deduction. Each problem is accompanied by a detailed reasoning trace, typically ranging from 1K to 10K tokens in length. Due to the high cost of human annotation, these long reasoning trajectories are usually generated by LLMs, such as DeepSeek-R1 (Guo et al 2025a), OpenAI o1 (Jaech et al 2024), and Qwen3-235B (Yang et al 2025). As a result, they may contain the noise introduced by model hallucinations, such as factual errors, logical inconsistencies, and redundant or spurious reasoning steps.

4.2.2 Commonsense Reasoning

Commonsense reasoning involves leveraging implicit everyday knowledge and intuitive understanding of the world to make plausible inferences in situations that are not explicitly described. In the context of LLMs, it evaluates whether the model can apply background knowledge, causal reasoning, and social norms to generate coherent and contextually appropriate responses rather than generating answers based on mere pattern matching or factual recall. For example, SWAG (Zellers et al 2018) focuses on grounded commonsense inference, requiring LLMs to predict the plausible continuations of everyday scenarios. Building on relational knowledge, CommonsenseQA (Talmor et al 2019) evaluates whether LLMs can reason the relationships between different objects. For this task, DVD enhancement typically focuses on improving data diversity and the logical coherence of reasoning processes. For instance, KnowLogic (Zhan et al 2025) creates logically consistent and diverse questions through knowledge-driven data synthesis.

4.2.3 Others

Beyond logical and commonsense reasoning, a wide range of other datasets have been developed to evaluate and enhance the reasoning ability of LLMs from different perspectives. For instance, planning reasoning datasets (Valmeekam et al 2023; Xie et al 2024; Zheng et al 2024) require LLMs to decompose complex goals into executable action sequences while satisfying diverse environmental and commonsense constraints. Causal reasoning datasets (Jin et al 2023; Du et al 2022; Wang 2024) require LLMs to provide explainable causal rationales to evaluate whether LLMs can identify underlying causal relationships instead of merely relying on superficial statistical correlations. Moreover, several comprehensive reasoning datasets (Guha et al 2025; Kim et al 2023; Liu et al 2025a) combine multiple types of reasoning problems, aiming to assess the overall reasoning ability of LLMs. These efforts collectively broaden the scope of reasoning enhancement for LLMs.

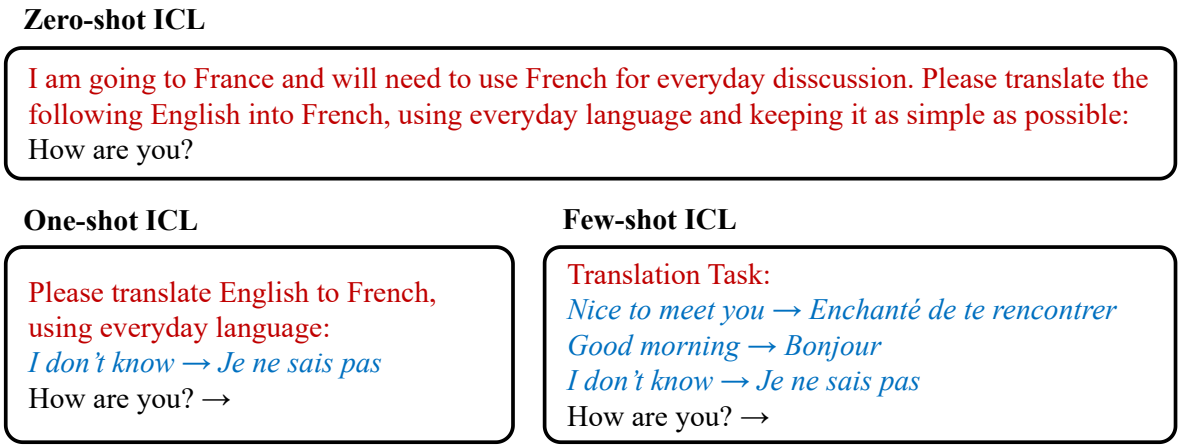


Figure 5: Three types of ICL methods, *i.e.*, zero-shot ICL, one-shot ICL, and few-shot ICL. Among them, zero-shot ICL provides a detailed task description without any specific examples, aiming to give the model a comprehensive understanding of the task requirements. One-shot ICL offers a task description along with a single example which helps clarify the task by illustrating how it can be solved. Few-shot ICL typically presents only multiple examples, allowing the model to understand the specific task requirements by analyzing these examples.

4.3 Vertical Domain Task

While LLMs have achieved remarkable progress on general tasks, their performance in vertical domains remains unsatisfactory. To advance research in this area, a variety of domain-specific datasets have been proposed for fields such as medicine (Zhang et al 2022; Zhu et al 2023; Gu et al 2024b), finance (Islam et al 2023; Guo et al 2025b; Xie et al 2023a; Lu et al 2023; Lei et al 2023b), and law (Dai et al 2025a; Fei et al 2024; Guha et al 2023; Chalkidis et al 2022; Niklaus et al 2023). These datasets are typically employed during the post-training stage to inject domain-specific expertise into LLMs, thereby improving the performance of LLMs on downstream tasks of different vertical domains.

Given the scarcity and limited availability of data in vertical domains, DVD enhancement plays a crucial role in improving the performance of LLMs in vertical domain tasks. Despite its importance, systematic exploration of DVD enhancement strategies in vertical domains, such as data scheduling and data generation, is still in early stages. Designing targeted strategies based on the unique knowledge structures and linguistic characteristics of different vertical domains represents a promising and essential direction for future research.

5 Related Topics

In this section, we explore several research topics related to DVD enhancement and highlight how it interacts with and diverges from these paradigms. We particularly focus on in-context learning, capacity density, sample efficiency, and active learning that share similar optimization objectives or challenges with DVD enhancement for LLM training. This discussion not only clarifies the distinctions among these research paradigms but also reveals their potential complementarities. Understanding these relationships is crucial for effectively leveraging the advantages of DVD enhancement in diverse scenarios and promoting synergy with other strategies. Through these comparative analyses, we aim to position DVD enhancement for LLM training within the broader landscape of artificial intelligence, providing novel opportunities for methodological innovation and practical application.

5.1 In-Context Learning

In-Context Learning (ICL) (Brown et al 2020) is a paradigm that improves the task performance of LLMs by providing additional information in the input context, such as knowledge and experience in solving this kind

of task. This paradigm is typically applied at inference time of LLMs and requires no additional training.

Formally, given a question q, ICL constructs a demonstration set $\mathcal{S}$ that is incorporated into the input context of the LLM, which is:

$$\mathcal{S} = (Ins, (x_1, y_1), \ldots, (x_k, y_k)), \ k \geq 0, \tag{5.1}$$

where Ins, x_k, and y_k denote the task instruction, an example question, and the correct answer of x_k. As shown in Fig. 5, ICL can be categorized into three types based on the definition of Eq. (5.1), which are: zero-shot ICL when $k = 0$, one-shot ICL when $k = 1$, and few-shot ICL when $k > 1$.

Conditioned on $\mathcal{S}$, the LLM is required to generate an answer $\hat{y}$ for q. The objective of ICL is to construct a compact demonstration set $\mathcal{S}$ that maximizes the probability of the LLM generating the correct answer. Intuitively, since both ICL and DVD enhancement for LLM training aim to improve the task performance of LLMs using limited data, ICL can be viewed as a form of DVD enhancement at the reasoning stage of LLMs. Strategies from DVD enhancement for LLM training, such as data synthesis and high-value data selection, can be adapted to ICL. Conversely, the methodologies developed in ICL, such as inter-relationship modeling among examples and reconstruction of example representations, can inspire new insights in DVD enhancement for LLM training.

5.2 Capacity Density

While both Capacity Density (CD) (Xiao et al 2025) and DVD enhancement for LLM training introduce the notion of "density", the evaluation object of the former is the model parameters, whereas the latter evaluates training data. Formally, for an LLM $\mathcal{M}$ with $N_{\mathcal{M}}$ parameters, its task performance is $Acc_{\mathcal{M}}$. The CD of $\mathcal{M}$ is defined as:

$$f_{CD}(\mathcal{M}) = \frac{\hat{N}(Acc_{\mathcal{M}})}{N_{\mathcal{M}}}, \tag{5.2}$$

where $\hat{N}(Acc_{\mathcal{M}})$ denotes the minimum number of parameters required for LLMs to achieve the same task performance $Acc_{\mathcal{M}}$. Notably, CD and DVD evaluate the trade-off between the effectiveness and efficiency of LLM training from the complementary viewpoints of model parameters and training data, respectively. The combination of techniques from these two research areas can yield synergistic effects, enabling the training of capable LLMs under limited data and computational resources. For instance, in the process of transferring knowledge from a teacher LLM to a student LLM, which serves as a strategy to enhance the CD of the student LLM, Guo et al (2025a) improve the quality of knowledge by incorporating DVD enhancement strategies such as data selection and data augmentation. As a result, the resulting 7B-parameter LLM outperforms QwQ-32B-Preview (Team 2024).

