Abstract

Instruction tuning has emerged as a critical paradigm for improving the capabilities and alignment of large language models (LLMs). However, existing iterative model-aware data selection methods incur significant computational overhead, as they rely on repeatedly performing full-dataset model inference to estimate sample utility for subsequent training iterations, creating a fundamental efficiency bottleneck. In this paper, we propose LEAD, an efficient iterative data selection framework that accurately estimates sample utility entirely within the standard training loop, eliminating the need for costly additional model inference. At its core, LEAD introduces Instance-Level Dynamic Uncertainty (IDU), a theoretically grounded utility function combining instantaneous training loss, gradient-based approximation of loss changes, and exponential smoothing of historical loss signals. To further scale efficiently to large datasets, LEAD employs a two-stage, coarse-to-fine selection strategy, adaptively prioritizing informative clusters through a multi-armed bandit mechanism, followed by precise fine-grained selection of high-utility samples using IDU. Extensive experiments across four diverse benchmarks show that LEAD significantly outperforms state-of-the-art methods, improving average model performance by 6.1%-10.8% while using only 2.5% of the training data and reducing overall training time by 5-10×.

1 Introduction

Instruction tuning has emerged as a powerful paradigm to improve the performance and alignment of large language models (LLMs) by fine-tuning them on instruction-response pairs [2, 11, 32, 51, 52]. Recent studies indicate that data quality, rather than quantity alone, is crucial for substantial performance gains [2, 30, 39, 65]. Consequently, recent research has focused on automatically selecting informative subsets of training data, guided by selection metrics such as data diversity and data quality [4, 10, 46, 58, 63]. However, since these methods do not directly leverage feedback from the model, they fail to dynamically adapt data selection to the model's evolving state and specific learning needs throughout training.

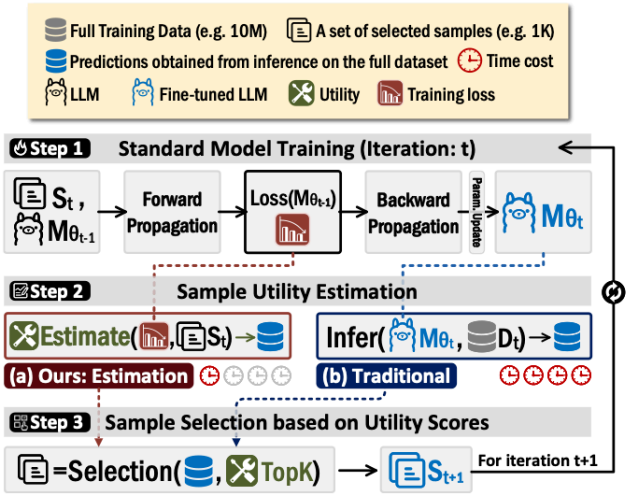

Figure 1: Comparison of Iterative Model-Aware Solutions.

In response, recent efforts have shifted toward *model-aware data selection*, which explicitly utilizes model-derived signals to dynamically identify informative training examples [50, 55]. These model-aware methods broadly fall into two categories: *non-iterative* and *iterative*. Non-iterative methods select data once based on initial model predictions before iterative training [32, 59]. However, since they do not adapt to the model evolvement during training, their effectiveness is inherently limited [60]. In contrast, iterative methods interleave model fine-tuning and data selection across multiple rounds, iteratively choosing new informative samples based on the model's latest feedback [59]. As shown in Figure 1-**Step 2-(b)**, most existing iterative model-aware methods typically rely on explicit model inference to assess the utility of samples. Specifically, after each training iteration, these methods perform inference on *every* sample in the training set to derive feedback signals (*e.g.,* model uncertainty scores) for utility estimation. Although effective at adapting data selection to the model's evolving state, repeatedly performing full-dataset inference significantly increases computational overhead. For example, the recent IFD method [32] spends approximately 98 GPU-hours selecting data from a pool of only $600K$ samples in a single round.

This predicament leads to a natural **research question**: *Can we retain the benefits of iterative model-aware data selection without repeatedly performing costly full-dataset inference?* In other words, can

we effectively determine "select what to learn next" by exclusively utilizing information already computed during standard training, without any additional model inference overhead?

In this work, we posit that the answer is yes. As shown in Figure 1-Step 1, our key insight is that during standard training, the model first conducts a forward propagation step using the current mini-batch of samples, computes the per-sample losses based on its predictions, and subsequently updates its parameters via backward propagation. Crucially, this training process naturally produces a per-sample loss for each training instance in the mini-batch. Intuitively, this loss indicates how challenging a sample is for the model—higher losses reflect greater difficulty and thus greater potential informativeness for future learning. Hence, these training-time losses inherently serve as valuable indicators of a sample's utility. Indeed, they provide an effective proxy for explicit utility metrics (*e.g.,* model uncertainty) typically obtained through costly, separate inference steps [22].

If we can cleverly harness these inherent training signals across the whole dataset, we could **estimate** the utility of each sample **without additional inference (inference-free)** (see Figure 1-Step 2-(a)). This idea – leveraging training-time loss signals to guide data selection – offers the potential to eliminate the full-dataset inference stage while still adapting to the model's training state.

Challenges. Realizing this idea in practice is non-trivial.

First, although using training-time losses allows us to avoid explicit inference, a subtle yet fundamental issue arises due to a timing misalignment. Specifically, as shown in Figure 1-Step 1, the training loss observed at iteration t reflects the model's performance *before* updating parameters (model state $M_{\theta_{t-1}}$), whereas the utility of selecting samples ideally should consider their usefulness *after* the parameter update (*i.e.,* M_{θ_t} at iteration $t+1$). This temporal mismatch means that naively reusing pre-update loss signals may not accurately reflect true sample utility after the next parameter update. We term this issue as the *temporal mismatch* challenge (**C1**).

Second, raw loss signals can be noisy or unstable – they fluctuate from one update to the next due to randomness (*e.g.,* varying batch composition) and the non-stationary nature of training, thus naively trusting instantaneous loss values might lead to suboptimal choices. This issue highlights the *instability of loss signals* challenge (**C2**).

Third, even if we successfully eliminate separate inference steps, individually estimating utility and selecting informative samples remains inefficient for large-scale datasets (*e.g.,* containing millions of samples). We refer to this as the *sample-level selection efficiency challenge* (**C3**). Thus, we need an effective mechanism that can rapidly narrow down candidate samples while prioritizing those most likely to substantially improve the model.

Our Methodology: Iterative Data Selection with Inference-free Utility Estimation. To address the above challenges, we propose LEAD, a theoretically-grounded iterative data selection framework that integrates seamlessly into the model training loop, accurately estimating sample utility without incurring additional inference overhead. The core theoretical insight behind inference-free yet accurate utility estimation lies in effectively addressing two critical challenges: (**C1**) the temporal mismatch between loss computation and parameter updates, and (**C2**) the inherent instability of instantaneous loss signals.

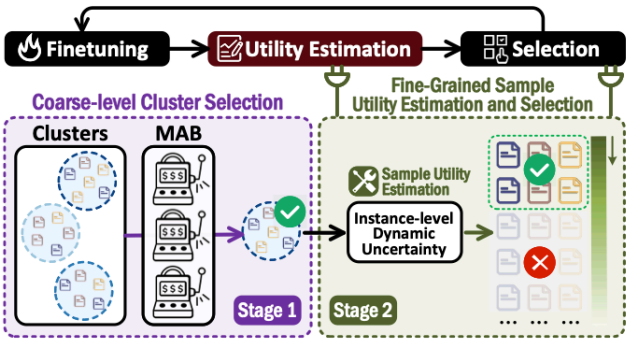

Figure 2: A High-level Overview of LEAD.

To achieve this, we propose a novel sample utility estimation function called *Instance-Level Dynamic Uncertainty* (**IDU**). IDU explicitly implements the `Estimate` step depicted in Figure 1-Step 2-(a) by combining three naturally available training signals: (1) the current training loss for each sample, (2) gradient-based approximation, derived from gradient correlation approximations, to anticipate loss changes at the next parameter update (addressing **C1**), and (3) historical loss trends via exponential smoothing to reduce random noise and improve stability (addressing **C2**). Importantly, IDU is computed entirely using training-time signals naturally available during model updates (losses and logits), thus incurring no additional inference overhead. Finally, we conduct a Lagrangian function and utilize complementary slackness conditions to rigorously derive optimal parameters for IDU, ensuring both theoretical soundness and practical effectiveness.

Guided by this theoretical foundation, our LEAD framework employs a practical coarse-to-fine data selection strategy (Figure 2).

Stage 1: Coarse-level Cluster Selection. Recall our third challenge (**C3**) – efficient candidate selection at scale. To address this, we first partition the dataset offline into clusters based on two widely-used metrics: (1) *instruction-following difficulty*, measuring how challenging each instruction is for the model [32], and (2) *task-level similarity*, grouping semantically related instructions [34]. This clustering step is performed only once per dataset. During training, LEAD employs a multi-armed bandit (MAB) algorithm [54] to dynamically identify and prioritize clusters likely to yield higher rewards – clusters containing samples with greater potential to significantly enhance the model's performance (addressing **C3**).

Stage 2: Fine-Grained Sample Utility Estimation and Selection. Within each selected cluster, LEAD utilizes the IDU function to estimate the utility of individual samples precisely. Specifically, given the IDU scores computed based on the previously discussed training signals (losses, historical trends, and gradient predictions), LEAD prioritizes and selects samples with the highest IDU values. Therefore, samples predicted to yield higher improvements for the model after subsequent parameter updates are selected preferentially.

Contributions. This paper makes the following contributions:

(1) Problem Formulation. We formally introduce the problem of Iterative Data Selection with Inference-Free Utility Estimation, defining a scenario where iterative model-aware selection is performed without incurring additional inference overhead (Section 2).

(2) Instance-Level Dynamic Uncertainty (IDU). We develop a new sample utility estimation function, IDU, which effectively addresses temporal mismatch and instability in loss signals by integrating current losses, gradient-based approximations of loss changes, and exponential smoothing of historical loss signals. All components are computed directly from naturally available training signals without requiring additional model inference (Section 3).

(3) LEAD Framework. We propose LEAD, a theoretically grounded and efficient iterative data selection framework seamlessly integrated into the standard model training process, eliminating repeated costly inference steps (Section 4 and Section 5).

(4) Theoretical Analysis. We rigorously ground our framework in a Lagrangian optimization formulation, employing complementary slackness conditions and gradient correlation approximations to derive theoretically optimal parameters for the IDU function, ensuring both soundness and practical effectiveness (Section 6).

(5) Extensive Experiments. Extensive experiments across four diverse benchmarks show that LEAD significantly outperforms state-of-the-art methods, improving average model performance by 6.1%-10.8% while using only 2.5% of the training data and reducing overall training time by 5-10× (Section 7).

2 Preliminary and Problem Formulation

2.1 Instruction Tuning for LLMs

Instruction tuning fine-tunes pretrained large language models using instruction-response pairs, enabling them to generalize to new tasks by interpreting diverse instructions [56]. Formally, given instruction-response pairs (x, y) from dataset $\mathcal{D}$, instruction tuning optimizes model θ by minimizing the expected loss:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[L(\mathcal{M}_\theta(x), y) \right] \tag{1}$$

where L is a task-specific loss function such as cross-entropy.

2.2 Data Selection for Instruction Tuning

In practice, datasets often originate from vast and noisy sources. Given limited computational budgets and data quality concerns, selecting the most informative samples for instruction tuning becomes crucial. We formalize this as the data selection problem, categorized into two groups: static and iterative data selection.

Static Data Selection for Instruction Tuning. Given a dataset $\mathcal{D}$, it selects a fixed subset $\mathcal{D}^* \subseteq \mathcal{D}$ under budget constraint B:

$$\min_{\mathcal{D}^*\subseteq\mathcal{D},\ |\mathcal{D}^*|\le B} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{target}}} \left[L(\mathcal{M}_\theta(x), y) \right], \tag{2}$$

where $\mathcal{D}_{\text{target}}$ denotes the target distribution. However, static methods cannot adaptively select samples based on the model's evolving capabilities to maximize learning effectiveness during training [2].

Iterative Data Selection for Instruction Tuning. Iterative data selection interleaves model fine-tuning and data selection across multiple iterations. Formally, given the model parameters θ_t at iteration t, we adaptively select a subset $S_t \subseteq \mathcal{D}$ based on a utility function $f(\theta_t, x)$, which estimates the expected contribution of each sample x to future model improvement (*e.g.,* loss reduction). The iterative selection problem can thus be formulated as:

$$\max_{\{S_1,\ldots,S_T\}} \sum_{t=1}^{T} \sum_{x\in S_t} f_t(\theta_t, x), \quad \text{s.t.} \quad \sum_{t=1}^{T} |S_t| \le B, \tag{3}$$

where B is the total sample selection budget allowed during training.

Existing methods typically estimate the utility $f_t(\theta_t, x)$ by performing full-dataset inference at each iteration. Specifically, after fine-tuning the model on selected samples S_t, traditional methods explicitly run inference on the entire dataset $\mathcal{D}$ using the updated model parameters θ_t to compute utility scores:

$$f_t(\theta_t, x) = g(\text{Infer}(\theta_t, x)), \quad \forall x \in \mathcal{D}, \tag{4}$$

where $\text{Infer}(\theta_t, x)$ denotes inference (*e.g.,* loss or uncertainty computation) and $g(\cdot)$ maps inference results to utility values.

Consequently, the next subset S_{t+1} is selected as:

$$S_{t+1} = \arg\max_{S_t\subseteq\mathcal{D},\ |S_t|\le k} \sum_{x\in\mathcal{D}} f_t(\theta_t, x), \quad \text{s.t.} \quad |S_t| \le k, \quad T\cdot k \le B. \tag{5}$$

Note that in iterative data selection, we typically assume a fixed selection size k per iteration, constrained by the total selection budget B. Thus, the number of iterations T and the selection size per iteration k satisfy the relation $T \cdot k \le B$.

2.3 Problem Formulation

Existing iterative model-aware methods rely heavily on repeated full-dataset inference for sample utility estimation, leading to significant computational costs. To eliminate this, we define the problem of *Iterative Data Selection with Inference-Free Utility Estimation.*

*Definition 2.1 (**Iterative Data Selection with Inference-Free Utility Estimation**). Given a total sample selection budget B, our objective is to identify subsets $\{S_t\}_{t=1}^{T}$ that maximize the cumulative estimated utility, where the utility function $f_t(\theta_{t-1}, x)$ is computed exclusively from training-time signals (e.g., training losses or logits) without incurring additional inference overhead:*

$$\max_{\{S_1,\ldots,S_T\}} \sum_{t=1}^{T} \sum_{x\in S_t} f_t(\theta_{t-1}, x), \quad \text{s.t.} \quad \sum_{t=1}^{T} |S_t| \le B, \tag{6}$$

Specifically, at each iteration t, the utility estimation $f_t(\theta_{t-1}, x)$ utilizes the loss signal computed using model parameters θ_{t-1} immediately after the forward propagation step, but before the backward propagation (parameter update). Thus, no additional inference is required to estimate utilities for data selection at iteration t.

Our goal, therefore, is to design accurate and stable inference-free utility estimation methods. For simplicity, we use $f_t(\theta_{t-1}, x)$ and $f(\theta_{t-1}, x)$ interchangeably when the context clearly refers to data selection at iteration t.

3 Instance-Level Dynamic Uncertainty Utility

Designing an effective inference-free utility function $f(\theta_{t-1}, x)$ requires addressing two fundamental challenges as discussed in Section 1: **(C1)** the temporal mismatch between pre-update loss signals and their actual post-update utility, and **(C2)** the instability of instantaneous loss signals due to random fluctuations and noise.