5.3 Sample Efficiency

Sample efficiency is a concept that commonly appears in the Reinforcement Learning (RL) setting (D'Oro et al 2022; Wang et al 2025e). One RL method is considered sample-efficient if it can substantially reduce the number of interactions between the model and the environment while maintaining comparable training effectiveness. Formally, a sample-efficient RL method RL_{eff} needs to satisfy:

$$\begin{cases} \mathcal{I}_{RL_{base}} - \mathcal{I}_{RL_{eff}} > \alpha \\ \left| Acc_{RL_{base}} - Acc_{RL_{eff}} \right| < \beta \\ \alpha, \beta > 0 \end{cases}. \tag{5.3}$$

In the above equation, for the strategy $* \in \{RL_{base}, RL_{eff}\}$, $\mathcal{I}_*$ and Acc_* denote the number of interactions between the model and the environment and the task performance of the model after training, respectively. Moreover, α is typically chosen to be a relatively large value to signify a substantial reduction in interactions, whereas β is set to a much smaller value to guarantee minimal performance degradation. At a high level, both DVD enhancement methods and sample-efficient RL methods aim to train strong LLMs with limited

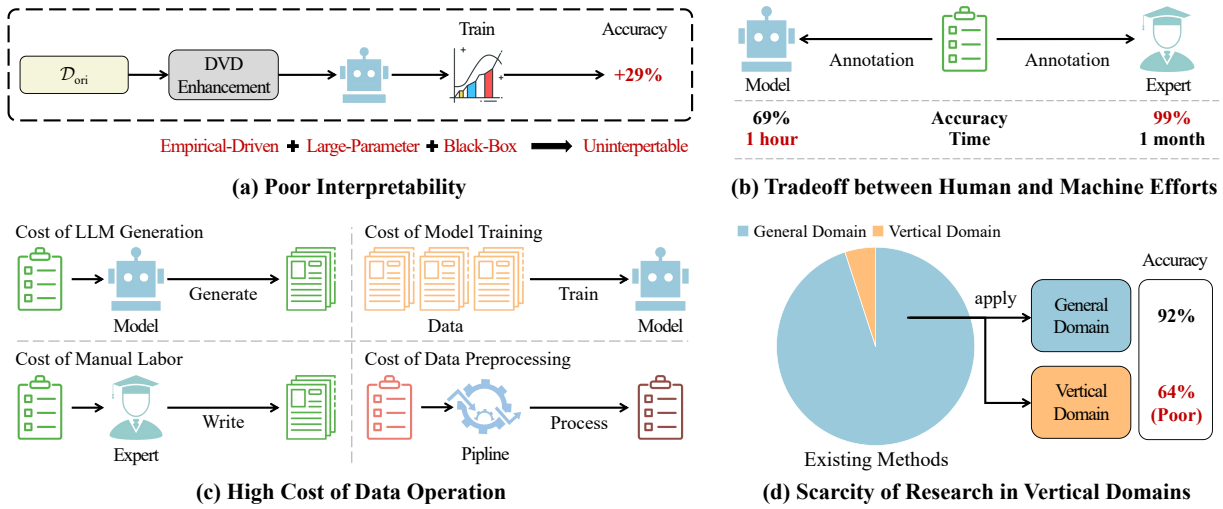


Figure 6: Four main challenges faced by existing research of DVD enhancement for LLM training, *i.e.*, poor interpretability, tradeoff between human and machine efforts, high cost of data operation, and scarcity of research in vertical domains.

training resources. However, to achieve this goal, DVD enhancement for LLM training focuses exclusively on optimizing the training data, while sample-efficient RL emphasizes the optimization of training strategies. For instance, as a representative of the DVD enhancement paradigm, Guo et al (2025a) generate long chain-of-thought reasoning processes for training data to teach LLMs how to perform complex reasoning questions. Conversely, from the perspective of sample-efficient RL, Yu et al (2025) propose the token-level policy gradient loss to improve the learning effectiveness of LLMs on long chain-of-thoughts data.

5.4 Active Learning

Active learning (Sener and Savarese 2018) is typically an iterative process. In each iteration, a subset of data points is selected from an unlabeled data pool based on the characteristics of the trained LLM. Then, these selected data points are annotated and used to train the LLM. The goal of active learning is to design an effective strategy to choose the data points that can maximize the task performance of the LLM from a large pool of unlabeled data. This paradigm is particularly useful in large-scale supervised learning scenarios where obtaining data labels is costly or difficult. Although active learning is rarely discussed directly in the context of LLM training, several of its core ideas have influenced the development of DVD enhancement for LLM training, such as high-value data selection and data scheduling. However, there are key differences between active learning and DVD enhancement for LLM training. The former is mostly designed for the supervised learning paradigm, whereas the latter is applicable to a broader range of model training paradigms. Additionally, active learning requires online data labeling from human at each iteration. Since the scale of the training data for LLMs tends to be large, online data labeling is costly, making active learning less suitable for LLM training.

6 Challenges and Future Directions

As the research on DVD enhancement for LLM training continues to deepen, it has encompassed increasingly diverse tasks, objectives, and strategies. Consequently, the research in this field is likely to encounter various challenges in future exploration and practice. In this section, we discuss potential avenues for future research that could significantly influence the development of DVD enhancement. The goals of this discussion are not only to motivate new methods but also to identify novel research directions that can expand and accelerate this research area in both academic and industrial scenarios. The overview of the main challenges is shown in Fig. 6.

6.1 Poor Interpretability

Although deep learning has driven substantial progress in artificial intelligence, its poor interpretability has long been a major concern. As a methodological branch in deep learning, DVD enhancement for LLM training inherits this issue and even exacerbates it. Specifically, this research area faces two major interpretability challenges. First, LLMs operate as black-box models. Their internal capabilities change throughout training, while the mechanisms underlying such changes remain unknown. Second, existing DVD enhancement methods typically rely on empirical heuristics to model the relationship between the training data and the LLM performance. They lack rigorous theoretical grounding for explaining how data points interact with one another or how data contributes to the development of a particular ability of LLMs. As a consequence, the generalization and stability of current approaches are limited. Even small variations in training context or training configurations may lead to substantial fluctuations in effectiveness. To bridge this gap, future research on DVD enhancement for LLM training should move beyond the current empirical-driven paradigm and develop theoretical frameworks that can formally characterize interactions among training data points and rigorously analyze how data characteristics affect model capabilities. Such theoretical frameworks will facilitate the understanding of the training mechanisms of LLMs and significantly enhance the interpretability of DVD enhancement.

6.2 Tradeoff between Human and Machine Efforts

DVD enhancement typically involves complex data operations such as annotation, generation, and evaluation. These operations are typically carried out through either human effort or machine automation, which exhibit complementary strengths and weaknesses. Specifically, processing data through human effort tends to be of high quality, as domain experts can design complex tasks grounded in professional knowledge, and identify semantic noise, systematic bias, and reasoning errors that machines may fail to detect. Training LLMs on such data enables stable optimization and reliable capability acquisition. As a result, LLMs typically exhibit strong generalization and great robustness after training, especially in complex or high-stakes tasks. Unfortunately, this high-precision process incurs substantial time and labor costs, which limit its feasibility in scenarios of LLM training that demand large-scale data. Differently, machine automation offers automatic and time-efficient data processing, making it suited to LLM training scenarios. However, due to inherent limitations of machines (*e.g.*, insufficient domain knowledge and incomplete training), the quality of data processed by machines cannot be guaranteed, which may contain factual errors, irrelevant content, and other types of noise. When used for training, such data may weaken the learning quality of LLMs, thereby limiting performance gains and increasing the risk of error propagation. Therefore, achieving an appropriate trade-off between human effort and machine automation in the process of data processing is one of the major challenges faced by DVD enhancement. Future research needs to build effective and efficient human–machine collaboration mechanisms. These mechanisms can reasonably allocate the involvement forms and contents of experts and machines during the process of data processing, enabling the time-efficient construction of high-quality data.

6.3 High Implementation Cost

In existing methods of DVD enhancement, achieving higher DVD typically requires additional processing steps, such as generating long chain-of-thought reasoning traces and training auxiliary selectors. As illustrated in Fig. 7, these steps introduce non-negligible cost, which can be grouped into four primary categories:

- **Cost of LLM Generation.** This type of cost typically arises from two processing steps: data generation for quality enhancement and model evaluation for data selection. For the former, prior studies (Muennighoff et al 2025; Li et al 2025a; Guo et al 2025a; Hsieh et al 2023) employ large teacher models to generate or refine training data to improve data quality, such as producing long chain-of-thought reasoning traces and iterative solution refinements, or structured answer formats. For the latter, many methods (Ye et al 2025; Kung et al 2023; Attendu and Corbeil 2023; Li et al 2024c) require testing the target LLM on additional datasets to estimate data utility. To obtain reliable estimations, the additional datasets used are usually large in scale, which requires extensive output sampling from the target LLM. This cost is further amplified in dynamic optimization strategies, where

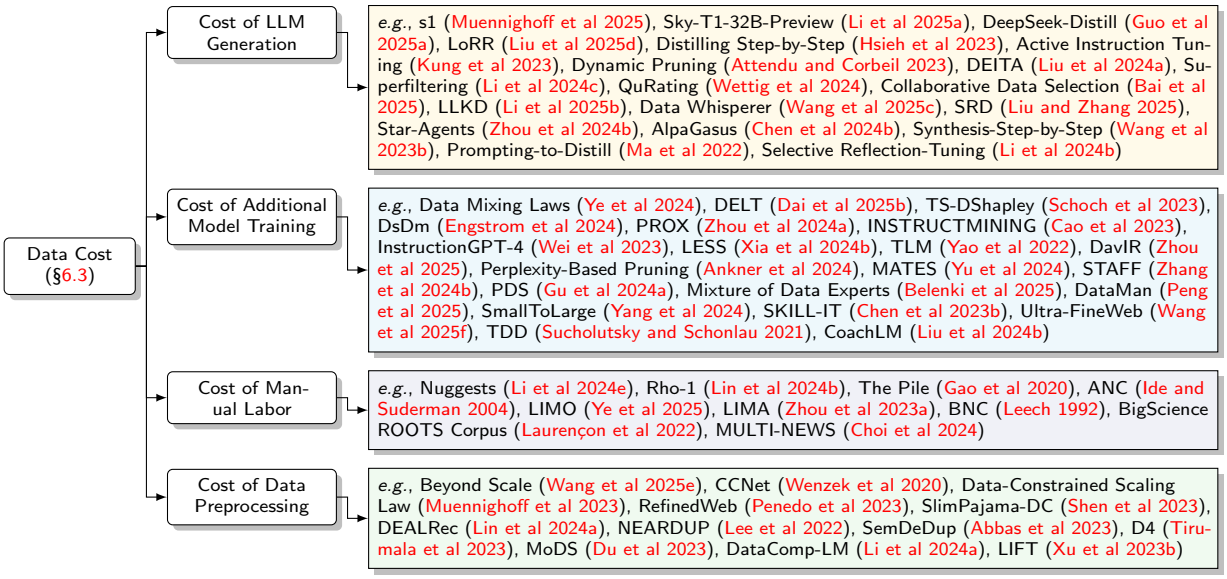


Figure 7: An overview of the main forms of implementation cost introduced by existing methods of DVD enhancement.

repeated evaluations are conducted throughout the training process to track the evolving characteristics of the target LLM.

- **Cost of Additional Model Training.** A mainstream strategy in DVD enhancement is to estimate data value by training the target LLM on each data point and measure the resulting change in training loss or LLM performance (Engstrom et al 2024; Zhou et al 2025; Yu et al 2024; Gu et al 2024a). Under this paradigm, the number of additional training runs required to evaluate the dataset $\mathcal{D}$ typically scales linearly with $\mu(\mathcal{D})$.

- **Cost of Manual Labor.** Since the abilities of current LLMs are limited (*e.g.*, insufficient reasoning ability and lack of domain knowledge), they are unable to reliably support many DVD enhancement strategies, such as verifying domain-specific knowledge and generating answers for vertical-domain tasks. Consequently, substantial expert involvement is required to execute these strategies (Lin et al 2024b; Zhou et al 2023a; Muennighoff et al 2025; Ma et al 2025), resulting in significant human labor costs.

- **Cost of Data Preprocessing.** It typically arises from several essential steps in preparing the pre-training corpora for LLMs, including deduplication, filtering, and privacy removal. As these corpora often span billions or even trillions of tokens, the computational resources required for these processes are enormous. For instance, deduplication and filtering demand substantial storage and processing power. Privacy removal introduces additional complexity, as sensitive information needs to be safely eliminated without compromising data integrity.

Therefore, to obtain a more complete understanding of DVD enhancement, it is necessary to not only focus on the improvement of DVD but also explicitly account for the costs incurred by achieving that improvement. Developing strategies that construct datasets with high DVD at low cost can significantly broaden the applicability of DVD enhancement and enhance its practical value.

6.4 Scarcity of Research in Vertical Domains

Vertical domains, such as healthcare, law, and finance, commonly show unique characteristics, including data scarcity, distinctive linguistic characteristics, and high demands for safety and accuracy. These features increase the difficulty of acquiring a large amount of data. Therefore, DVD enhancement holds particularly high practical value in vertical domains, as it can improve the performance of LLMs under limited training

data. However, the exploration of DVD enhancement in vertical domains is still scarce. Most of the existing methods are designed for general corpora. Since they typically ignore the special data characteristics of vertical domains, their performance is often poor when applied to vertical domains. For example, data cleaning strategies often remove texts with similar semantics under the assumption that such texts contain redundant information. However, in legal domains, two terms exhibiting high semantic similarity may represent entirely different legal meanings. If they are treated as redundant content and are then removed, it may lead to the loss of critical information. Therefore, future research on DVD enhancement for LLM training should place greater emphasis on developing methods for vertical domains. Such methods should carefully consider special data characteristics of vertical domains, thereby advancing the development and application of DVD enhancement for LLM training in various vertical domains.

7 Conclusion

This survey introduces the concept of DVD enhancement for LLM training to characterize the emerging research field of maximizing the training effect of limited data from a data-centric perspective. Based on this concept, we establish a unified framework for this field through a novel taxonomy, marking a critical step toward systematizing its fragmented methodologies. By comprehensively interpreting and organizing state-of-the-art methods, we highlight their design principles and emphasize their potential applications in various scenarios, such as medicine, law, and electronic design automation. In the future, to break through current limitations, research in this field needs to establish a comprehensive theoretical foundation, explore data processing strategies that balance quality and cost, and recognize the distinct properties of vertical-domain data. These advancements will not only improve the effectiveness and efficiency of LLM training, but also significantly broaden the applicability of DVD enhancement for LLM training in both academic and industrial scenarios, particularly in domains where data is scarce. This survey clarifies the development roadmap and core technologies in this field, serving as a catalyst to unlock the full potential of DVD enhancement for LLM training in the era of data-driven artificial intelligence.

References

Abbas AKM, Tirumala K, Simig D, Ganguli S, Morcos AS (2023) Semdedup: Data-efficient learning at web-scale through semantic deduplication. In: ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al (2023) Gpt-4 technical report. arXiv preprint arXiv:230308774

Albalak A, Pan L, Raffel C, Wang WY (2023) Efficient online data mixing for language model pre-training. arXiv preprint arXiv:231202406

Albalak A, Elazar Y, Xie SM, Longpre S, Lambert N, Wang X, Muennighoff N, Hou B, Pan L, Jeong H, Raffel C, Chang S, Hashimoto T, Wang WY (2024) A survey on data selection for language models. Transactions on Machine Learning Research pp 1–81

Alt C, Gabryszak A, Hennig L (2020) Tacred revisited: A thorough evaluation of the tacred relation extraction task. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

An C, Gong S, Zhong M, Zhao X, Li M, Zhang J, Kong L, Qiu X (2024) L-eval: Instituting standardized evaluation for long context language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Ankner Z, Blakeney C, Sreenivasan K, Marion M, Leavitt ML, Paul M (2024) Perplexed by perplexity: Perplexity-based data pruning with small reference models. In: Proceedings of the Thirteenth International Conference on Learning Representations

Arnett C, Jones E, Yamshchikov IP, Langlais PC (2024) Toxicity of the commons: Curating open-source pre-training data. arXiv preprint arXiv:241022587

Attendu Jm, Corbeil JP (2023) Nlu on data diets: Dynamic data subset selection for nlp classification tasks. In: Proceedings of the 40th Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)

Bai T, Yang L, Wong ZH, Sun F, Zhuang X, Peng J, Zhang C, Wu L, Jiantao Q, Zhang W, et al (2025) Efficient pretraining data selection for language models via multi-actor collaboration. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Bai Y, Lv X, Zhang J, Lyu H, Tang J, Huang Z, Du Z, Liu X, Zeng A, Hou L, et al (2023) Longbench: A bilingual, multitask benchmark for long context understanding. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Bai Y, Tu S, Zhang J, Peng H, Wang X, Lv X, Cao S, Xu J, Hou L, Dong Y, et al (2024) Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. arXiv preprint arXiv:241215204

Bai Y, Si S, Luo K, Wang Q, Li W, Chen G, Qi F, Sun M (2026) Infi-check: Interpretable and fine-grained fact-checking of LLMs. arXiv preprint arXiv:260106666

Balunović M, Dekoninck J, Petrov I, Jovanović N, Vechev M (2025) Matharena: Evaluating LLMs on uncontaminated math competitions. arXiv preprint arXiv:250523281

Belenki L, Agarwal A, Shi T, Toutanova K (2025) Optimizing pre-training data mixtures with mixtures of data expert models. arXiv preprint arXiv:250215950

Bendinelli T, Dox A, Holz C (2025) Exploring LLM agents for cleaning tabular machine learning datasets. arXiv preprint arXiv:250306664

Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: Proceedings of the 26th International Conference on Machine Learning

Besta M, Blach N, Kubicek A, Gerstenberger R, Podstawski M, Gianinazzi L, Gajda J, Lehmann T, Niewiadomski H, Nyczyk P, et al (2024) Graph of thoughts: Solving elaborate problems with large language models. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence

Bhatt G, Chen Y, Das A, Zhang J, Truong S, Mussmann S, Zhu Y, Bilmes J, Du S, Jamieson K, et al (2024) An experimental design framework for label-efficient supervised finetuning of large language models. In: Findings of the Association for Computational Linguistics: ACL 2024

Borgeaud S, Mensch A, Hoffmann J, Cai T, Rutherford E, Millican K, Van Den Driessche GB, Lespiau JB, Damoc B, Clark A, et al (2022) Improving language models by retrieving from trillions of tokens. In: Proceedings of the 39th International conference on machine learning

Broder AZ (1997) On the resemblance and containment of documents. In: Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al (2020) Language models are few-shot learners. Advances in Neural Information Processing Systems

Busa-Fekete RI, Zimmert J, Zheng AX, Gentile C, Gyorgy A (2026) Tbdfiltering: Sample-efficient tree-based data filtering. arXiv preprint arXiv:260122016

Cao Y, Kang Y, Wang C, Sun L (2023) Instruction mining: Instruction data selection for tuning large language models. arXiv preprint arXiv:230706290