To tackle these challenges, we first define a baseline utility function based on a *loss-based uncertainty metric*, and then introduce an

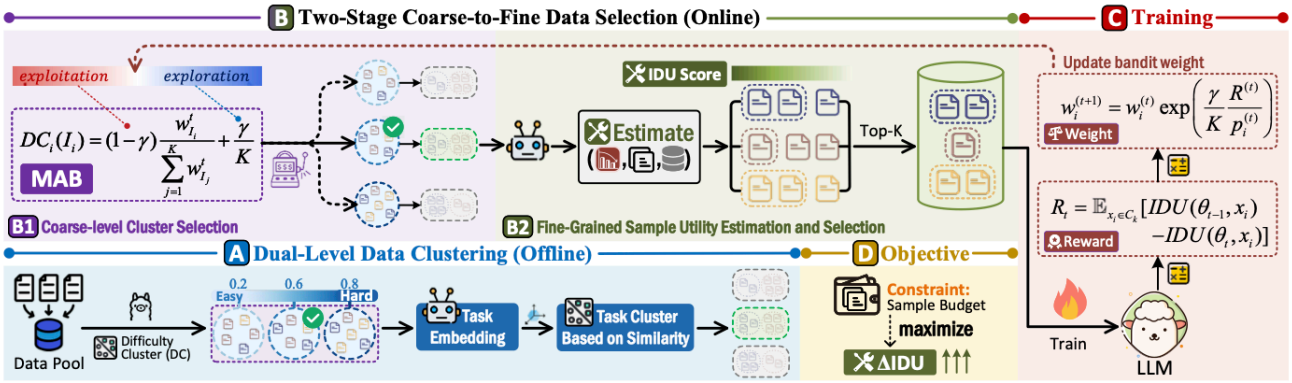

Figure 3: An Overview of the LEAD Framework.

improved formulation, termed *Instance-Level Dynamic Uncertainty (IDU)* utility function, which explicitly addresses these limitations.

Loss-based Uncertainty Estimation. Specifically, our approach begins by formalizing Instance-level uncertainty through a loss-based formulation. Formally, given an instruction-response pair (x, y), we define the Instance-level Uncertainty (IU) [20] at training iteration t as the empirical cross-entropy between the model's current predictive distribution and the ground-truth response:

$$IU(\theta_t, y|x) = L(\theta_t, x) = -\frac{1}{T}\sum_{j=1}^{T} \log p_{\theta_t}(t_j^y | x, t_1^y, \ldots, t_{j-1}^y), \quad (7)$$

where T is response length, t_j^y refers to the j-th response token, and p_{θ_t} the model's token-level predictive probability distribution.

IU naturally corresponds to the training-time negative log-likelihood loss, providing a direct and computationally free baseline. However, IU alone cannot effectively handle challenges **(C1)** and **(C2)**.

Instance-Level Dynamic Uncertainty (IDU). To explicitly mitigate both temporal mismatch **(C1)** and instability **(C2)** of loss signals, we introduce the Instance-Level Dynamic Uncertainty (IDU), which incorporates exponential smoothing of historical losses and gradient-based approximation of loss changes. Formally, given subset S_t at iteration t, IDU for sample x is recursively defined as:

$$\begin{aligned} f(\theta_{t-1}, x) &= IDU(\theta_{t-1}, x) \\ &= (1-b)\cdot \underbrace{[\underbrace{L(\theta_{t-1}, x)}_{\text{IU at } \theta_{t-1}} + \underbrace{\Delta L'(\theta_t, x)}_{\text{Utility Change}}]}_{\textbf{Estimated} \text{ Utility at } \theta_t} + b \cdot \underbrace{IDU(\theta_{t-2}, x)}_{\text{Historical Utility}}, \end{aligned} \quad (8)$$

where $b \in [0, 1)$ controls the balance between current and historical signals, $L(\theta_{t-1}, x)$ is the IU computed using model parameters θ_{t-1}, and $\Delta L'(\theta_t, x)$ is an approximation of the expected utility change, defined as: $\Delta L'(\theta_t, x) = L(\theta_t, x) - L(\theta_{t-1}, x)$.

We have the following key clarifications regarding Eq. (8):

- The instantaneous loss $L(\theta_{t-1}, x)$ is computed naturally during forward propagation at iteration t, requiring no extra inference.
- The $\Delta L'(\theta_t, x)$ denotes the anticipated loss change from θ_{t-1} to θ_t. Importantly, this estimation leverages only readily available gradient and historical loss information collected at iteration $t-1$, ensuring no extra inference is performed at iteration t.

IDU effectively resolves both fundamental challenges through two carefully designed components:

- **Utility Change Estimation (Gradient-Based approximation).** To address temporal mismatch **(C1)**, IDU explicitly estimates the expected utility change ($\Delta L'(\theta_t, x)$) between consecutive iterations. Instead of performing additional inference passes with updated parameters (θ_t), we leverage gradient-based approximations derived from backward propagation at iteration $t-1$ to estimate the loss at iteration t.
- **Historical Utility (Exponential Smoothing).** To tackle instability **(C2)**, IDU incorporates historical uncertainty signals using an exponential smoothing mechanism. Rather than depending solely on instantaneous IU values, IDU maintains an exponential moving average of previous utility estimates ($IDU(\theta_{t-2}, x)$). This significantly reduces fluctuations caused by random noise and local minima encountered during training.

We will elaborate on the details of computing IDU and optimizing the coefficient b of the IDU utility function in Section 5.1.

4 LEAD: LEArning to Iteratively Select Data

We first present an overview of LEAD (Section 4.1), followed by the three key components enabling inference-free iterative data selection (Section 4.2). Finally, we describe how these components systematically interact during iterative training (Section 4.3).

4.1 LEAD Framework: An Overview

Figure 3 provides a high-level overview of LEAD, illustrating its coarse-to-fine approach guided by a theoretically grounded IDU utility function. The framework comprises two key phases: offline dual-level clustering and online adaptive selection.

Dual-Level Data Clustering (Offline). As shown in Figure 3-(A), we first perform an offline preprocessing step to systematically partition the dataset into clusters based on two complementary dimensions: instruction-following difficulty [32] and task similarity [34]. This dual-level clustering is conducted offline, incurring no additional computational overhead during online training.

(1) Difficulty-aware Instance-level Clustering. We use the Instruction-Following Difficulty (IFD) metric [32] to evaluate instance-level difficulty. Given an instruction-response pair (x, y), the IFD is computed as: $IFD(y \mid x) = \frac{PPL(y|x)}{PPL(y)}$, where $PPL(y \mid x)$ and $PPL(y)$ denote the perplexities of generating the y with and without the x, respectively. Using these IFD scores, we group training samples into clusters through sliding intervals (*e.g.,* intervals of 0.1).

(2) Similarity-based Task-level Clustering. Within each difficulty cluster, we further conduct finer-grained clustering based on task similarity. Specifically, we extract task-specific embeddings from instructions by emphasizing task-defining terms (*e.g.,* key verbs and nouns), following the approach in [34]. We then apply the K-means algorithm [43] to group instructions by task similarity.

Coarse-to-Fine Data Selection (Online). During the training, as shown in Figure 3-(B), LEAD implements a coarse-to-fine selection process designed to maximize utility and training effectiveness under a given total sample budget.

(1) Coarse-Level Cluster Selection (via MAB). At each training iteration t, we first employ a Multi-Armed Bandit (MAB) algorithm (specifically EXP3, detailed in Section 5.2) to dynamically select one difficulty-level cluster that is most beneficial to the current model state. The MAB algorithm leverages a self-guided IDU-based reward signal, directly measuring the reduction in IDU scores derived from training on previously selected clusters.

(2) Fine-Grained Sample Selection (via IDU). After identifying the optimal difficulty-level cluster, we distribute the selection budget across its finer-grained task clusters. Specifically, we select the most informative samples from each task cluster based on their current IDU values (see Section 5.1), thus ensuring efficient fine-grained selection of training data at iteration t.

These selected samples form the subset S_t used to fine-tune the model at iteration t. After training, the model parameters are updated from θ_{t-1} to θ_t, and the MAB rewards are updated accordingly, ensuring the LEAD framework continuously improves its data selection strategy.

4.2 LEAD Framework: Core Components

LEAD has three carefully designed core components.

(1) Instance-Level Dynamic Uncertainty (IDU) Utility. To estimate sample utility efficiently without additional inference, we introduce the Instance-Level Dynamic Uncertainty (IDU) metric. IDU combines exponential smoothing of historical losses and a gradient-based approximation of loss change, effectively addressing the temporal instability and inference overhead challenges inherent in traditional iterative selection methods (see Section 5.1).

(2) Adaptive Data Selection via MAB-Integrated Training Scheduler. To integrate coarse and fine-grained selections seamlessly, we employ the MAB-EXP3 algorithm to dynamically balance exploration and exploitation among clusters. The MAB scheduler dynamically prioritizes clusters demonstrating higher historical utility gains, thus efficiently adapting to the model's evolving learning capabilities (further described in Section 5.2).

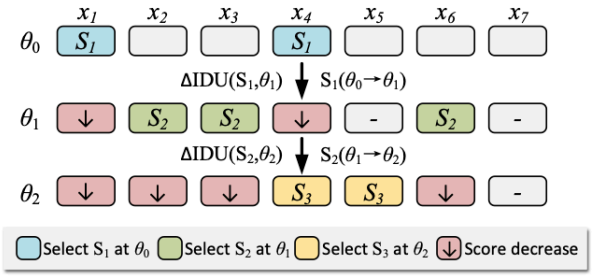

Figure 4: Iterative Sample Selection Guided by IDU Scores.

(3) Self-Guided IDU-Based Reward. To guide the coarse-level cluster selection via MAB, we propose a novel reward function based on the reduction of IDU achieved by training on a given cluster without the need for external validation steps and additional inference (Please refer to Section 5.3 for details).

Next, we illustrate how these components interact seamlessly in the iterative training workflow.

4.3 Training Iteration Workflow of LEAD

The LEAD integrates iterative data selection with LLM instruction tuning. Each training iteration t within LEAD comprises four steps.

Step 1: Difficulty-Aware Cluster Selection. Select the optimal coarse-level difficulty cluster C_{i^*} via the MAB-EXP3 algorithm, guided by the reward derived from previous training iterations, reflecting the cluster's historical effectiveness.

Step 2: Fine-Grained Sample Selection. Within the cluster C_{i^*}, utilize the IDU function to select the top n_{i^*} most informative samples. These samples form the training subset S_t. For example, in Figure 4, at iteration θ_0, samples with the highest initial IDU scores (labeled as S_1) are chosen for training.

Step 3: LLM Instruction Tuning. The selected samples (S_t) are used to fine-tune the model parameters, transitioning from the current parameters θ_{t-1} to the updated parameters θ_t.

Step 4: Reward and Utility Updates. After fine-tuning, trained samples typically show decreased IDU scores, reflecting reduced informativeness. This reduction serves as the training reward. As shown in Figure 4, lowered IDU scores of previously selected samples (*e.g.,* S_1 at θ_0 and S_2 at θ_1) prompt dynamic selection of new, more informative samples for subsequent iterations (*e.g.,* S_2 to S_3). Finally, both IDU scores and the MAB weights are updated accordingly, guiding the sample selection process in future iterations

Through this structured workflow, LEAD continuously and adaptively selects the most beneficial samples at each training step.

5 The Design Details of LEAD

We first show how to optimize our IDU utility under a budget constraint (Section 5.1), followed by an adaptive data selection scheduler via MAB algorithms (Section 5.2), and finally, a self-guided IDU-based reward for cluster evaluation (Section 5.3).

5.1 Instance-Level Dynamic Uncertainty Optimization under the Budget Constraint

In Section 3, we introduced the *IDU* utility (Eq. (8)) for estimating sample utilities in iterative data selection. Note that our LEAD aims to iteratively select subsets of samples with the highest cumulative utility gain, defined as the expected reduction in average *IDU* at each iteration (ΔIDU_t) under a total budget constraint B. Formally, our optimization problem can be defined as follows.

Problem 1 (Budget-Constrained IDU Utility Optimization). *Given a total selection budget B, our goal is to maximize the cumulative expected utility over T training iterations:*

$$\max_{b,T} \sum_{t=1}^{T} \mathbb{E}[\Delta IDU_t], \quad s.t. \sum_{t=1}^{T} \mathbb{E}[n_t] \leq B \tag{9}$$

$$where \quad \mathbb{E}[n_t] = \alpha \cdot (1-b) \cdot |\overline{C}| \cdot (1 + CV^2) \cdot (1 + O(\gamma)) \tag{10}$$

Here, n_t denotes the number of samples selected at iteration t, α is the sampling ratio, $b \in [0,1)$ is the smoothing parameter controlling the influence of historical utility, $|\overline{C}|$ is the average cluster size, and $CV^2 = \frac{1}{K}\sum_{i=1}^{K} \frac{(|C_i| - |\overline{C}|)^2}{|\overline{C}|^2}$ quantifies variability among cluster sizes.

To solve this problem, we construct a Lagrangian function incorporating the budget constraint and apply the complementary slackness condition to derive the optimal smoothing parameter b^*. Specifically, the optimal smoothing coefficient b^* that maximizes cumulative utility gain under the budget constraint is given by: $b^* = 1 - \frac{B}{\alpha \cdot |\overline{C}| \cdot T \cdot (1+CV^2)}$. The detailed derivation and theoretical justification of b^* are provided in Theorem 6.1 (Section 6).

In practice, to effectively implement the optimal solution to our budget-constrained utility maximization problem, we first derive the optimal smoothing coefficient b^* from the theoretical analysis above. However, to fully instantiate our IDU utility function, we must also efficiently estimate the utility changes ($\Delta L'(\theta_t, S_t)$) between consecutive training iterations, as this term directly contributes to computing the cumulative utility gain ΔIDU_t. Directly calculating these utility changes would typically require additional inference steps, violating our zero-cost constraint.

To address this, we introduce the gradient-based approximation of utility change, as discussed below.

Gradient-Based Approximation of Utility Change. Our approach efficiently utilizes gradient information computed during standard model training, thus requiring no extra computational resources beyond regular forward-backward propagation.

Formally, consider a subset of samples S_i. When model parameters are updated from θ_{t-1} to θ_t, the average uncertainty change (utility change) $\Delta L(\theta_t, S_i)$ can be approximated as follows:

Theorem 5.1 (Utility Change Approximation). *For a given sample subset S_i, the utility change from parameter update θ_{t-1} to θ_t can be approximated as:*

$$\begin{aligned}\Delta L'(\theta_t, S_i) &\equiv \frac{1}{|S_i|} \sum_{x \in S_i} (L(\theta_t, x) - L(\theta_{t-1}, x)) \\ &\approx -\eta \left[\beta^2 \delta_{t_k} + (1-\beta)^2 \delta_{t-1} + 2\beta(1-\beta)\sqrt{\delta_{t_k}\delta_{t-1}} \cos\phi \right],\end{aligned} \tag{11}$$

where η is the learning rate, δ_{t_k} and δ_{t-1} denote historical gradient norms, and ϕ is the angle between consecutive gradient directions, given by: $\cos\phi = \frac{\Delta\theta_{t_k}^{\top} \Delta\theta_{t-1}}{\|\Delta\theta_{t_k}\| \cdot \|\Delta\theta_{t-1}\|}$.

This approach ensures that our utility estimation remains efficient, accurate, and fully integrated into standard model training workflows. The complete derivation of this gradient-based approximation method is presented in Theorem 6.4 (Section 6).

While the above approximation method significantly enhances efficiency, its accuracy critically depends on selecting an appropriate approximation coefficient β. To further refine our method, we analytically derive the optimal approximation weight β^* that minimizes approximation error.