Chalkidis I, Jana A, Hartung D, Bommarito M, Androutsopoulos I, Katz D, Aletras N (2022) Lexglue: A benchmark dataset for legal language understanding in english. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Chang E, Yeh HS, Demberg V (2021) Does the order of training samples matter? improving neural data-to-text generation with curriculum learning. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume

Chen D, Huang Y, Ma Z, Chen H, Pan X, Ge C, Gao D, Xie Y, Liu Z, Gao J, et al (2024a) Data-juicer: A one-stop data processing system for large language models. In: Proceedings of the 2024 International Conference on Management of Data

Chen H, Zhang Y, Zhang Q, Yang H, Hu X, Ma X, Yanggong Y, Zhao J (2023a) Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning. arXiv preprint arXiv:230509246

Chen L, Li S, Yan J, Wang H, Gunaratna K, Yadav V, Tang Z, Srinivasan V, Zhou T, Huang H, et al (2024b) Alpagasus: Training a better alpaca with fewer data. In: Proceedings of the Twelfth International Conference on Learning Representations

Chen M, Roberts N, Bhatia K, Wang J, Zhang C, Sala F, Ré C (2023b) Skill-it! a data-driven skills framework for understanding and training language models. Advances in Neural Information Processing Systems

Chen S, Wong S, Chen L, Tian Y (2023c) Extending context window of large language models via positional interpolation. arXiv preprint arXiv:230615595

Chen S, Li B, Niu D (2024c) Boosting of thoughts: Trial-and-error problem solving with large language models. In: Proceedings of the Twelfth International Conference on Learning Representations

Chen W, Wang H, Chen J, Zhang Y, Wang H, Li S, Zhou X, Wang WY (2019) Tabfact: A large-scale dataset for table-based fact verification. In: Proceedings of the Eighth International Conference on Learning Representations

Choi J, Yun J, Jin K, Kim Y (2024) Multi-news+: Cost-efficient dataset cleansing via LLM-based data annotation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing

Choi WS, Han DS, Lee H, Park J, Zhang BT (2022) Duel: Adaptive duplicate elimination on working memory for self-supervised learning. arXiv preprint arXiv:221017052

Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, Plappert M, Tworek J, Hilton J, Nakano R, et al (2021) Training verifiers to solve math word problems. arXiv preprint arXiv:211014168

Dai Y, Feng D, Huang J, Jia H, Xie Q, Zhang Y, Han W, Tian W, Wang H (2025a) Laiw: A chinese legal large language models benchmark. In: Proceedings of the 31st International Conference on Computational Linguistics

Dai Y, Huang Y, Zhang X, Wu W, Li C, Lu W, Cao S, Dong L, Li S (2025b) Data efficacy for language model training. arXiv preprint arXiv:250621545

Dasigi P, Lo K, Beltagy I, Cohan A, Smith NA, Gardner M (2021) A dataset of information-seeking questions and answers anchored in research papers. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

Deb R, Thekumparampil KK, Kalantari K, Hiranandani G, Sabach S, Kveton B (2025) Fishersft: Data-efficient supervised fine-tuning of language models using information gain. In: Proceedings of the 42nd International Conference on Machine Learning

Ding J, Ma S, Dong L, Zhang X, Huang S, Wang W, Zheng N, Wei F (2023) Longnet: Scaling transformers to 1,000,000,000 tokens. arXiv preprint arXiv:230702486

D'Oro P, Schwarzer M, Nikishin E, Bacon PL, Bellemare MG, Courville A (2022) Sample-efficient reinforcement learning by breaking the replay ratio barrier. In: Proceedings of the Eleventh International Conference on Learning Representations

Du L, Ding X, Xiong K, Liu T, Qin B (2022) e-care: a new dataset for exploring explainable causal reasoning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Du Q, Zong C, Zhang J (2023) Mods: Model-oriented data selection for instruction tuning. arXiv preprint arXiv:231115653

Dua D, Wang Y, Dasigi P, Stanovsky G, Singh S, Gardner M (2019) Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)

Elgaar M, Amiri H (2026) Curriculum learning for LLM pretraining: An analysis of learning dynamics. arXiv preprint arXiv:260121698

Engstrom L, Feldmann A, Madry A (2024) Dsdm: Model-aware dataset selection with datamodels. In: Proceedings of the 41st International Conference on Machine Learning

Fan S, Grangier D, Ablin P (2024a) Dynamic gradient alignment for online data mixing. arXiv preprint arXiv:241002498

Fan S, Pagliardini M, Jaggi M (2024b) Doge: domain reweighting with generalization estimation. In: Proceedings of the 41st International Conference on Machine Learning

Fayyaz M, Aghazadeh E, Modarressi A, Pilehvar MT, Yaghoobzadeh Y, Kahou SE (2022) Bert on a data diet: Finding important examples by gradient-based pruning. arXiv preprint arXiv:221105610

Fei Z, Shen X, Zhu D, Zhou F, Han Z, Huang A, Zhang S, Chen K, Yin Z, Shen Z, et al (2024) Lawbench: Benchmarking legal knowledge of large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing

Feng J, Meng F, Long C, Cong P, Wang D, Zheng Y, Zhang Y, Gao X, Yuan Y, Ma Y, et al (2025) Jt-safe: Intrinsically enhancing the safety and trustworthiness of LLMs. arXiv preprint arXiv:251017918

Foundation CC (2023) Common crawl. https://commoncrawl.org/

Gadre SY, Smyrnis G, Shankar V, Gururangan S, Wortsman M, Shao R, Mercat J, Fang A, Li J, Keh S, et al (2025) Language models scale reliably with over-training and on downstream tasks. In: Proceedings of the Thirteenth International Conference on Learning Representations

Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, et al (2020) The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:210100027

Gao Z, Kim J, Sun W, Joachims T, Wang S, Pang RY, Tan L (2025) Prompt curriculum learning for efficient LLM post-training. arXiv preprint arXiv:251001135

Ge C, Ma Z, Chen D, Li Y, Ding B (2024) Bimix: A bivariate data mixing law for language model pretraining. arXiv preprint arXiv:240514908

Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Vaughan A, et al (2024) The llama 3 herd of models. arXiv preprint arXiv:240721783

Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) Learning word vectors for 157 languages. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)

Gu Y, Dong L, Wang H, Hao Y, Dong Q, Wei F, Huang M (2024a) Data selection via optimal control for language models. In: Proceedings of the Thirteenth International Conference on Learning Representations

Gu Z, Zhu X, Ye H, Zhang L, Wang J, Zhu Y, Jiang S, Xiong Z, Li Z, Wu W, et al (2024b) Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In: Proceedings of the 38th AAAI conference on artificial intelligence

Guha E, Marten R, Keh S, Raoof N, Smyrnis G, Bansal H, Nezhurina M, Mercat J, Vu T, Sprague Z, et al (2025) Openthoughts: Data recipes for reasoning models. arXiv preprint arXiv:250604178

Guha N, Nyarko J, Ho D, Ré C, Chilton A, Chohlas-Wood A, Peters A, Waldon B, Rockmore D, Zambrano D, et al (2023) Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in Neural Information Processing Systems

Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, Zhu Q, Ma S, Wang P, Bi X, et al (2025a) Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:250112948

Guo W, Yang J, Yang K, Li X, Rao Z, Xu Y, Niu D (2024) Instruction fusion: advancing prompt evolution through hybridization. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Guo X, Xia H, Liu Z, Cao H, Yang Z, Liu Z, Wang S, Niu J, Wang C, Wang Y, et al (2025b) Fineval: A chinese financial domain knowledge evaluation benchmark for 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)

Gyawali B, Anastasiou L, Knoth P (2020) Deduplication of scholarly documents using locality sensitive hashing and word embeddings. In: Proceedings of the Twelfth Language Resources and Evaluation Conference

Hamming RW (1950) Error detecting and error correcting codes. The Bell System Technical Journal 29(2):147–160

Hancock B, Bordes A, Mazare PE, Weston J (2019) Learning from dialogue after deployment: Feed yourself, chatbot! arXiv preprint arXiv:190105415

Hao X, Shen K, Li C (2025) Maga: Massive genre-audience reformulation to pretraining corpus expansion. arXiv preprint arXiv:250204235

He Y, Wang Z, Shen Z, Sun G, Dai Y, Wu Y, Wang H, Li A (2024) Shed: Shapley-based automated dataset refinement for instruction fine-tuning. Advances in Neural Information Processing Systems

He Z, Liang T, Xu J, Liu Q, Chen X, Wang Y, Song L, Yu D, Liang Z, Wang W, et al (2025) Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. arXiv preprint arXiv:250411456

Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J (2021a) Measuring massive multitask language understanding. In: Proceedings of the Ninth International Conference on Learning Representations

Hendrycks D, Burns C, Kadavath S, Arora A, Basart S, Tang E, Song D, Steinhardt J (2021b) Measuring mathematical problem solving with the math dataset. Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)

Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, Casas DdL, Hendricks LA, Welbl J, Clark A, et al (2022) Training compute-optimal large language models. Advances in Neural Information Processing Systems

Hsieh CY, Li CL, Yeh CK, Nakhost H, Fujii Y, Ratner A, Krishna R, Lee CY, Pfister T (2023) Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In: Findings of the Association for Computational Linguistics: ACL 2023

Hu S, Zhou L, Liu S, Chen S, Meng L, Hao H, Pan J, Liu X, Li J, Sivasankaran S, et al (2024) Wavllm: Towards robust and adaptive speech large language model. In: Findings of the Association for Computational Linguistics: EMNLP 2024

Hu Z, Huang X, Mendoza P, Alghamdi EA, Popa RA, Wagner D (2026) Gradshield: Alignment preserving finetuning. In: Proceedings of the Fourteenth International Conference on Learning Representations