Optimal Approximation Coefficient β^*. Formally, we define the approximation error function as: $J(\beta) = \|\Delta L(\theta_t, S_i) - \Delta L'(\theta_t, S_i)\|^2$. Minimizing this error function leads us to the theoretical β^*:

Theorem 5.2 (Optimal Weight β^*). *The optimal approximation weight β^* minimizing the error function $J(\beta)$ is given by:*

$$\beta^* = \frac{\delta_{t-1} - \sqrt{\delta_{t_k}\delta_{t-1}} \cos\phi}{\delta_{t_k} + \delta_{t-1} - 2\sqrt{\delta_{t_k}\delta_{t-1}} \cos\phi}. \tag{12}$$

Detailed proofs and analyses regarding the derivation of this optimal coefficient are provided in Theorem 6.4 (Section 6).

Finally, to rigorously evaluate the theoretical guarantees and practical utility of our gradient-based approximation, we establish a formal approximation error bound as follows.

Approximation Error Bound. We bound the approximation error between the approximated loss L' and the true loss L.

Theorem 5.3 (Approximation Error Bound). *With the optimal weight β^*, the error between the approximated loss L' and the true loss L satisfies:*

$$\|L'(\theta_t, x) - L(\theta_t, x)\| \leq \epsilon_{taylor} + \epsilon_{approx},$$

where:

- $L'(\theta_i, x) = L(\theta_{i-1}, x) + \Delta L'(\theta_t, S_t)$
- $\epsilon_{taylor} = \frac{1}{2}\eta^2 \cdot \max_\theta \|\nabla^2 L(\theta, x)\| \cdot \|\nabla L(S_i, \theta_{i-1})\|^2$ *is the error from Taylor expansion.*
- $\epsilon_{approx} = \eta \cdot \|\nabla L(S_i, \theta_{i-1}) - (\beta^* \cdot \nabla L(S_{i_k}, \theta_{i_k - 1}) + (1-\beta^*) \cdot \nabla L(S_{i-1}, \theta_{i-2}))\|^2$ *is the error from gradient approximation.*

5.2 Adaptive Data Selection via MAB-Integrated Training Scheduler

In this section, we propose a novel training scheduler for the LEAD framework that integrates the Multi-Armed Bandit (MAB) algorithm with our IDU utility function. The scheduler adaptively selects training data clusters based on their evolving informativeness.

Step 1: Difficulty-Aware Cluster Selection. Initially, we set the weights $W = \{w_1, w_2, ..., w_K\}$ for all clusters categorized by difficulty level, where w_i denotes the weight of cluster C_i and K is the number of clusters. To assess the difficulty score of each cluster, we employ the EXP3 [3] algorithm, a well-established method within the MAB framework, for the cluster selection. Specifically, for each iteration t, we first calculate the cluster score $DC_t(i)$ of the cluster

C_i based on the cluster weight w_i, and then select a cluster (arm) DC_t^* with the highest score DC. The $DC_t(i)$ can be computed as:

$$DC_t(i) = (1-\gamma)\frac{w_i^{(t)}}{\sum_{j=1}^{K} w_j^{(t)}} + \frac{\gamma}{K} \tag{13}$$

where γ controls the exploration-exploitation trade-off.

The selected cluster at iteration t is the one with the highest probability: $C_{i^*} = \arg\max_{i\in[1,K]} DC_t(i)$.

Step 2: Sample Selection with IDU. After selecting a cluster C_i with the highest DC score, we apply our previously introduced IDU utility function to sample the most informative subset B_{C_i} within the selected cluster C_i. Specifically, we select samples with the highest IDU scores to maximize utility gain at each iteration.

Step 3: Model Training and Reward Computation. Using the selected subset B_{C_i}, we train the large language model during iteration t. Once training is complete, we compute a reward $r_i^{(t)}$ to quantify the model's improvement resulting from the selected samples (Please refer to Section 5.3 for details).

Step 4: Cluster Weight Updates for Next Round Selection. After obtaining the reward $r_i^{(t)}$, we update the cluster weights $w_i^{(t+1)}$ according to EXP3 update rule:

$$w_i^{(t+1)} = \begin{cases} w_i^{(t)} \exp\left(\frac{\gamma}{K}\frac{r^{(t)}}{DC_t(i)}\right), & i = i_t \\ w_i^{(t)}, & \text{otherwise} \end{cases} \tag{14}$$

This adaptive weight-update mechanism ensures clusters that consistently yield high utility are progressively favored in subsequent iterations, achieving adaptive training data selection.

5.3 Self-Guided IDU-Based Reward

An effective reward function is critical to guiding effective cluster selection within the MAB framework. Ideally, such a reward should precisely capture each cluster's direct contribution to model improvement, while remaining computationally efficient and fully integrated into the training process.

To achieve this, we propose a *Self-Guided IDU-Based Reward*, leveraging our previously defined IDU utility to efficiently quantify each cluster's contribution to model improvement without additional inference overhead. Formally, the reward for training on cluster C_i at iteration t is computed as:

$$r_i^{(t)} = InfoGain(C_i, t) = \mathbb{E}_{x_i\in C_i}\left[IDU(\theta_{t-1}, x_i) - IDU(\theta_t, x_i)\right], \tag{15}$$

where θ_{t-1} and θ_t represent the model parameters before and after training, respectively. To maintain numerical stability and consistent scaling, rewards are further normalized to the range $[-1, 1]$ via min-max normalization.

Compared to traditional reward designs [8], our self-guided reward naturally integrates into the standard training loop, accurately reflects dynamic model improvements at no additional inference cost, and significantly simplifies the reward computation.

6 Theoretical Guarantees

In this section, we analyze the theoretical guarantees of our IDU utility and the LEAD framework.

6.1 Optimal Smoothing Coefficient

We now analyze the optimal smoothing coefficient for the budget-constrained IDU optimization (Problem 1, presented in Section 5.1).

Theorem 6.1 (**Optimal Smoothing Coefficient**). *The optimal smoothing coefficient b^* that maximizes the cumulative utility gain under the budget constraint is:*

$$b^* = 1 - \frac{B}{n_0 T \cdot (1 + CV^2)} \tag{16}$$

where $n_0 = \alpha \cdot \overline{|C|}$ is the expected batch size without smoothing and heterogeneity effects, B is the total budget, T is the number of training steps, and CV^2 quantifies cluster size variability.

Under a total budget B, we propose the optimization problem:

$$\max_{b,T} \sum_{t=1}^{T} \Delta IDU_t, \quad \text{s.t.} \sum_{t=1}^{T} n_t \le B \tag{17}$$

The overall goal is to maximize the cumulative utility gain, and the cumulative utility gain depends on the $\Delta IDU_t(x)$ of each round.

$$R^{(t)} = \Delta IDU_t = \sum_{x\in S_t} \left(IDU(\theta_t, x) - IDU(\theta_{t-1}, x)\right) \tag{18}$$

We take $\Delta IDU_t(x)$ of each round as the reward of the current round to guide the selection of new groups in the next round.

As the selection rounds typically exceed 5, the utility-based reward for cluster C_t simplifies to:

$$R^{(t)} = \Delta IDU_t = -(1-b)\eta_t |S_t| \Psi_t. \tag{19}$$

The specific simplification process can be referred to as Lemma 6.2. Here ΔIDU_t depends on the size of $|S_t| = n_t$. Therefore, before estimating ΔIDU_t, we need to estimate n_t. We get it in four steps.

Step 1: Estimate sample size selected in the t-th round n_t. The probability of all clusters being selected in the initial round is the same, so the clusters are randomly selected in the first round. According to the Eq. (14) and Eq. (13), which cluster is selected in the next round depends on which cluster was selected in the previous round. So we can only estimate the expectation of n_t. Then $\mathbb{E}[n_t]$ can be simplified as follows (see Lemma 6.3 for details):

$$\mathbb{E}[n_t] = \alpha \cdot (1-b) \cdot \frac{\sum_{i=1}^{K} |C_i|^2}{\sum_{i=1}^{K} |C_i|} \cdot (1 + O(\gamma)) \tag{20}$$

Step 2: Estimate the expectation of utility gain ΔIDU_t. Since the utility gain ΔIDU_t in the $T-th$ round depends on nt, and for n_t, due to the randomness of the MAB when selecting the cluster, we can only estimate the expectations. Therefore, it is necessary to further solve the expectations of ΔIDU_t. According to the Eq. (19) and Eq. (20), we can further obtain $\mathbb{E}[\Delta IDU_t]$.

$$\sum_{t=1}^{T} \mathbb{E}[\Delta IDU_t] = -\sum_{t=1}^{T} (1-b)\eta_t \cdot \mathbb{E}[|S_t|\Psi_t] \tag{21}$$

$$= -n_0 \cdot (1-b)^2 \cdot (1 + \mathrm{CV}^2) \cdot \sum_{t=1}^{T} \eta_t \delta_t \tag{22}$$

where $n_0 = \alpha \cdot \overline{|C|}$ represents the expected sample size without smoothing, $\mathbb{E}[\Psi_t \cdot |S_t|] = \delta_t \cdot \mathbb{E}[n_t]$, δ_t represents the average per-sample utility contribution.

Step 3: Redefine objective and constrained condition. Having derived the expected sample size and utility gain, we now reformulate our optimization problem by incorporating these expectations.

$$\max_{b,T} \sum_{t=1}^{T} \mathbb{E}[\Delta IDU_t], \quad \text{s.t.} \sum_{t=1}^{T} \mathbb{E}[n_t] \le B \tag{23}$$

$$\textit{where} \quad \mathbb{E}[n_t] = \alpha \cdot (1-b) \cdot |\overline{C}| \cdot (1+\mathrm{CV}^2) \cdot (1+O(\gamma)) \tag{24}$$

Let $\bar{\eta}\delta = \frac{1}{T}\sum_{t=1}^{T} \eta_t \delta_t$, The budget constraint becomes:

$$\sum_{t=1}^{T} \mathbb{E}[n_t] = \sum_{t=1}^{T} n_0 \cdot (1-b) \cdot (1+\mathrm{CV}^2) \le B \tag{25}$$

Step 4: Solving optimal b^* and T^*. We formulate the Lagrangian:

$$\mathcal{L}(b,\lambda) = \mathbb{E}[\Delta IDU_t] - \lambda(\mathbb{E}[n_t] - B) \tag{26}$$

$$= -n_0 \cdot T \cdot \bar{\eta}\delta \cdot (1-b)^2 \cdot (1+\mathrm{CV}^2) + \tag{27}$$

$$\lambda(n_0 \cdot T \cdot (1-b) \cdot (1+\mathrm{CV}^2) - B) \tag{28}$$

Taking the partial derivative with respect to b and setting it to zero:

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow 2\bar{\eta}\delta \cdot (1-b) = \lambda \tag{29}$$

The complementary slackness condition states $\lambda(n_0 \cdot T \cdot (1-b) \cdot (1+\mathrm{CV}^2) - B) = 0$. Since $\lambda \neq 0$ (as verified by the optimality condition), the budget constraint must be tight:

$$n_0 \cdot T \cdot (1-b) \cdot (1+\mathrm{CV}^2) = B \Rightarrow b^* = 1 - \frac{B}{n_0 \cdot T \cdot (1+\mathrm{CV}^2)} \tag{30}$$

We require $0 \le b^* < 1$, which implies:

$$T_{\min} = \left\lceil \frac{B}{n_0 \cdot (1+\mathrm{CV}^2)} \right\rceil + 1 \tag{31}$$

Lemma 6.2 (**Batch Utility Change Decomposition**). *The utility change for batch S_t under the smoothed utility function can be expressed as:*

$$\Delta IDU_t = \begin{cases} -(1-b)\eta_t |S_t| \Psi_t + b|S_t|\delta_{t-1}(1-b^{t-1}), & t \le 5 \\ -(1-b)\eta_t |S_t| \Psi_t, & t > 5 \end{cases} \tag{32}$$

where Ψ_t denotes the gradient alignment term:

$$\Psi_t = \beta_t^2 \delta_{t_k} + (1-\beta_t)^2 \delta_{t-1} + 2\beta_t(1-\beta_t)\sqrt{\delta_{t_k}\delta_{t-1}}\cos\phi_t \tag{33}$$

Proof. For any $x \in S_t$, $\Delta IDU_t(x)$ can be decomposed as:

$$\Delta IDU_t(x) = (1-b)\Delta L(\theta_t, x) + b(1-b)\sum_{k=0}^{t-3} b^k \Delta L(\theta_{t-2-k}, x) + (1-b)b^{t-1} IDU(\theta_0, x) \tag{34}$$

For the historical cumulative terms when $t \le 5$, we apply finite-order approximation:

$$\sum_{k=0}^{t-3} b^k \Delta L(\theta_{t-2-k}, x) \approx \delta_{t-1} \frac{1-b^{t-2}}{1-b} \tag{35}$$

The initial utility term $IDU(\theta_0, x)$ becomes a constant C_0 after aggregation. Summing over batch S_t gives:

$$\Delta IDU_t = -(1-b)\eta_t |S_t| \Psi_t + b|S_t|\delta_{t-1}(1-b^{t-1}) + (1-b)b^{t-1}|S_t|C_0 \tag{36}$$

When $t > 5$, the exponential decay term b^{t-1} becomes negligible:

$$\Delta IDU_t \approx -(1-b)\eta_t |S_t| \Psi_t \tag{37}$$

□

Lemma 6.3 (**Expected Sample Size Under MAB mechanism**). *In the MAB framework using EXP3 for cluster selection with smoothed utility, the expected sample size per round $\mathbb{E}[n_t]$ satisfies:*

$$\mathbb{E}[n_t] = \alpha \cdot (1-b) \cdot \overline{|C|} \cdot (1+CV^2) \cdot (1+O(\gamma)) \tag{38}$$

where α is the sampling rate, b is the smoothing coefficient, $|C_i|$ is the size of cluster i, and γ is the exploration rate in function 13.

Proof. We analyze cluster selection probabilities in the EXP3 algorithm when used with our smoothed utility rewards. The reward signal for selecting cluster i at time t is:

$$R_i^{(t)} = \Delta IDU_t \propto (1-b)|C_i| \tag{39}$$

This relationship follows directly from Lemma 6.2. Since $|S_t|$ is proportional to cluster size $|C_i|$ when cluster i is selected, and assuming Ψ_t and η_t are approximately constant across clusters, we derive $R_i^{(t)} \propto (1-b)|C_i|$.

From the weight update Eq. (13) and Eq. (14) in the MAB EXP3 algorithm. As the algorithm converges to steady state, the weights stabilize such that:

$$\frac{w_i^{(t)}}{\sum_{j=1}^{K} w_j^{(t)}} \propto \exp\left(\sum_{\tau=1}^{t-1} \frac{\gamma}{K} \frac{R_i^{(\tau)}}{p_i^{(\tau)}}\right) \tag{40}$$

In the fully converged regime, assuming small γ and ϵ, and sufficiently heterogeneous cluster sizes, we can derive a fixed-point equation. At this fixed point, the ratio $\frac{R_i^{(t)}}{p_i^{(t)}}$ becomes approximately constant across arms, leading to:

$$p_i^{(t)} \approx \frac{(1-\gamma)(1-b)|C_i|}{\sum_{j=1}^{K}(1-b)|C_j|} + \frac{\gamma}{K} \approx \frac{(1-b)|C_i|}{\sum_{j=1}^{K}|C_j|} + O(\gamma) \tag{41}$$

The expected sample size in round t is:

$$\mathbb{E}[n_t] = \alpha \sum_{i=1}^{K} p_i^{(t)} |C_i| = \alpha(1-b)\frac{\sum_{i=1}^{K}|C_i|^2}{\sum_{j=1}^{K}|C_j|} + \alpha \cdot O(\gamma) \sum_{i=1}^{K} |C_i| \tag{42}$$

Since $\sum_{i=1}^{K} |C_i| = N$ (total dataset size), we can express this as:

$$\mathbb{E}[n_t] = \alpha \cdot (1-b) \cdot \frac{\sum_{i=1}^{K}|C_i|^2}{\sum_{i=1}^{K}|C_i|} \cdot (1+O(\gamma)) \tag{43}$$

Let $\overline{|C|} = \frac{1}{K}\sum_{i=1}^{K}|C_i|$ be the average cluster size. Using the relation between variance and second moment:

Substituting into our expected sample size formula:

$$\mathbb{E}[n_t] = \alpha \cdot (1-b) \cdot \overline{|C|} \cdot (1+\mathrm{CV}^2) \cdot (1+O(\gamma)) \tag{44}$$

□

6.2 Loss Changes in Gradient-Based Approximation

Recap that we have introduced utility function Eq. (8) in Section 3, In this section, we try to approximate the loss reduction $\Delta L'(\theta_t, x)$.