Ide N, Suderman K (2004) The american national corpus first release. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation

Islam P, Kannappan A, Kiela D, Qian R, Scherrer N, Vidgen B (2023) Financebench: A new benchmark for financial question answering. arXiv preprint arXiv:231111944

Ivison H, Smith NA, Hajishirzi H, Dasigi P (2023) Data-efficient finetuning using cross-task nearest neighbors. In: Findings of the Association for Computational Linguistics: ACL 2023

Jaech A, Kalai A, Lerer A, Richardson A, El-Kishky A, Low A, Helyar A, Madry A, Beutel A, Carney A, et al (2024) Openai o1 system card. arXiv preprint arXiv:241216720

Jia Y, Zhang C, Diao X, Yuan X, Ouyang Z, Ma C, Vosoughi S (2025) What makes a good curriculum? disentangling the effects of data ordering on LLM mathematical reasoning. arXiv preprint arXiv:251019099

Jin Z, Chen Y, Leeb F, Gresele L, Kamal O, Lyu Z, Blin K, Gonzalez Adauto F, Kleiman-Weiner M, Sachan M, et al (2023) Cladder: Assessing causal reasoning in language models. Advances in Neural Information Processing Systems

Jung J, Jung S (2025) Reasoning steps as curriculum: Using depth of thought as a difficulty signal for tuning LLMs. arXiv preprint arXiv:250818279

Kandpal N, Wallace E, Raffel C (2022) Deduplicating training data mitigates privacy risks in language models. In: Proceedings of the 39th International conference on machine learning

Kang F, Sun Y, Wen B, Chen S, Song D, Mahmood R, Jia R (2024) Autoscale: Scale-aware data mixing for pre-training LLMs. arXiv preprint arXiv:240720177

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020) Scaling laws for neural language models. arXiv preprint arXiv:200108361

Kim J, Lee J (2024) Strategic data ordering: Enhancing large language model performance through curriculum learning. arXiv preprint arXiv:240507490

Kim S, Joo SJ, Kim D, Jang J, Ye S, Shin J, Seo M (2023) The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. arXiv preprint arXiv:230514045

Kočiskỳ T, Schwarz J, Blunsom P, Dyer C, Hermann KM, Melis G, Grefenstette E (2018) The narrativeqa reading comprehension challenge. Transactions of the Association for Computational Linguistics 6:317–328

Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2022) Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems

Korbak T, Shi K, Chen A, Bhalerao R, Buckley CL, Phang J, Bowman SR, Perez E (2023) Pretraining language models with human preferences. In: Proceedings of the 40th International Conference on Machine Learning

Kung PN, Yin F, Wu D, Chang KW, Peng N (2023) Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing

Lai H, Liu X, Iong IL, Yao S, Chen Y, Shen P, Yu H, Zhang H, Zhang X, Dong Y, et al (2024) Autowebglm: Bootstrap and reinforce a large language model-based web navigating agent. CoRR

Laurençon H, Saulnier L, Wang T, Akiki C, Villanova del Moral A, Le Scao T, Von Werra L, Mou C, González Ponferrada E, Nguyen H, et al (2022) The bigscience roots corpus: A 1.6 tb composite multilingual dataset. Advances in Neural Information Processing Systems

Lee BW, Cho H, Yoo KM (2024a) Instruction tuning with human curriculum. In: Findings of the Association for Computational Linguistics: NAACL 2024

Lee K, Ippolito D, Nystrom A, Zhang C, Eck D, Callison-Burch C, Carlini N (2022) Deduplicating training data makes language models better. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Lee N, Wattanawong T, Kim S, Mangalam K, Shen S, Anumanchipalli G, Mahoney M, Keutzer K, Gholami A (2024b) Llm2llm: Boosting LLMs with novel iterative data enhancement. In: Findings of the Association for Computational Linguistics: ACL 2024

Leech GN (1992) 100 million words of english: the british national corpus (bnc). Language Research

Lei B, Liao C, Ding C, et al (2023a) Boosting logical reasoning in large language models through a new framework: The graph of thought. arXiv preprint arXiv:230808614

Lei Y, Li J, Cheng D, Ding Z, Jiang C (2023b) Cfbenchmark: Chinese financial assistant benchmark for large language model. arXiv preprint arXiv:231105812

Li D, Cao S, Griggs T, Liu S, Mo X, Tang E, Hegde S, Hakhamaneshi K, Patil SG, Zaharia M, et al (2025a) Llms can easily learn to reason from demonstrations structure, not content, is what matters! arXiv preprint arXiv:250207374

Li J, Fang A, Smyrnis G, Ivgi M, Jordan M, Gadre SY, Bansal H, Guha E, Keh SS, Arora K, et al (2024a) Datacomp-lm: In search of the next generation of training sets for language models. Advances in Neural Information Processing Systems

Li J, Nag S, Liu H, Tang X, Sarwar SM, Cui L, Gu H, Wang S, He Q, Tang J (2025b) Learning with less: Knowledge distillation from large language models via unlabeled data. In: Findings of the Association for Computational Linguistics: NAACL 2025

Li M, Chen L, Chen J, He S, Gu J, Zhou T (2024b) Selective reflection-tuning: Student-selected data recycling for LLM instruction-tuning. In: Findings of the Association for Computational Linguistics: ACL 2024

Li M, Zhang Y, He S, Li Z, Zhao H, Wang J, Cheng N, Zhou T (2024c) Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Li M, Zhang Y, Li Z, Chen J, Chen L, Cheng N, Wang J, Zhou T, Xiao J (2024d) 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)

Li Y, Li W (2021) Data distillation for text classification. arXiv preprint arXiv:210408448

Li Y, Hui B, Xia X, Yang J, Yang M, Zhang L, Si S, Chen LH, Liu J, Liu T, et al (2024e) One-shot learning as instruction data prospector for large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Li Y, Liu Z, Xing E (2025c) Data mixing optimization for supervised fine-tuning of large language models. arXiv preprint arXiv:250811953

Li Y, Lu T, Li Y, Chen Y, Huang WC, Jiang W, Wang H, Zheng HT, Yu PS (2025d) Teaching according to talents! instruction tuning LLMs with competence-aware curriculum learning. In: Findings of the Association for Computational Linguistics: EMNLP 2025

Li Y, Liu Z, Li Z, Lin Z, Zhang J (2026) Token-level data selection for safe LLM fine-tuning. In: Proceedings of the Fourteenth International Conference on Learning Representations

Lin X, Wang W, Li Y, Yang S, Feng F, Wei Y, Chua TS (2024a) Data-efficient fine-tuning for LLM-based recommendation. In: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval

Lin Z, Gou Z, Gong Y, Liu X, Shen Y, Xu R, Lin C, Yang Y, Jiao J, Duan N, et al (2024b) Rho-1: Not all tokens are what you need. arXiv preprint arXiv:240407965

Lin Z, Lin M, Lin L, Ji R (2025) Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence

Ling W, Yogatama D, Dyer C, Blunsom P (2017) Program induction by rationale generation: Learning to solve and explain algebraic word problems. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Liu C, He S, Liu K, Zhao J (2018) Curriculum learning for natural answer generation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence

Liu C, Wang Z, Shen S, Peng J, Zhang X, Du Z, Wang Y (2025a) The chinese dataset distilled from deepseek-r1-671b

Liu L, Zhang M (2025) Less is more: Selective reflection for compatible and efficient knowledge distillation in large language models. arXiv preprint arXiv:250806135

Liu Q, Zheng X, Muennighoff N, Zeng G, Dou L, Pang T, Jiang J, Lin M (2025b) Regmix: Data mixture as regression for language model pre-training. In: Proceedings of the Thirteenth International Conference on Learning Representations

Liu T, Chen Z, Fang Z, Luo W, Tian M, Liu Z (2025c) Matheval: A comprehensive benchmark for evaluating large language models on mathematical reasoning capabilities. Frontiers of Digital Education 2(2):16–53

Liu W, Zeng W, He K, Jiang Y, He J (2024a) What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In: Proceedings of the Twelfth International Conference on Learning Representations

Liu Y, Tao S, Zhao X, Zhu M, Ma W, Zhu J, Su C, Hou Y, Zhang M, Zhang M, et al (2024b) Coachlm: Automatic instruction revisions improve the data quality in LLM instruction tuning. In: Proceeding of the 2024 IEEE 40th International Conference on Data Engineering

Liu Z, Wang J, Song L, Bian J (2025d) Sample-efficient LLM optimization with reset replay. arXiv preprint arXiv:250806412

Lu D, Wu H, Liang J, Xu Y, He Q, Geng Y, Han M, Xin Y, Xiao Y (2023) Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. arXiv preprint arXiv:230209432

Lu K, Yuan H, Yuan Z, Lin R, Lin J, Tan C, Zhou C, Zhou J (2024) # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In: Porceedings of the Twelfth International Conference on Learning Representations

Luo H, Sun Q, Xu C, Zhao P, Lou JG, Tao C, Geng X, Lin Q, Chen S, Tang Y, et al (2023) Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In: Proceedings of the Thirteenth International Conference on Learning Representations

Luo J, Wu B, Luo X, Xiao Z, Jin Y, Tu RC, Yin N, Wang Y, Yuan J, Ju W, et al (2025a) 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)

Luo K, Ding Z, Weng Z, Qiao L, Zhao M, Li X, Yin D, Shu J (2025b) Let's be self-generated via step by step: A curriculum learning approach to automated reasoning with large language models. In: Findings of the Association for Computational Linguistics: ACL 2025