THEOREM 6.4 (**IU CHANGE APPROXIMATION**). *For any sample set S_t, the average uncertainty change $\Delta L'(\theta_t, S_t)$ when model parameters update from θ_{t-1} to θ_t can be approximated as:*

$$\delta_t \equiv \Delta L'(\theta_t, S_t) \tag{45}$$

$$= -\eta\left[\beta^2\delta_{t_k} + (1-\beta)^2\delta_{t-1} + 2\beta(1-\beta)\sqrt{\delta_{t_k}\delta_{t-1}}\cos\phi\right] \tag{46}$$

where ϕ is the angle between parameter update directions $\Delta\theta_{t_k}$ and $\Delta\theta_{t-1}$, with $\cos\phi = \frac{\Delta\theta_{t_k}^\top \Delta\theta_{t-1}}{\|\Delta\theta_{t_k}\|\|\Delta\theta_{t-1}\|}$.

$$\beta^* = \frac{\delta_{t-1} - \sqrt{\delta_{t_k}\delta_{t-1}}\cos\phi}{\delta_{t_k} + \delta_{t-1} - 2\sqrt{\delta_{t_k}\delta_{t-1}}\cos\phi} \tag{47}$$

Step 1: Simplify the loss change. Assume at iteration t, model parameters are updated via gradient descent: $\theta_t = \theta_{t-1} - \eta_t \nabla L(S_t, \theta_{t-1})$, where $\nabla L(S_t, \theta_{t-1}) = \frac{1}{|S_t|}\sum_{x\in S_t}\nabla L(x, \theta_{t-1})$ is the average gradient of subset S_t. For each sample $x \in S_t$, the loss function $L(\theta, x)$ is expanded using first-order Taylor expansion at θ_{t-1}:

$$L(\theta_t, x) \approx L(\theta_{t-1}, x) + \nabla L(\theta_{t-1}, x)^\top(\theta_t - \theta_{t-1}) \tag{48}$$

Averaging over all samples in S_t:

$$\delta_t = \Delta L'(\theta_t, S_t) \approx -\eta_t \frac{1}{|S_t|}\sum_{x\in S_t}\nabla L(\theta_{t-1}, x)^\top \nabla L(\theta_{t-1}, S_t) \tag{49}$$

$$= -\eta_t \|\nabla L(\theta_{t-1}, S_t)\|^2 \tag{50}$$

It can be concluded that the loss reduction is related to the gradient.

Step 2: Approximate the gradient. To further approximate the loss, we need to approximate the gradient. Here we consider that the gradient at the current moment is related to the gradient at the previous moment and the gradient when the cluster used at the current moment was first selected.

$$\nabla L'(S_t, \theta_{t-1}) \equiv \beta \cdot \nabla L(S_{t_k}, \theta_{t_k-1}) + (1-\beta)\cdot \nabla L(S_{t-1}, \theta_{t-2}), \tag{51}$$

where t_k is the most recent step when C_k was previously selected, C_k is the cluster selected at step t, where $\beta \in [0, 1]$ is a weighting coefficient measuring the relative importance of cluster-specific historical information versus recent optimization direction.

Step 3: Solving optimal β^* to obtain final IU Change Approximation $\Delta L'(\theta_t, S_t)$. The β^* can be solved by minimizing the difference between the current gradient and the approximate gradient.

$$J(\beta) = \|\nabla L_t - (\beta\nabla L_{t_k} + (1-\beta)\nabla L_{t-1})\|^2 \tag{52}$$

Using the gradient descent update rule $\Delta\theta_t = -\eta\,\nabla L_t$, we rewrite in terms of parameter updates:

$$J(\beta) = \frac{1}{\eta^2}\left\|\Delta\theta_t - \left(\beta\,\Delta\theta_{t_k} + (1-\beta)\,\Delta\theta_{t-1}\right)\right\|^2. \tag{53}$$

Since $\|\Delta\theta_{t_k}\|^2 \approx -\eta\,\delta_{t_k}$ and $\cos\phi = \frac{\Delta\theta_{t_k}^\top\Delta\theta_{t-1}}{\|\Delta\theta_{t_k}\|\|\Delta\theta_{t-1}\|}$.

Setting $\frac{d\tilde{J}}{d\beta} = 0$ yields the optimal coefficient:

$$\beta^* = \frac{\delta_{t-1} - \sqrt{\delta_{t_k}\delta_{t-1}}\cos\phi}{\delta_{t_k} + \delta_{t-1} - 2\sqrt{\delta_{t_k}\delta_{t-1}}\cos\phi}. \tag{54}$$

The loss change is then approximated as:

$$\delta_t = -\eta\left[(\beta^*)^2\,\delta_{t_k} + (1-\beta^*)^2\,\delta_{t-1} + 2\,\beta^*(1-\beta^*)\sqrt{\delta_{t_k}\delta_{t-1}}\,\cos\phi\right]. \tag{55}$$

7 Experiments

8 Related Work

Data Selection for Instruction Tuning. Previous works on data selection [9, 23, 59, 65] can be broadly categorized into two key approaches: model-agnostic methods and model-aware methods.

Model-agnostic methods operate independently of the target model, including rule-based approaches [5, 6, 28, 40, 45, 49, 66] that are computationally efficient but lack semantic understanding. Advanced model-based methods [13, 14, 35] like GPT-4 [1] that provide nuanced assessment at high computational cost, and proxy model-based methods [31, 61] that balance efficiency and quality. However, these methods cannot adapt to the specific learning characteristics of the target model. Model-aware methods [5, 7, 8, 36, 41, 42, 64] address this limitation by customizing selection based on the model's learning dynamics, though they introduce higher computational costs through required model inference or fine-tuning. In contrast, LEAD proposes a two-stage adaptive approach that efficiently combines model-aware adaptiveness with zero computational overhead, effectively addressing the challenge of balancing effectiveness and efficiency in instruction tuning data selection.

Sample Utility Scores. Sample utility scoring plays a critical role in data selection, employing various predefined metrics [7, 47, 57]. Perplexity-based metrics [31, 44] favor simpler patterns, while diversity-aware selection [58, 63] ensures broad coverage but depends heavily on pre-trained embedding quality. Quality-based metrics incorporating influence scoring [16, 21, 29, 59] and external model [33] evaluation are theoretically sound but require expensive gradient computations. Complexity-based selection [32, 38] risks including noisy samples that hinder convergence, while uncertainty-driven metrics [22, 37] suffer from instability due to loss landscape irregularities. A common limitation across these approaches is their significant computational overhead. Although recent efforts have improved data efficiency in utility estimation, they still incur additional costs. We propose IDU, a novel utility function achieving zero-cost estimation while maintaining selection effectiveness.

9 Conclusion

In this paper, we proposed LEAD, an efficient iterative data selection framework for instruction tuning of LLMs. LEAD introduces Instance-Level Dynamic Uncertainty utility function, enabling accurate utility estimation without extra inference. In addition, we developed a coarse-to-fine selection approach guided by a multi-armed bandit mechanism. Experiments show LEAD achieves 6.1%-10.8% performance improvement using only 2.5% training data and reduces training costs by 5-10×.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. [n. d.]. A Survey on Data Selection for Language Models. *Transactions on Machine Learning Research* ([n. d.]).