Luo Z, Xu C, Zhao P, Sun Q, Geng X, Hu W, Tao C, Ma J, Lin Q, Jiang D (2024) Wizardcoder: Empowering code large language models with evol-instruct. In: Proceedings of the Twelfth International Conference on Learning Representations

Luo Z, Zhang X, Liu X, Li H, Gong Y, Chen Q, Cheng P (2025c) Velocitune: A velocity-based dynamic domain reweighting method for continual pre-training. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Lv W, Xia X, Huang SJ (2025) Data-efficient LLM fine-tuning for code generation. arXiv preprint arXiv:250412687

Lyu W, Huang SJ, Xia X (2025) Efficient code LLM training via distribution-consistent and diversity-aware data selection. arXiv preprint arXiv:250702378

Ma R, Wang P, Liu C, Liu X, Chen J, Zhang B, Zhou X, Du N, Li J (2025) S ^{2} r: Teaching LLMs to self-verify and self-correct via reinforcement learning. arXiv preprint arXiv:250212853

Ma X, Wang X, Fang G, Shen Y, Lu W (2022) Prompting to distill: Boosting data-free knowledge distillation via reinforced prompt. In: Proceedings of the 41st International Joint Conference on Artificial Intelligence

Ma Y, Zang Y, Chen L, Chen M, Jiao Y, Li X, Lu X, Liu Z, Ma Y, Dong X, et al (2024) Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. Advances in Neural Information Processing Systems

MAA (2023) American mathematics competition 2023

MAA (2024) American invitational mathematics examination 2024

Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human language technologies

Maekawa A, Kobayashi N, Funakoshi K, Okumura M (2023) Dataset distillation with attention labels for fine-tuning bert. In: Proceedings of the 61st Annual Meeting Of The Association For Computational Linguistics

Maini P, Goyal S, Sam D, Robey A, Savani Y, Jiang Y, Zou A, Fredrikson M, Lipton ZC, Kolter JZ (2025) Safety pretraining: Toward the next generation of safe ai. Advances in Neural Information Processing Systems

Majumdar S, Noroozi V, Samadi M, Narenthiran S, Ficek A, Ahmad W, Huang J, Balam J, Ginsburg B (2025) Genetic instruct: Scaling up synthetic generation of coding instructions for large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)

Manber U, Myers G (1993) Suffix arrays: a new method for on-line string searches. siam Journal on Computing 22(5):935–948

Manku GS, Jain A, Das Sarma A (2007) Detecting near-duplicates for web crawling. In: Proceedings of the 16th international conference on World Wide Web

Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, Gao J (2024) Large language models: A survey. arXiv preprint arXiv:240206196

Miranda B, Lee A, Sundar S, Casasola A, Koyejo S (2023) Beyond scale: The diversity coefficient as a data quality metric for variability in natural language data. arXiv preprint arXiv:230613840

Mitra A, Del Corro L, Zheng G, Mahajan S, Rouhana D, Codas A, Lu Y, Chen Wg, Vrousgos O, Rosset C, et al (2024) Agentinstruct: Toward generative teaching with agentic flows. arXiv preprint arXiv:240703502

Mo K, Shi Y, Weng W, Zhou Z, Liu S, Zhang H, Zeng A (2025) Mid-training of large language models: A survey. arXiv preprint arXiv:251006826

Muennighoff N, Rush A, Barak B, Le Scao T, Tazi N, Piktus A, Pyysalo S, Wolf T, Raffel CA (2023) Scaling data-constrained language models. Advances in Neural Information Processing Systems

Muennighoff N, Yang Z, Shi W, Li XL, Fei-Fei L, Hajishirzi H, Zettlemoyer L, Liang P, Candès E, Hashimoto T (2025) s1: Simple test-time scaling. arXiv preprint arXiv:250119393

Nagatsuka K, Broni-Bediako C, Atsumi M (2023) Length-based curriculum learning for efficient pre-training of language models. New Generation Computing 41(1):109–134

Nguyen D, Li Z, Bateni M, Mirrokni V, Razaviyayn M, Mirzasoleiman B (2025) Synthetic text generation for training large language models via gradient matching. In: Proceedings of the 42nd International Conference on Machine Learning

Nie Y, Williamson M, Bansal M, Kiela D, Weston J (2021) I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)

van Niekerk C, Vukovic R, Ruppik BM, Lin Hc, Gašić M (2025) Post-training large language models via reinforcement learning from self-feedback. arXiv preprint arXiv:250721931

Niklaus J, Matoshi V, Rani P, Galassi A, Stürmer M, Chalkidis I (2023) Lextreme: A multi-lingual and multi-task benchmark for the legal domain. In: Findings of the Association for Computational Linguistics: EMNLP 2023

Oren Y, Sagawa S, Hashimoto T, Liang P (2019) Distributionally robust language modeling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing

Pan Z, Wu Q, Jiang H, Xia M, Luo X, Zhang J, Lin Q, Rühle V, Yang Y, Lin CY, et al (2024) Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In: Findings of the Association for Computational Linguistics: ACL 2024

Park C, Park S, Ahn Y, Kim J, Park J, Lee J (2025) Beyond line-level filtering for the pretraining corpora of LLMs. arXiv preprint arXiv:251024139

Parkar RS, Kim J, Park JI, Kang D (2024) SelectLLM: Can LLMs select important instructions to annotate? arXiv preprint arXiv:240116553

Penedo G, Malartic Q, Hesslow D, Cojocaru R, Cappelli A, Alobeidli H, Pannier B, Almazrouei E, Launay J (2023) The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:230601116

Penedo G, Kydlíček H, Lozhkov A, Mitchell M, Raffel CA, Von Werra L, Wolf T, et al (2024) The fineweb datasets: Decanting the web for the finest text data at scale. Advances in Neural Information Processing Systems

Peng R, Yang K, Zeng Y, Lin J, Liu D, Zhao J (2025) Dataman: Data manager for pre-training large language models. arXiv preprint arXiv:250219363

Platanios EA, Stretcu O, Neubig G, Poczos B, Mitchell T (2019) Competence-based curriculum learning for neural machine translation. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)

Pleiss G, Zhang T, Elenberg E, Weinberger KQ (2020) Identifying mislabeled data using the area under the margin ranking. Advances in Neural Information Processing Systems

Pontiki M, Galanis D, Papageorgiou H, Manandhar S, Androutsopoulos I (2015) SemEval-2015 task 12: Aspect based sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation

Press O, Smith N, Lewis M (2021) Train short, test long: Attention with linear biases enables input length extrapolation. In: Proceedings of the Tenth International Conference on Learning Representations

Qin Z, He Z, Prakriya N, Cong J, Sun Y (2025) Dynamic-width speculative beam decoding for LLM inference. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence

Qiu C, Chen Q, Li J, Wang C, Hua R, Li M, Hu S, Zhang Y (2025) WISDOM: Progressive curriculum synthesis makes LLMs better mathematical reasoner. URL https://openreview.net/forum?id=hFFAg5Dmw9

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021) Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International conference on machine learning

Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, Aslanides J, Henderson S, Ring R, Young S, et al (2021) Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:211211446

Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C (2023) Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research 21(140):1–67

Rajani NF, McCann B, Xiong C, Socher R (2019) Explain yourself! leveraging language models for commonsense reasoning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics

Rein D, Hou BL, Stickland AC, Petty J, Pang RY, Dirani J, Michael J, Bowman SR (2023) Gpqa: A graduate-level google-proof q&a benchmark. In: Proceedings of the 1st Conference on Language Modeling

Ross AS, Hughes MC, Doshi-Velez F (2017) Right for the right reasons: training differentiable models by constraining their explanations. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence

Sachdeva N, Coleman B, Kang WC, Ni J, Hong L, Chi EH, Caverlee J, McAuley J, Cheng DZ (2024) How to train data-efficient LLMs. arXiv preprint arXiv:240209668

Saha S, Levy O, Celikyilmaz A, Bansal M, Weston J, Li X (2024) Branch-solve-merge improves large language model evaluation and generation. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)

Sahni S, Patel H (2023) Exploring multilingual text data distillation. arXiv preprint arXiv:230804982

Schoch S, Mishra R, Ji Y (2023) Data selection for fine-tuning large language models using transferred shapley values. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)

Sebastian Nagel JN (2016) News-crawl. https://github.com/commoncrawl/news-crawl/

Sener O, Savarese S (2018) Active learning for convolutional neural networks: A core-set approach. In: Proceedings of the Sixth International Conference on Learning Representations

Seo Y, Kim G, Kim J, Yeo J (2025) Prior-based noisy text data filtering: Fast and strong alternative for perplexity. arXiv preprint arXiv:250918577

Shaham U, Segal E, Ivgi M, Efrat A, Yoran O, Haviv A, Gupta A, Xiong W, Geva M, Berant J, et al (2022) Scrolls: Standardized comparison over long language sequences. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing

Shaham U, Ivgi M, Efrat A, Berant J, Levy O (2023) Zeroscrolls: A zero-shot benchmark for long text understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2023

Shao Z, Wang P, Zhu Q, Xu R, Song J, Bi X, Zhang H, Zhang M, Li Y, et al (2024) Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:240203300

Shapley L (1953) A value for n-person games. Contributions to the Theory of Games

Shen X, Hu S, Zhang X, Han X, Meng X, Wei J, Liu Z, Sun M (2025) Autoclean: LLMs can prepare their training corpus. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)