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32, 1 (2002), 48–77.

[4] Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. 2024. Data Diversity Matters for Robust Instruction Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 3411–3425.

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

[6] Chengliang Chai, Lei Cao, Guoliang Li, Jian Li, Yuyu Luo, and Samuel Madden. 2020. Human-in-the-loop Outlier Detection. In *SIGMOD Conference*. ACM, 19–33.

[7] Chengliang Chai, Jiabin Liu, Nan Tang, Ju Fan, Dongjing Miao, Jiayi Wang, Yuyu Luo, and Guoliang Li. 2023. Goodcore: Data-effective and data-efficient machine learning through coreset selection over incomplete data. *Proceedings of the ACM on Management of Data* 1, 2 (2023), 1–27.

[8] Chengliang Chai, Jiabin Liu, Nan Tang, Guoliang Li, and Yuyu Luo. 2022. Selective data acquisition in the wild for model charging. *Proceedings of the VLDB Endowment* 15, 7 (2022), 1466–1478.

[9] Chengliang Chai, Nan Tang, Ju Fan, and Yuyu Luo. 2023. Demystifying Artificial Intelligence for Data Preparation. In *SIGMOD Conference Companion*. ACM, 13–20.

[10] Chengliang Chai, Jiayi Wang, Yuyu Luo, Zeping Niu, and Guoliang Li. 2023. Data Management for Machine Learning: A Survey. *IEEE Trans. Knowl. Data Eng.* 35, 5 (2023), 4646–4667.

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

[12] Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation.

[13] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. [n. d.]. AlpaGasus: Training a Better Alpaca with Fewer Data. In *The Twelfth International Conference on Learning Representations*.

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

[15] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[16] Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, et al. 2024. What is your data worth to gpt? llm-scale data valuation with influence functions. *arXiv preprint arXiv:2405.13954* (2024).

[17] Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics* 8 (2020), 454–470.

[18] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[19] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 3029–3051.

[20] Everette S Gardner Jr. 1985. Exponential smoothing: The state of the art. *Journal of forecasting* 4, 1 (1985), 1–28.

[21] Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*. PMLR, 2242–2251.

[22] Jindong Han, Hao Liu, Jun Fang, Naiqiang Tan, and Hui Xiong. [n. d.]. Automatic Instruction Data Selection for Large Language Models via Uncertainty-Aware Influence Maximization. In *THE WEB CONFERENCE 2025*.

[23] LIU Hanmo, DI Shimin, LI Haoyang, LI Shuangyin, CHEN Lei, and ZHOU Xiaofang. 2024. Effective Data Selection and Replay for Unsupervised Continual Learning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 1449–1463.

[24] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. [n. d.]. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

[25] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* (2021).

[26] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689* (2022).

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[28] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems* 36 (2023), 47669–47681.

[29] Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. [n. d.]. DataInf: Efficiently Estimating Data Influence in LoRA-tuned LLMs and Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

[30] Boyan Li, Yuyu Luo, Chengliang Chai, Guoliang Li, and Nan Tang. 2024. The Dawn of Natural Language to SQL: Are We Fully Ready? [Experiment, Analysis & Benchmark]. *Proc. VLDB Endow.* 17, 11 (2024), 3318–3331.

[31] Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024. Superfiltering: Weak-to-Strong Data Filtering for Fast Instruction-Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14255–14273.

[32] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 7595–7628.

[33] Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, et al. 2024. 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)*. 4586–4601.

[34] Yiwei Li, Jiayi Shi, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Boyuan Pan, Heda Wang, and Yao Hu. 2024. Instruction Embedding: Latent Representations of Instructions Towards Task Identification. *Advances in Neural Information Processing Systems* 37 (2024), 87683–87711.

[35] W Lian et al. 2023. SlimOrca: An Open Dataset of GPT-4 Augmented FLAN Reasoning Traces, with Verification.

[36] Jiabin Liu, Chengliang Chai, Yuyu Luo, Yin Lou, Jianhua Feng, and Nan Tang. 2022. Feature augmentation with reinforcement learning. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 3360–3372.

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

[38] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=BTKAeLqLMw

[39] Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. 2025. A Survey of NL2SQL with Large Language Models: Where are we, and where are we going? arXiv:2408.05109 [cs.DB] https://arxiv.org/abs/2408.05109

[40] Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. [n. d.]. # InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

[41] Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. 2020. Interactive Cleaning for Progressive Visualization through Composite Questions. In *ICDE*. IEEE, 733–744.

[42] Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. 2020. VisClean: Interactive Cleaning for Progressive Visualization. *Proc. VLDB Endow.* 13, 12 (2020), 2821–2824.

[43] Yuyu Luo, Yihui Zhou, Nan Tang, Guoliang Li, Chengliang Chai, and Leixian Shen. 2023. Learned Data-aware Image Representations of Line Charts for Similarity Search. *Proc. ACM Manag. Data* 1, 1 (2023), 88:1–88:29.

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

[45] Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro Von Werra, and Shayne Longpre. 2023. Octopack: Instruction tuning code large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

[46] Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2020. Making data visualization more efficient and effective: a survey. *VLDB J.* 29, 1 (2020), 93–117.

[47] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB endowment. International conference on very large data bases*, Vol. 11. 269.

[48] Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

[49] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15725–15788.

[50] Jielin Song, Siyu Liu, Bin Zhu, and Yanghui Rao. 2024. IterSelectTune: An Iterative Training Framework for Efficient Instruction-Tuning Data Selection. *arXiv preprint arXiv:2410.13464* (2024).

[51] Wangtao Sun, Haotian Xu, Xuanqing Yu, Pei Chen, Shizhu He, Jun Zhao, and Kang Liu. 2024. ItD: Large Language Models Can Teach Themselves Induction through Deduction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2719–2731.

[52] Nan Tang, Chenyu Yang, Ju Fan, Lei Cao, Yuyu Luo, and Alon Y. Halevy. 2024. VerifAI: Verified Generative AI. In *CIDR*. www.cidrdb.org.

[53] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

[54] Joannes Vermorel and Mehryar Mohri. 2005. Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*. Springer, 437–448.

[55] Jiachen Tianhao Wang, Tong Wu, Dawn Song, Prateek Mittal, and Ruoxi Jia. 2024. GREATS: Online selection of high-quality data for llm training in every iteration. *Advances in Neural Information Processing Systems* 37 (2024), 131197–131223.

[56] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems* 36 (2023), 74764–74786.

[57] Yong Wang, Kaiyu Li, Yuyu Luo, Guoliang Li, Yunyan Guo, and Zhuo Wang. 2024. Fast, Robust and Interpretable Participant Contribution Estimation for Federated Learning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2298–2311.

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

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

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

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

[62] Mingjia Yin, Chuhan Wu, Yufei Wang, Hao Wang, Wei Guo, Yasheng Wang, Yong Liu, Ruiming Tang, Defu Lian, and Enhong Chen. 2024. Entropy law: The story behind data compression and llm performance. *arXiv preprint arXiv:2407.06645* (2024).

[63] Simon Yu, Liangyu Chen, Sara Ahmadian, and Marzieh Fadaee. 2024. Diversify and Conquer: Diversity-Centric Data Selection with Iterative Refinement. *arXiv preprint arXiv:2409.11378* (2024).

[64] Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Jiantao Qiu, Lei Cao, Ju Fan, et al. 2024. Harnessing Diversity for Important Data Selection in Pretraining Large Language Models. *arXiv preprint arXiv:2409.16986* (2024).

[65] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems* 36 (2023), 55006–55021.

[66] Terry Yue Zhuo, Armel Zebaze, Nitchakarn Suppattarachai, Leandro von Werra, Harm de Vries, Qian Liu, and Niklas Muennighoff. 2024. Astraios: Parameter-efficient instruction tuning code large language models. *arXiv preprint arXiv:2401.00788* (2024).