Shen Z, Tao T, Ma L, Neiswanger W, Liu Z, Wang H, Tan B, Hestness J, Vassilieva N, Soboleva D, et al (2023) Slimpajama-dc: Understanding data combinations for LLM training. arXiv preprint arXiv:230910818

Shengyu Z, Linfeng D, Xiaoya L, Sen Z, Xiaofei S, Shuhe W, Jiwei L, Hu R, Tianwei Z, Wu F, et al (2023) Instruction tuning for large language models: A survey. arXiv preprint arXiv:230810792

Shi F, Chen X, Misra K, Scales N, Dohan D, Chi EH, Schärli N, Zhou D (2023) Large language models can be easily distracted by irrelevant context. In: Proceedings of the 40th International Conference on Machine Learning

Shi W, Min S, Lomeli M, Zhou C, Li M, Lin XV, Smith NA, Zettlemoyer L, Yih Wt, Lewis M (2024) In-context pretraining: Language modeling beyond document boundaries. In: Proceedings of the Twelfth International Conference on Learning Representations

Silcock E, D'Amico-Wong L, Yang J, Dell M (2021) Noise-robust de-duplication at scale. In: Proceedings of the Eleventh International Conference on Learning Representations

Soldaini L, Kinney R, Bhagia A, Schwenk D, Atkinson D, Authur R, Bogin B, Chandu K, Dumas J, Elazar Y, et al (2024) Dolma: an open corpus of three trillion tokens for language model pretraining research. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, Brown AR, Santoro A, Gupta A, Garriga-Alonso A, et al (2023) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on machine learning research pp 1460–1480

Su D, Hou J, Gao W, Tian Y, Tang B (2024) Dˆ 4: Dataset distillation via disentangled diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Suarez PO, Romary L, Sagot B (2020) A monolingual approach to contextualized word embeddings for mid-resource languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

Sucholutsky I, Schonlau M (2021) Soft-label dataset distillation and text dataset distillation. In: Proceedings of the 2021 International Joint Conference on Neural Networks

Sun H, Liu L, Li J, Wang F, Dong B, Lin R, Huang R (2024a) Conifer: Improving complex constrained instruction-following ability of large language models. arXiv preprint arXiv:240402823

Sun P, Shi B, Yu D, Lin T (2024b) On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

Sun Y, Dong L, Patra B, Ma S, Huang S, Benhaim A, Chaudhary V, Song X, Wei F (2023) A length-extrapolatable transformer. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Sun Y, Shen J, Wang Y, Chen T, Wang Z, Zhou M, Zhang H (2025a) Improving data efficiency for LLM reinforcement fine-tuning through difficulty-targeted online data selection and rollout replay. arXiv preprint arXiv:250605316

Sun Y, Zhang Y, Zhao Z, Wan S, Tao D, Gong C (2025b) Fast-slow-thinking: Complex task solving with large language models. arXiv preprint arXiv:250408690

Sun Y, Zhao Z, Wan S, Gong C (2025c) Cortexdebate: Debating sparsely and equally for multi-agent debate. In: Findings of the Association for Computational Linguistics: ACL 2025

Suzgun M, Scales N, Schärli N, Gehrmann S, Tay Y, Chung HW, Chowdhery A, Le Q, Chi E, Zhou D, et al (2023) Challenging big-bench tasks and whether chain-of-thought can solve them. In: Findings of the Association for Computational Linguistics: ACL 2023

Talmor A, Herzig J, Lourie N, Berant J (2019) Commonsenseqa: A question answering challenge targeting commonsense knowledge. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)

Talmor A, Yoran O, Le Bras R, Bhagavatula C, Goldberg Y, Choi Y, Berant J (2021) Commonsenseqa 2.0: Exposing the limits of ai through gamification. Advances in Neural Information Processing Systems

Tan Q, Xu L, Bing L, Ng HT, Aljunied SM (2022) Revisiting docred-addressing the false negative problem in relation extraction. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing

Tao Y, Kong L, Kan A, Callot L (2024) Textual dataset distillation via language model embedding. In: Findings of the Association for Computational Linguistics: EMNLP 2024

Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, Schalkwyk J, Dai AM, Hauth A, Millican K, et al (2023) Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805

Team G, Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, Sifre L, Rivière M, Kale MS, Love J, et al (2024) Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:240308295

Team Q (2024) Qwq: Reflect deeply on the boundaries of the unknown. https://qwenlm.github.io/blog/qwq-32b-preview

Thangarasa V, Venkatesh G, Lasby M, Sinnadurai N, Lie S (2024) Self-data distillation for recovering quality in pruned large language models. arXiv preprint arXiv:241009982

Tirumala K, Simig D, Aghajanyan A, Morcos A (2023) D4: Improving LLM pretraining via document de-duplication and diversification. Advances in Neural Information Processing Systems

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, et al (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288

Özgür Uğur, Göksu M, Çimen M, Yılmaz M, Şavirdi E, Demir AT, Güllüce R, İclal Çetin, Ömer Can Sağbaş (2026) Mecellem models: Turkish models trained from scratch and continually pre-trained for the legal domain. arXiv preprint arXiv:260116018

Valmeekam K, Marquez M, Olmo A, Sreedharan S, Kambhampati S (2023) Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. Advances in Neural Information Processing Systems

Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M (2022) Will we run out of data? limits of LLM scaling based on human-generated data. arXiv preprint arXiv:221104325

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) Glue: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the Seventh International Conference on Learning Representations

Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman S (2019) Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in Neural Information Processing Systems

Wang A, Song L, Tian Y, Peng B, Yu D, Mi H, Su J, Yu D (2025a) Litesearch: Efficient tree search with dynamic exploration budget for math reasoning. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence

Wang G, Cheng S, Zhan X, Li X, Song S, Liu Y (2024a) Openchat: Advancing open-source language models with mixed-quality data. In: Proceedings of the Twelfth International Conference on Learning Representations

Wang J, Tian C, Chen K, Liu Z, Mao J, Zhao WX, Zhang Z, Zhou J (2026) Mergemix: Optimizing mid-training data mixtures via learnable model merging. arXiv preprint arXiv:260117858

Wang K, Zhu J, Ren M, Liu Z, Li S, Zhang Z, Zhang C, Wu X, Zhan Q, Liu Q, et al (2024b) A survey on data synthesis and augmentation for large language models. arXiv preprint arXiv:241012896

Wang K, Li Z, Cheng ZQ, Khaki S, Sajedi A, Vedantam R, Plataniotis KN, Hauptmann A, You Y (2025b) Emphasizing discriminative features for dataset distillation in complex scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Wang L, Xu W, Lan Y, Hu Z, Lan Y, Lee RKW, Lim EP (2023a) Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Wang R, Zhou W, Sachan M (2023b) Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. In: Findings of the Association for Computational Linguistics: EMNLP 2023

Wang S, Jin X, Wang Z, Wang J, Zhang J, Li K, Wen Z, Li Z, He C, Hu X, et al (2025c) Data whisperer: Efficient data selection for task-specific LLM fine-tuning via few-shot in-context learning. arXiv preprint arXiv:250512212

Wang S, Yang Y, Liu Z, Sun C, Hu X, He C, Zhang L (2025d) Dataset distillation with neural characteristic function: A minmax perspective. In: Proceedings of the Computer Vision and Pattern Recognition Conference

Wang S, Yu L, Gao C, Zheng C, Liu S, Lu R, Dang K, Chen X, Yang J, Zhang Z, et al (2025e) Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. arXiv preprint arXiv:250601939

Wang X, Chen Y, Zhu W (2021) A survey on curriculum learning. IEEE transactions on pattern analysis and machine intelligence 44(9):4555–4576

Wang Y, Liu Y, Shi C, Li H, Chen C, Lu H, Yang Y (2024c) Inscl: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)

Wang Y, Ma X, Zhang G, Ni Y, Chandra A, Guo S, Ren W, Arulraj A, He X, Jiang Z, et al (2024d) Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. Advances in Neural Information Processing Systems

Wang Y, Fu Z, Cai J, Tang P, Lyu H, Fang Y, Zheng Z, Zhou J, Zeng G, Xiao C, et al (2025f) Ultra-fineweb: Efficient data filtering and verification for high-quality LLM training data. arXiv preprint arXiv:250505427

Wang Y, Liu B, Liu F, Guo Y, Deng J, Wu X, Zhou W, Zhou X, Wang T (2025g) Tikmix: Take data influence into dynamic mixture for language model pre-training. arXiv preprint arXiv:250817677

Wang Z (2024) Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In: Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)

Weber M, Fu D, Anthony Q, Oren Y, Adams S, Alexandrov A, Lyu X, Nguyen H, Yao X, Adams V, et al (2024) Redpajama: an open dataset for training large language models. Advances in Neural Information Processing Systems

Wei J, Bosma M, Zhao V, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV (2021) Finetuned language models are zero-shot learners. In: Proceedings of the Tenth International Conference on Learning Representations

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D, et al (2022) Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems

Wei L, Jiang Z, Huang W, Sun L (2023) Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. arXiv preprint arXiv:230812067

Wenzek G, Lachaux MA, Conneau A, Chaudhary V, Guzmán F, Joulin A, Grave É (2020) Ccnet: Extracting high quality monolingual datasets from web crawl data. In: Proceedings of the Twelfth Language Resources and Evaluation Conference

Wettig A, Gupta A, Malik S, Chen D (2024) Qurating: Selecting high-quality data for training language models. In: Proceedings of the 41st International Conference on Machine Learning

Wu S, Lu K, Xu B, Lin J, Su Q, Zhou C (2023) Self-evolved diverse data sampling for efficient instruction tuning. arXiv preprint arXiv:231108182

Wu Y, Rabe MN, Hutchins D, Szegedy C (2022) Memorizing transformers. In: Proceedings of the Tenth International Conference on Learning Representations

Xia M, Gao T, Zeng Z, Chen D (2024a) Sheared llama: Accelerating language model pre-training via structured pruning. In: Proceedings of the Twelfth International Conference on Learning Representations

Xia M, Malladi S, Gururangan S, Arora S, Chen D (2024b) Less: Selecting influential data for targeted instruction tuning. In: Proceedings of the 41st International Conference on Machine Learning

Xiao C, Cai J, Zhao W, Lin B, Zeng G, Zhou J, Zheng Z, Han X, Liu Z, Sun M (2025) Densing law of LLMs. Nature Machine Intelligence pp 1–11

Xie J, Zhang K, Chen J, Zhu T, Lou R, Tian Y, Xiao Y, Su Y (2024) Travelplanner: a benchmark for real-world planning with language agents. In: Proceedings of the 41st International Conference on Machine Learning

Xie Q, Han W, Zhang X, Lai Y, Peng M, Lopez-Lira A, Huang J (2023a) Pixiu: a large language model, instruction data and evaluation benchmark for finance. Advances in Neural Information Processing Systems

Xie SM, Pham H, Dong X, Du N, Liu H, Lu Y, Liang PS, Le QV, Ma T, Yu AW (2023b) Doremi: Optimizing data mixtures speeds up language model pretraining. Advances in Neural Information Processing Systems

Xie SM, Santurkar S, Ma T, Liang PS (2023c) Data selection for language models via importance resampling. Advances in Neural Information Processing Systems

Xu B, Zhang L, Mao Z, Wang Q, Xie H, Zhang Y (2020) Curriculum learning for natural language understanding. In: Proceedings of the 58th annual meeting of the association for computational linguistics

Xu B, Yang A, Lin J, Wang Q, Zhou C, Zhang Y, Mao Z (2023a) Expertprompting: Instructing large language models to be distinguished experts. arXiv preprint arXiv:230514688

Xu C, Sun Q, Zheng K, Geng X, Zhao P, Feng J, Tao C, Lin Q, Jiang D (2024a) Wizardlm: Empowering large pre-trained language models to follow complex instructions. In: Proceedings of the Twelfth International Conference on Learning Representations

Xu G, Jin P, Wu Z, Li H, Song Y, Sun L, Yuan L (2024b) Llava-cot: Let vision language models reason step-by-step. arXiv preprint arXiv:241110440

Xu Y, Yao Y, Huang Y, Qi M, Wang M, Gu B, Sundaresan N (2023b) Rethinking the instruction quality: Lift is what you need. arXiv preprint arXiv:231211508

Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C (2021) mt5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies

Yang A, Li A, Yang B, Zhang B, Hui B, Zheng B, Yu B, Gao C, Huang C, Lv C, et al (2025) Qwen3 technical report. arXiv preprint arXiv:250509388

Yang Y, Mishra S, Chiang J, Mirzasoleiman B (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

Yao X, Zheng Y, Yang X, Yang Z (2022) Nlp from scratch without large-scale pretraining: A simple and efficient framework. In: Proceedings of the 39th International Conference on Machine Learning

Yao Y, Ye D, Li P, Han X, Lin Y, Liu Z, Liu Z, Huang L, Zhou J, Sun M (2019) Docred: A large-scale document-level relation extraction dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Yaswanth M, Singh V, Maheshwari A, Krishna A, Ramakrishnan G (2025) Arise: Iterative rule induction and synthetic data generation for text classification. In: Findings of the Association for Computational Linguistics: NAACL 2025

Ye J, Liu P, Sun T, Zhan J, Zhou Y, Qiu X (2024) Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In: Proceedings of the Thirteenth International Conference on Learning Representations

Ye Y, Huang Z, Xiao Y, Chern E, Xia S, Liu P (2025) Limo: Less is more for reasoning. arXiv preprint arXiv:250203387

Yu Q, Zhang Z, Zhu R, Yuan Y, Zuo X, Yue Y, Dai W, Fan T, Liu G, Liu L, et al (2025) Dapo: An open-source LLM reinforcement learning system at scale. arXiv preprint arXiv:250314476

Yu W, Jiang Z, Dong Y, Feng J (2020) Reclor: A reading comprehension dataset requiring logical reasoning. In: Proceedings of the 8th International Conference on Learning Representations

Yu Z, Das S, Xiong C (2024) Mates: Model-aware data selection for efficient pretraining with data influence models. Advances in Neural Information Processing Systems

Zellers R, Bisk Y, Schwartz R, Choi Y (2018) Swag: A large-scale adversarial dataset for grounded commonsense inference. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing

Zellers R, Holtzman A, Rashkin H, Bisk Y, Farhadi A, Roesner F, Choi Y (2019) Defending against neural fake news. Advances in Neural Information Processing Systems

Zeng W, Ren X, Su T, Wang H, Liao Y, Wang Z, Jiang X, Yang Z, Wang K, Zhang X, et al (2021) Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. arXiv preprint arXiv:210412369

Zeng W, Xu C, Zhao Y, Lou JG, Chen W (2024) Automatic instruction evolving for large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing

Zhan W, Wang Y, Hu N, Xiao L, Ma J, Qin Y, Li Z, Yang Y, Deng S, Ding J, et al (2025) Knowlogic: A benchmark for commonsense reasoning via knowledge-driven data synthesis. arXiv preprint arXiv:250306218

Zhang C, Zhong H, Zhang K, Chai C, Wang R, Zhuang X, Bai T, Jiantao Q, Cao L, Fan J, et al (2024a) Harnessing diversity for important data selection in pretraining large language models. In: Proceedings of the Thirteenth International Conference on Learning Representations

Zhang E, Yan X, Lin W, Zhang T, Qianchun L (2025a) Learning like humans: Advancing LLM reasoning capabilities via adaptive difficulty curriculum learning and expert-guided self-reformulation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing

Zhang J, Liu Y, Zhang CX, Liu Y, Jin YX, Guo LZ, Li YF (2026) Data selection for LLM alignment using fine-grained preferences. In: Proceedings of the Fourteenth International Conference on Learning Representations

Zhang N, Chen M, Bi Z, Liang X, Li L, Shang X, Yin K, Tan C, Xu J, Huang F, et al (2022) Cblue: A chinese biomedical language understanding evaluation benchmark. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Zhang X, Kumar G, Khayrallah H, Murray K, Gwinnup J, Martindale MJ, McNamee P, Duh K, Carpuat M (2018) An empirical exploration of curriculum learning for neural machine translation. arXiv preprint arXiv:181100739

Zhang X, Zhai J, Ma S, Shen C, Li T, Jiang W, Liu Y (2024b) Speculative coreset selection for task-specific fine-tuning. arXiv preprint arXiv:241001296

Zhang Y, Zhong V, Chen D, Angeli G, Manning CD (2017) Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing

Zhang Y, Liu H, Xiao Y, Amoon M, Zhang D, Wang D, Yang S, Quek C (2024c) Llm-enhanced multi-teacher knowledge distillation for modality-incomplete emotion recognition in daily healthcare. IEEE Journal of Biomedical and Health Informatics 29(9):6406–6416

Zhang Y, Sun Y, Zhan Y, Tao D, Tao D, Gong C (2025b) Large language models as an indirect reasoner: Contrapositive and contradiction for automated reasoning. In: Proceedings of the 31st International Conference on Computational Linguistics

Zhao W, Fan H, Hu SX, Zhou W, Lane N (2024a) Clues: Collaborative private-domain high-quality data selection for LLMs via training dynamics. Advances in Neural Information Processing Systems

Zhao Y, Yu B, Hui B, Yu H, Li M, Huang F, Zhang NL, Li Y (2024b) Tree-instruct: A preliminary study of the intrinsic relationship between complexity and alignment. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Zheng HS, Mishra S, Zhang H, Chen X, Chen M, Nova A, Hou L, Cheng HT, Le QV, Chi EH, et al (2024) Natural plan: Benchmarking LLMs on natural language planning. arXiv preprint arXiv:240604520

Zhou C, Liu P, Xu P, Iyer S, Sun J, Mao Y, Ma X, Efrat A, Yu P, Yu L, et al (2023a) Lima: Less is more for alignment. Advances in Neural Information Processing Systems

Zhou D, Schärli N, Hou L, Wei J, Scales N, Wang X, Schuurmans D, Cui C, Bousquet O, Le QV, et al (2023b) Least-to-most prompting enables complex reasoning in large language models. In: Proceedings of the Eleventh International Conference on Learning Representations

Zhou F, Wang Z, Liu Q, Li J, Liu P (2024a) Programming every example: Lifting pre-training data quality like experts at scale. In: Proceedings of the 42nd International Conference on Machine Learning

Zhou H, Tang Y, Qin H, Yang Y, Jin R, Xiong D, Han K, Wang Y (2024b) Star-agents: Automatic data optimization with LLM agents for instruction tuning. Advances in Neural Information Processing Systems

Zhou H, Liu T, Ma Q, Zhang Y, Yuan J, Liu P, You Y, Yang H (2025) Davir: Data selection via implicit reward for large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Zhu W, Wang X, Zheng H, Chen M, Tang B (2023) Promptcblue: A chinese prompt tuning benchmark for the medical domain. arXiv preprint arXiv:231014151

Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision