Data-Efficient Selection via Grammatical Complexity in Continual Pre-training of Domain-Specific LLMs

Yizhou Ying[1], **Geng Zhang**[1], **Danxin Cui**[3], **Chengyu Du**[1], **Guanglei Yue**[2]
Sihang Jiang[1*], **Jiaqing Liang**[2], **Yifei Fu**[4], **Hailin Hu**[4], **Yanghua Xiao**[1*]
[1]College of Computer Science and Artificial Intelligence, [2]School of Data Science, Fudan University
[3]College of Foreign Languages and Literature, Fudan University; [4]Huawei Noah's Ark Lab
{yzying24, gzhang24, dxcui22, cydu24, glyue24}@m.fudan.edu.cn;
{shjiang20, liangjiaqin, shawyh}@fudan.edu.cn;{fuyifei1, hailin.hu}@huawei.com

Abstract

Data efficiency is crucial in domain-specific continual pre-training (CPT) of large language models (LLMs), especially under resource constraints. Aiming for "small data, big impact," this work addresses the limitations of existing domain-specific data selection strategies, which often rely on scarce labeled data or computationally expensive LLMs. We introduce CDF Sampling with Grammatical Complexity (CDF-GC)[1], an annotation-independent, efficient and interpretable data selection framework for CPT. Our approach comprehensively evaluates grammatical complexity using lexical diversity and syntactic complexity, and employs a cumulative distribution function (CDF)-based sampling strategy to balance complexity and diversity. To validate the effectiveness of CDF-GC, we conducted experiments on a financial dataset. The results demonstrate that CDF-GC significantly outperforms baselines, achieving 2.0% improvement in financial benchmark at the same selection ratio and even surpassing full-data training by 1.7% using only 20% of the data.

1 Introduction

With the gradual maturation of the Transformer architecture (Vaswani et al., 2017; Touvron et al., 2023a,b) and continuous improvements in training methodologies such as pre-training (PT) (Brown et al., 2020; Chowdhery et al., 2022), continual pre-training (CPT) (Zhou et al., 2024; Wu et al., 2023), instruction fine-tuning (IFT) (Ouyang et al., 2022; Wei et al., 2022), and reinforcement learning (RL) (Shao et al., 2024; Bai et al., 2022; Schulman et al., 2017), data-centric AI has increasingly become a research focus in both academia and industry (Zha et al., 2025; Jakubik et al., 2024). While PT equips LLMs with general language capabilities, domain-specific CPT is a crucial step in building domain-expert models. This research focuses on the problem of efficiently leveraging domain-specific data in CPT and investigates data screening techniques to achieve this goal.

While the rapid growth of data resources (Gao et al., 2021; Weber et al., 2024; Xue et al., 2022) offers opportunities for training LLMs, it concurrently introduces critical challenges, notably the substantial computational cost associated with large-scale training (Covert et al., 2024; Hoffmann et al., 2022) and the issue of low quality in web-sourced data. Research demonstrates the negative impact of low-quality data on LLMs (Iskander et al., 2024), highlighting that a small volume of high-quality data can outperform a large amount of raw data (Yin and Rush, 2024; Xie et al., 2024). Therefore, efficiently selecting the most valuable samples from massive datasets to maximize model performance has emerged as a significant challenge for improving LLM training efficiency.

Recent domain-specific data selection approaches mainly measure the relevance between the original data and the target domain, employing metrics such as embedding similarity (Xie et al., 2024; Gururangan et al., 2020), loss difference (Moore and Lewis, 2010), and N-gram feature similarity (Xie et al., 2023). These methods are suitable for identifying the most relevant data to the target domain from large-scale unlabeled datasets. However, after we have a sufficient source of domain-specific data, further filtering the most valuable samples for training domain-specific expert models **relies on a large amount of labeled high quality target domain data**, limiting their generalizability and applicability in scenarios where annotated data is insufficient.

Concurrently, domain-agnostic data selection methods also present their own limitations. Rule-based data quality filtering offers an efficient initial data cleaning approach for immense PT datasets (Penedo et al., 2024a,b), but it has **limitations in processing complex semantic information**. Conversely, LLM-based strategies, including perplexity evaluation (Marion et al., 2023; Yin and Rush, 2024), high-quality data synthesis (Luo et al., 2025), data distillation (Hsieh et al., 2023) and prompting for data assessment (Liu et al., 2024), effectively filter and synthesize information-dense data, significantly enhancing the IFT and RL stages of LLM training. However, **the substantial inference cost associated with LLMs** makes it not suitable for large-scale CPT data screening in resource-constrained settings. Besides, data scoring methods based on reward models (Lozhkov et al., 2024) quantify data quality with relatively acceptable computational costs. But these methods have an inherent **lack of objectivity and**

[*] The corresponding author.

[1] The source code is publicly available in the https://github.com/PPMark0712/CDF-GC repository.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22055–22069
November 4-9, 2025 ©2025 Association for Computational Linguistics

interpretability, as their scoring outcomes are susceptible to reward model biases, making it difficult to provide clear and reliable decision-making rationale.

To address these challenges, we propose to leverage inherent linguistic properties for data quality assessment. Research has demonstrated that lexical diversity and syntactic complexity serve as effective indicators for assessing the quality of data (Havrilla et al., 2024; Tsvetkov et al., 2016), and these two aspects constitute key dimensions of **grammatical complexity** (**GC**) (O'Leary and Steinkrauss, 2022; Donnelly et al., 2025). Considering that lexical diversity reflects the scope and precision of information, and syntactic complexity embodies logical relations and abstract thinking, we hypothesize that data with high GC is likely to contain richer domain-specific knowledge. So we utilize GC metrics (combination of lexical diversity and syntactic complexity) for data screening. Compared to traditional methods that rely on basic metrics such as token types, sentence length, and verb ratio, this framework innovatively introduces content word entropy and dependency tag entropy to capture deeper GC features, while also incorporating empirically validated metrics like part-of-speech entropy (Xie et al., 2024), dependency distance, and dependency tree height for a comprehensive assessment from multiple angles. To ensure a unified scale and allow for effective aggregation across these diverse metrics, these individual metrics are first normalized per feature, and their average serves as the overall GC score.

Furthermore, to mitigate potential training convergence difficulties and reduced generalization arising from excessively high data complexity, we introduce a novel Cumulative Distribution Function (CDF)-based balanced sampling method, which adaptively assigns higher sampling probabilities to data with greater GC, while still ensuring that data with lower GC has a non-negligible chance of being selected. This strategy aims to elevate the overall GC distribution of the selected data while simultaneously preserving the balance and representativeness of the original data distribution.

The main contributions of this work are threefold:

- We construct an efficient (requiring only 100M-scale syntactic parsing models), objective and comprehensive grammatical complexity evaluation framework that incorporates multiple linguistic dimensions, innovatively introducing content word entropy and dependency tag entropy, providing a reliable foundation for data assessment.

- We propose a novel CDF-based sampling method that balances grammatical complexity and data diversity, enabling effective selection of linguistically challenging samples while maintaining broad coverage of the original data distribution.

- We selected a 2B tokens subset from a 15B tokens (10B after cleaning) financial dataset. Continual pre-training a 1B LLM on this subset outperformed models trained on baseline-selected subsets or the full 10B dataset in both domain and general question answering.

2 Methodology

2.1 Problem Setting

Given a dataset $D = \{x_1, x_2, \ldots, x_n\}$ containing n data points, and a target subset size $T < n$. Our goal is to find an optimal subset $S^* \subseteq D$ such that $|S^*| = T$, and the model trained on this subset achieves the best performance on the validation set $\mathcal{V}$. This problem can be formally expressed as:

$$S^* = \underset{S \subseteq D, |S|=T}{\arg\max} \; \mathcal{P}(f_S, \mathcal{V}),$$

where f_S denotes the model trained on subset S, $\mathcal{P}(f, \mathcal{V})$ is the performance metric on the validation set $\mathcal{V}$, and T is the predetermined subset size ($T < n$) constrained by computational resources.

2.2 CDF-GC Pipeline

Our CDF Sampling with Grammatical Complexity (CDF-GC) method achieves efficient data screening through two core steps (Figure 1): first, it conducts grammatical complexity (**GC**) evaluation by quantitatively analyzing lexical diversity and syntactic complexity to generate a GC score that serves as the screening basis; subsequently, it employs balanced sampling via cumulative distribution function (CDF) to shift the GC distribution towards higher-score regions while maintaining overall data balance, where the cumulative distribution-based strategy optimizes data distribution alignment through probabilistic reweighting of low/high-score samples. This integrated framework effectively combines complexity analysis with distribution-aware sampling to enhance data quality and model generalization.

2.3 Grammatical Complexity Evaluation

Building upon the established significance of lexical diversity and syntactic complexity in evaluating data quality (Havrilla et al., 2024; Tsvetkov et al., 2016), we systematically investigate the influence of **grammatical complexity** (**GC**) on domain-specific CPT of LLMs, aiming to automatically screen data containing deep domain knowledge through high-GC features. To this end, we construct a multi-dimensional evaluation framework that integrates computational linguistics features such as lexical diversity and syntactic complexity, and establish a GC quantification system based on normalized average. Specifically, we use *content word entropy* (H_{con}) and *part-of-speech entropy* (H_{pos}) to quantify lexical diversity, and *dependency tag entropy* (H_{dep}), *average dependency distance* ($\bar{d}_{\text{dep}}$), and *average dependency tree height* ($\bar{h}_{\text{dep}}$) to quantify syntactic complexity. After normalizing these five metrics, we take their mean as the comprehensive **GC** score.

- **Content Word Entropy**:

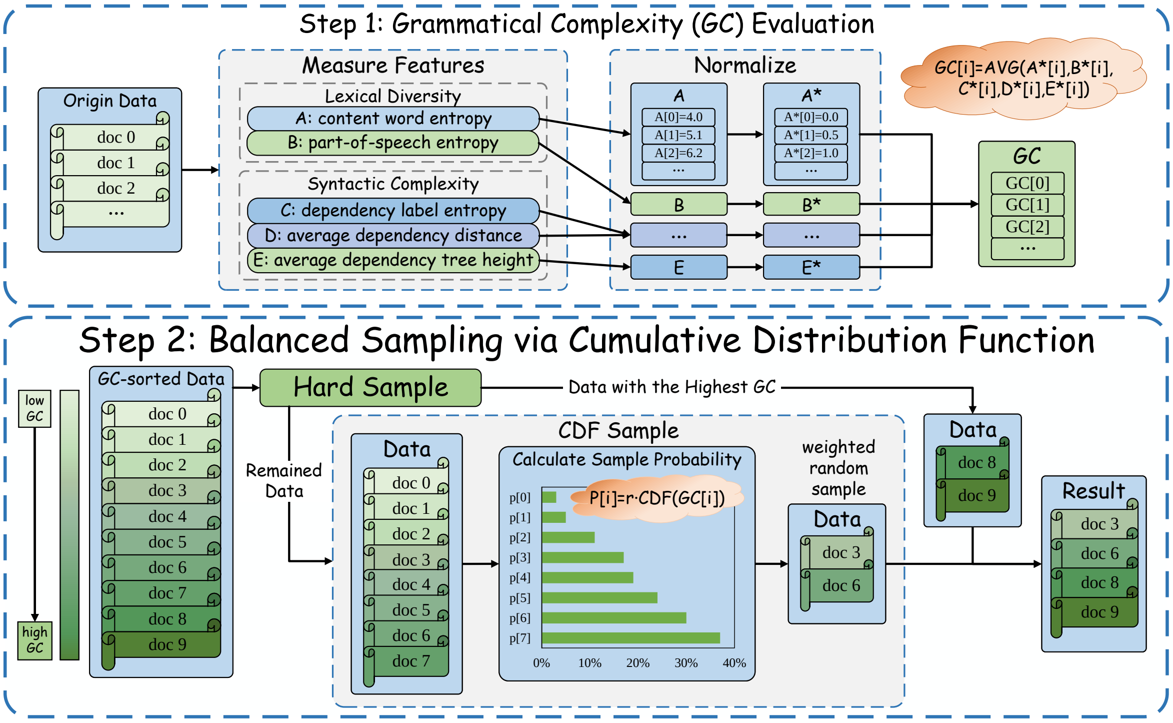

Figure 1: Pipeline of CDF Sampling with Grammatical Complexity (CDF-GC)

- **Formula**: $H_{\text{con}} = -\sum_{w \in \mathcal{V}_{\text{con}}} p(w) \log p(w)$.
- **Explanation**: $\mathcal{V}_{\text{con}}$ is the content word set, $p(w) = \frac{n_w}{N_{\text{con}}}$ is the probability of word w, with n_w being the frequency of w and N_{con} the total number of content words, quantifying lexical diversity.

- **Part-of-speech Entropy**:
  - **Formula**: $H_{\text{pos}} = -\sum_{t \in \mathcal{T}} p(t) \log p(t)$.
  - **Explanation**: $\mathcal{T}$ is the part-of-speech (POS) tag set, $p(t) = \frac{n_t}{N_{\text{pos}}}$ is the probability of POS tag t, with n_t being the frequency of t and N_{pos} the total number of POS tags, measuring POS tag diversity.

- **Dependency Label Entropy**:
  - **Formula**: $H_{\text{dep}} = -\sum_{r \in \mathcal{R}} p(r) \log p(r)$.
  - **Explanation**: $\mathcal{R}$ is the dependency relation set, $p(r) = \frac{n_r}{N_{\text{dep}}}$ is the probability of dependency relation r, with n_r being the frequency of r and N_{dep} the total number of dependency labels, reflecting syntactic relation diversity.

- **Average Dependency Distance**:
  - **Formula**: $\bar{d}_{\text{dep}} = \frac{1}{|\mathcal{E}|} \sum_{(w_i, w_j) \in \mathcal{E}} |i - j|$.
  - **Explanation**: $\mathcal{E}$ is the dependency edge set, $|i - j|$ is the absolute positional distance (in terms of word indices in the text) between dependent word w_i and its head w_j, measures syntactic dependency span complexity.

- **Average Dependency Tree Height**:
  - **Formula**: $\bar{h}_{\text{dep}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \text{depth}(T_s)$.
  - **Explanation**: $\mathcal{S}$ is the sentence set, $\text{depth}(T_s)$ is the height of the dependency tree for sentence s, Measures hierarchical structural complexity.

For a dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ with n samples, the feature vector of sample x_i is defined as

$$\mathbf{m}_i = \left[H_{\text{con}}^{(i)}, H_{\text{pos}}^{(i)}, H_{\text{dep}}^{(i)}, \bar{d}_{\text{dep}}^{(i)}, \bar{h}_{\text{dep}}^{(i)}\right]^T,$$

where the superscript (i) indicates the value for the i-th sample.

Since the five indicators have different units, each component of $\mathbf{m}_i$ is normalized using min-max normalization:

$$\mathbf{m}_i^* = \left[\frac{m_{ij} - \min_k\{m_{kj}\}}{\max_k\{m_{kj}\} - \min_k\{m_{kj}\}}\right]_{j=1}^{5},$$

where $\min_k\{m_{kj}\}$ and $\max_k\{m_{kj}\}$ denote, respectively, the minimum and maximum value of the j-th feature across all samples.

Finally, the **GC** score is defined as the mean of the normalized features:

$$\text{GC}(x_i) = \frac{1}{5} \sum_{j=1}^{5} m_{ij}^*.$$

2.4 Balanced Sampling via Cumulative Distribution Function

To obtain data with high grammatical complexity while maintaining diversity across the overall distribution, we utilize a staged sampling strategy that combines **hard sampling** with *cumulative distribution function-based sampling* (**CDF sampling**). The key approach involves first selecting high-complexity samples through hard sampling and then preserving data diversity by applying weighted random sampling. The pseudocode is provided in Algorithm 1 to facilitate a comprehensive understanding of its operational details.

Algorithm 1 Balanced Sampling via Cumulative Distribution Function

```
Require: Dataset $\mathcal{D} = \{x_1, \dots, x_n\}$, total token budget $T$, ratio of budget for hard sampling $p \in [0, 1]$
Ensure: Sampled subset $\mathcal{D}^* \subseteq \mathcal{D}$ within budget
$T_{\text{hard}} \leftarrow pT$, $T_{\text{cdf}} \leftarrow T - T_{\text{hard}}$    ▷ Budget for Hard Sample / CDF Sample
$\mathcal{D}_{\text{top}} \leftarrow$ HARD_SAMPLE($\mathcal{D}, T_{\text{hard}}$)
$\mathcal{D}' \leftarrow \mathcal{D} \setminus \mathcal{D}_{\text{top}}$
$\mathcal{D}'_s \leftarrow$ CDF_SAMPLE($\mathcal{D}', T_{\text{cdf}}$)
$\mathcal{D}^* \leftarrow \mathcal{D}_{\text{top}} \cup \mathcal{D}'_s$
return $\mathcal{D}^*$
Define GC($x$): Returns grammatical complexity of Sample $x$.
Define TC($x$): Returns token count of Sample $x$.
function HARD_SAMPLE($\mathcal{D}, T_{\text{hard}}$)
    Sort $\mathcal{D}$ by GC($x$) descending
    $\mathcal{D}_{\text{top}} \leftarrow \varnothing$, $t \leftarrow 0$
    for $x \in \mathcal{D}$ do
        if $t + \text{TC}(x) > T_{\text{hard}}$ then break
        end if
        $\mathcal{D}_{\text{top}} \leftarrow \mathcal{D}_{\text{top}} \cup \{x\}$
        $t \leftarrow t + \text{TC}(x)$
    end for
    return $\mathcal{D}_{\text{top}}$
end function
function CDF($\mathcal{D}', z$)
    $\mathcal{D}'_z \leftarrow \{x \mid x \in \mathcal{D}' \land \text{GC}(x) \le z\}$
    return $\sum_{x \in \mathcal{D}'_z} \text{TC}(x) / \sum_{x \in \mathcal{D}'} \text{TC}(x)$
end function
function CDF_SAMPLE($\mathcal{D}', T_{\text{cdf}}$)
    $E_t \leftarrow \sum_{x \in \mathcal{D}'} [\text{CDF}(\mathcal{D}', \text{GC}(x)) \cdot \text{TC}(x)]$
    $r \leftarrow T_{\text{cdf}} / E_t$    ▷ For the parameter calculation rule, see Appendix A
    $\mathcal{D}'_s \leftarrow \varnothing$
    for $x \in \mathcal{D}'$ do
        $p_x \leftarrow r \cdot \text{CDF}(\mathcal{D}', \text{GC}(x))$
        if Uniform$(0, 1) \le p_x$ then    ▷ Generate a random number for each sample
            $\mathcal{D}'_s \leftarrow \mathcal{D}'_s \cup \{x\}$
        end if
    end for
    return $\mathcal{D}'_s$
end function
```

Given a dataset $\mathcal{D}$ and a total token budget T, a portion of the budget, $T_{\text{hard}} = pT$ ($p \in [0, 1]$), is allocated to the hard sampling phase, while the remaining budget, $T_{\text{cdf}} = (1 - p)T$, is reserved for the CDF sampling phase. In the hard sampling phase, the samples with the highest grammatical complexity are selected from $\mathcal{D}$, forming the set $\mathcal{D}_{\text{top}}$, until the budget T_{hard} is exhausted. For clarity, we denote the token count of sample x as TC(x).

$$\mathcal{D}_{\text{top}} = \left\{ x_i \;\middle|\; x_i \in \mathcal{D}_{\text{sorted}} \;\land \sum_{j=1}^{i} \text{TC}(x_j) \le T_{\text{hard}} \right\},$$

where $\mathcal{D}_{\text{sorted}}$ represents the dataset $\mathcal{D}$ sorted in descending order based on the grammatical complexity score GC(x_i).

The remaining data, $\mathcal{D}' = \mathcal{D} \setminus \mathcal{D}_{\text{top}}$, enters the CDF sampling phase. Before performing CDF sampling, we first define the cumulative distribution function (CDF) of the grammatical complexity score GC(x) based on token count (rather than the number of documents), enabling precise quantification of data volume with GC values below any given threshold:

$$\text{CDF}(z) = \frac{\sum_{x_i \in \mathcal{D}'_z} \text{TC}(x_i)}{\sum_{x_i \in \mathcal{D}'} \text{TC}(x_i)},$$

$$\mathcal{D}'_z = \{x_i \in \mathcal{D}' \mid \text{GC}(x_i) \le z\}.$$

In the CDF sampling phase, the sampling probability $P(x_i)$ for each data sample x_i is adaptively set based on the CDF of the GC:

$$P(x_i) = \min\left(r \cdot \text{CDF}(\text{GC}(x_i)), 1\right).$$

The parameter r is used solely to adaptively control the sampling budget, with the detailed method described in Appendix A. Based on the sampling probabilities, samples are drawn from $\mathcal{D}'$ to form $\mathcal{D}'_s$. With this weighted random sampling approach, data with higher grammatical complexity are more likely to be selected. This method improves the overall grammatical complexity distribution of the sampled data while maintaining the diversity of the distribution, ensuring that even low-grammatical-complexity samples still have a certain probability of being selected.

$$\mathcal{D}'_s = \{x_i \mid x_i \in \mathcal{D}' \land u_i \le P(x_i), u_i \sim U(0, 1)\}.$$

Finally, the results from the hard sampling phase and the CDF sampling phase are integrated to obtain the complete sampling result, $\mathcal{D}^*$:

$$\mathcal{D}^* = \mathcal{D}_{\text{top}} \cup \mathcal{D}'_s.$$

3 Experiments

3.1 Experimental Setup

We primarily use the *Llama-3.2-1B* (Dubey et al., 2024) and *Qwen2.5-0.5B* (Yang et al., 2024a) as base model. The training approach mainly follows a continual pretraining (CPT) strategy, with specific training configurations detailed in Appendix B.1.

Training Datasets The training data in this study is derived from the domain-specific corpus *FinCorpus* (Duxiaoman-DI, 2023), which focuses on Chinese financial domain data. A more detailed description of the training data can be found in Appendix B.2.

Model	Domain	General Benchmarks				Avg.
	CFinBench	CMMLU	MMLU	Xiezhi	ICLEval	
Base Model	22.4	26.2	27.9	30.1	43.6	30.0
Full-Scale	28.0	26.4	27.5	28.3	38.5	29.7
Random	26.3	26.0	27.0	29.6	**38.9**	29.2
DSIR	27.7	26.2	27.9	28.2	36.3	29.3
ETA-DACP	27.1	26.4	28.7	30.4	35.0	29.5
PPL	27.7	26.8	29.0	27.9	36.5	29.6
CDF-GC (Ours)	**29.7**	**27.7**	**29.3**	**30.9**	35.1	**30.5**

Table 1: Performance comparison of data selection methods for continual pre-training in the financial domain. This table presents a comparative evaluation of CDF-GC against four baseline data selection strategies – Random, DSIR, ETA-DACP, and PPL – when applied to the *Fincorpus* financial dataset. All experiments were conducted using the *Llama-3.2-1B* model as the base and employed identical configurations in CPT and evaluation. The best and second-best performance metrics achieved by each method are highlighted in bold and underlined text, respectively.

Evaluation Benchmarks The model's performance evolution during the CPT process is evaluated from two perspectives: domain-specific question answering (Domain QA) and general question answering (General QA). For domain-specific QA, the *CFinBench* dataset (Nie et al., 2024) is used. For general QA, the evaluation utilizes the *MMLU* (Hendrycks et al., 2021), *CMMLU* (Li et al., 2024a), *Xiezhi* (Gu et al., 2024), and *ICLEval* (Chen et al., 2025) datasets. All evaluations adopt the *Few-shot* question answering paradigm, with *Exact Match* as the accuracy metric. Further details on the test sets and experimental setup are provided in Appendix B.3.

Baselines In our experiments, we compare the performance of our method against the **Base Model** (trained without any data selection) and the **Full-scale** model (trained on the entire dataset). Additionally, we include the following established baseline data selection methods for comparison: **Random** sampling, **DSIR** (Xie et al., 2023), **ETA-DACP** (Xie et al., 2024), and **PPL** (Marion et al., 2023; Yin and Rush, 2024). More comprehensive details regarding these baseline methods can be found in Appendix B.4.

3.2 Comparison of Data Selection Methods

To investigate the performance of CDF-GC in domain-specific data selection, we compared it against four baseline strategies. Specifically, we obtain a 20% (2B tokens) data subset from the *Fincorpus* dataset by applying our CDF-GC and four baseline methods individually. Subsequently, we continual pre-trained *Llama-3.2-1B* by each of these subsets, and evaluated the models after CPT. The results of the experiment are presented in Table 1.

The results show that, under the same selection budget, CDF-GC improves by 2.0% on domain-specific QA and 0.9% on the domain-general comprehensive capability compared to the best-performing baseline. When compared to full-scale training, CDF-GC achieves a 1.7% improvement in domain-specific QA and a 0.8% improvement in domain-general comprehensive capability, despite selecting only 20% of the data. We also report our results of generalizing our methods to *Qwen2.5-0.5B* model in Appendix C.

3.3 The Impact of Hard Sample Budget Ratio

In the Balanced Sampling step (subsection 2.4), we used a combination of Hard Sample and CDF Sample. To investigate the impact of the ratio of budget for hard sampling p ($p \in [0, 1]$) on model performance, this experiment extracted 20% of the data (1B tokens) from the *Fincorpus* dataset at hard sampling budget ratios of $p = [0, 0.2, 0.4, 0.6, 0.8, 1.0]$. Subsequently, we continual pre-trained *Llama-3.2-1B* by the different sampling results at different p value, and evaluated the models after CPT. We plotted line graphs showing the performance across different benchmarks for various values of p. The experimental results are shown in Figure 2.

The results indicate that in most scenarios, an appropriate value of p (ranging from 0.4 to 0.6) effectively balances data complexity and diversity, maximizing model capability enhancement. Both excessively high and low p values hinder model performance. For domain-specific task (*CFinBench*) and some general tasks (*MMLU*, *CMMLU*), $p = 0.4$ is a favorable choice. Notably, when $p = 1$, it corresponds to pure hard sampling, which can be regarded as an ablation experiment on the Balanced Sampling stage. However, optimal p values are task-specific and experimental, with *Xiezhi* requiring $p = 1.0$ to maximize GC retention. Our analysis of this phenomenon suggests that the reasons may be as follows: Excessively high p values may lead to overly complex data and insufficient diversity, making the model difficult to fit and reducing its generalization ability. On the other hand, too low a p value may result in insufficient data complexity, failing to include

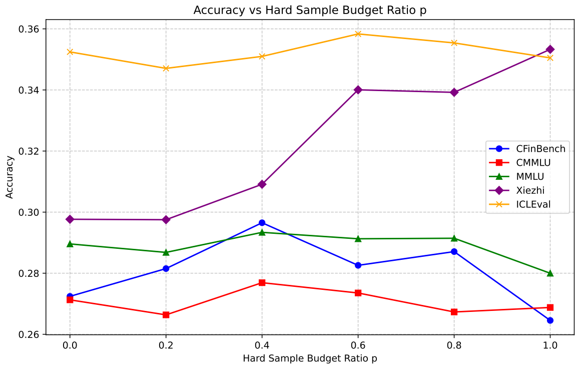

Figure 2: The impact of budget ratio for hard sampling. The x-axis represents the ratio of budget for hard sampling, and the y-axis represents the model performance (accuracy) on benchmarks after CPT. The five lines correspond to five different benchmarks.

enough knowledge.

3.4 Impact of Balanced Sampling on Data Distribution

To investigate the effect of this balanced sampling method on the distribution of grammatical complexity, we extracted a chunk of 128,437 samples from the `fin_articles` subset of *Fincorpus* and conducted a 20% token-budget (budget ratio for hard sampling $p = 0.4$) sampling experiment. We then computed the overall grammatical complexity and its individual dimensions for both the original slice and the sampled data. The results are shown in Figure 3.

The results indicate that the balanced sampling method effectively increased the overall grammatical complexity, causing the distribution curve to shift toward the higher score range, while simultaneously ensuring the balance of the data distribution.

3.5 Trend of Model Performance with Increasing Selection Ratio

To systematically evaluate the impact of the selection ratio on the model performance for the CDF-GC method, we designed the following experimental setup: Using *Fincorpus* as the training dataset, we constructed training subsets with six different selection ratios of $[5\%, 10\%, 15\%, 20\%, 25\%, 30\%]$. Based on the *Llama-3.2-1B* base model, we performed CPT on the data subsets for each selection ratio and evaluated the model performance on both domain and general benchmarks. Concurrently, we also compared the training curves of CDF-GC against those of Full-scale training. Figure 4 shows that the CDF-GC method achieves its peak performance at approximately a 20% data selection ratio, a performance that notably surpasses the full-scale training result.

Since CDF-GC uses an offline sampling strategy, it requires all data sampling to be completed at once, making it unsuitable for dynamic data loading methods that sample and train concurrently. This characteristic results in the need to restart continuous pre-training from scratch for each selection ratio during the experiment. Due to resource constraints, our experiment did not cover higher selection ratios.

3.6 The Relationship between Grammatical Complexity and Length

To investigate the potential correlation between the dimensions of grammatical complexity and text length (especially considering that longer texts often correspond to higher lexical entropy), we conducted a systematic analysis of their relationship using Pearsons correlation coefficient. The experimental data was sourced from the Fincorpus dataset, and 2,000 text samples were obtained via stratified sampling, with character lengths uniformly distributed in the range $[300, 8000]$. The correlation results for each component of grammatical complexity with text length are visualized in Figure 5.

The analysis reveals a significant positive correlation between lexical entropy and text length (r = 0.692), while other grammatical complexity components and the overall index show weaker correlations with length (|r| < 0.3). To control for potential bias introduced by text length when evaluating grammatical complexity, the data can be preprocessed by segmenting it based on text length prior to applying the CDF-GC method.

3.7 Correlation Analysis of Grammatical Complexity Components

To investigate the interrelationships among the individual components of our Grammatical Complexity (GC) framework, we performed a correlation analysis. We computed the pairwise Pearson correlation coefficients between the five GC components using 128,437 instances from the `fin_articles` subset of *Fincorpus*. The resulting correlation matrix, visualized as a heatmap in Figure 6, illustrates the linear relationships between various grammatical complexity measures within our financial domain dataset.

The analysis revealed that all pairwise Pearson correlation coefficients between the five grammatical complexity metrics were consistently below 0.7. This limited correlation, lacking strong linear relationships, empirically supports that these metrics, despite some conceptual overlap, predominantly capture distinct facets of grammatical complexity. For instance, even the highest observed correlations – 0.55 between *part-of-speech entropy* and *dependency label entropy*, and 0.65 between *average dependency distance* and *average dependency tree height* – indicate moderate rather than strong linear associations. This demonstrates that while these pairs might quantify related aspects of complexity, they do so from unique angles. This experimental finding robustly validates our framework's design: by integrating these diverse yet complementary metrics, we achieve a comprehensive and multi-dimensional assessment of grammatical complexity, moving beyond

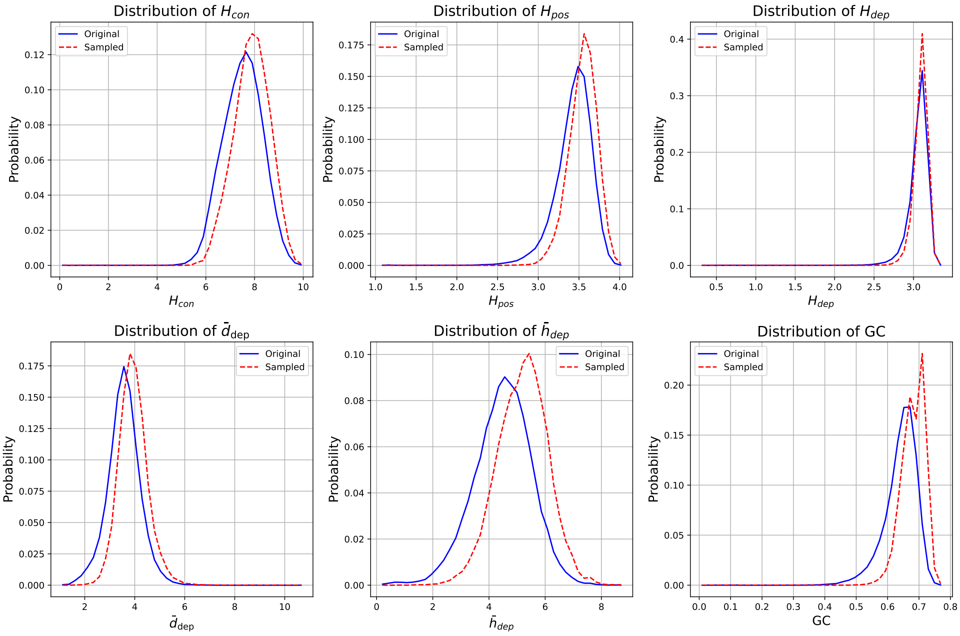

Figure 3: Distribution of Grammatical Complexity Components. The above figure shows the distribution of grammatical complexity and its five components before and after sampling. We divided the data range into 40 bins, calculated the probability distribution of each bin, and finally plotted the results as a line chart.

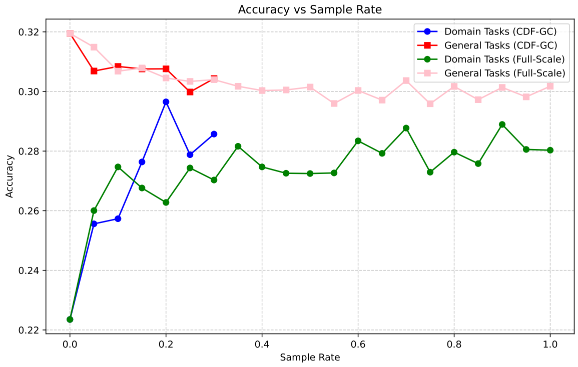

Figure 4: This figure plots the trend of model performance increasing with the growth of the data usage ratio. As indicated in the legend in the upper right corner, the four lines represent the model's performance on domain and general tasks after CPT using either the full dataset or data filtered by CDF-GC.

the limitations of a single, narrow perspective.

4 Related Works

Data Quality Data quality measures the "noisiness", "correctness" or distributional alignment of data (Havrilla et al., 2024). High-quality data enhances in-distribution generalization (Yu et al., 2024; Thrush et al., 2024). Many studies (Liu et al., 2024; Lozhkov et al., 2024; Du et al., 2023; Cao et al., 2024) employ LLMs or reward models for quality filtering. Uesato et al. (2022) proposed a method combining supervised learning and reward-model-based reinforcement learning to solve mathematical word problems, emphasizing the importance of high-quality process-supervised data in training. Meanwhile, other approaches select data based on distributional alignment. Traditional methods (Moore and Lewis, 2010; Axelrod, 2017) measure alignment scores via cross-entropy loss differences. Xie et al. (2023) quantifies alignment probability using Hashed N-gram features. Ni et al. (2022); Xie et al. (2024); Xia et al. (2024) assess alignment through document embedding similarity. For synthetic data, larger LLMs generally produce higher-quality outputs than smaller ones (Yang et al., 2024b; Qu et al., 2024; Du et al., 2024). However, quality-diversity trade-offs emerge during filtering, as quality improvements often reduce diversity (Longpre et al., 2024).

Data Diversity Data diversity measures the "self-similarity" and "coverage" of data (Havrilla et al., 2024). High-diversity data improves out-of-distribution generalization (Ye et al., 2024; Samvelyan et al., 2024). Information entropy serves as an effective metric for diversity quantification (Bengio et al., 2009). Lexical diversity can quantify text data diversity, with high lexical diversity data facilitating better word representation learning (Tsvetkov et al., 2016). Furthermore, Xie et al. (2024) demonstrated that selecting financial domain data with high part-of-speech entropy effectively enhances LLMs' domain QA capability. Regarding quality-diversity trade-offs, Liu et al. (2024); Xie et al. (2024); Du et al. (2023) achieve diverse sampling by selecting both high-quality core sets and their neighborhood data through document embedding vectors.

Data Complexity Data complexity measures the

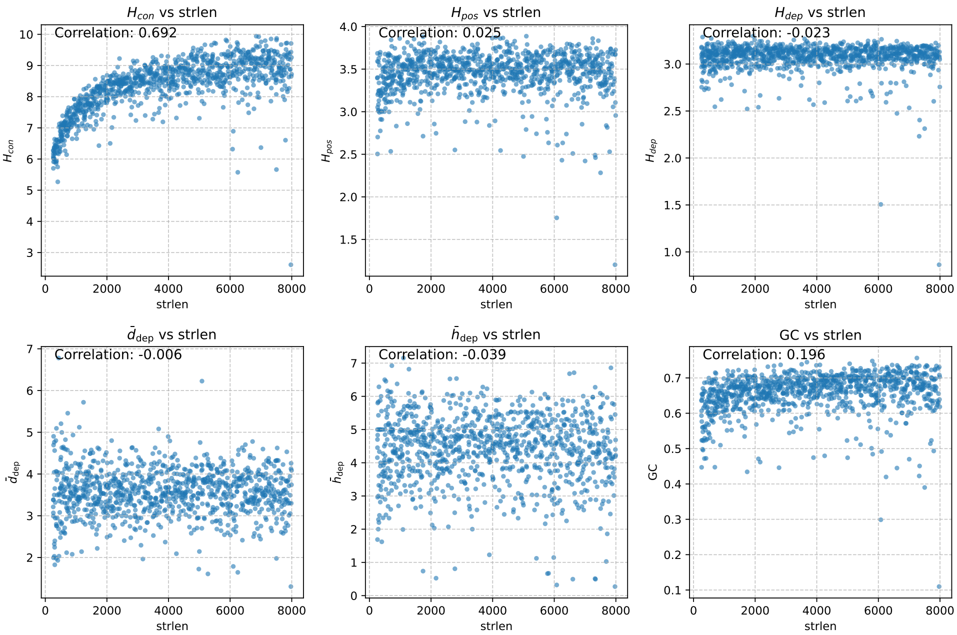

Figure 5: Scatter plot of the correlation between grammatical complexity components and text length, along with Pearson's correlation coefficient.

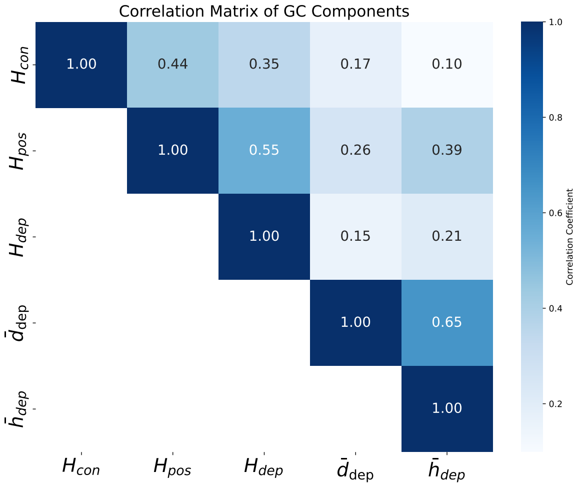

Figure 6: Correlation Matrix of Grammatical Complexity Components. This heatmap illustrates the pairwise Pearson correlation coefficients between the five individual components of our Grammatical Complexity framework, calculated on a chunk of the *Fincorpus* dataset.

"difficulty" or "compositionality" of samples (Havrilla et al., 2024), which can be quantified by metrics such as length (Cao et al., 2024), number of phrases , and syntactic tree depth (Tsvetkov et al., 2016). An appropriate level of complexity can enhance a model's in-distribution and out-of-distribution generalization, as complex data encourages the model to learn deeper features and patterns (Xu et al., 2024). Li et al. (2024b) improved instruction tuning by filtering complex instructions using Instruction Following Difficulty (IFD). Perplexity can measure data complexity for a specific model (Wenzek et al., 2019). Based on the hypothesis that "high-perplexity data has greater learning value for models" (Bengio et al., 2009), Marion et al. (2023) found that training models on the top 30% highest-perplexity data led to lower test loss compared to using the full dataset. Yin and Rush (2024) observed that perplexity-based filtering benefits large-scale models more significantly than smaller ones. However, excessive complexity may cause overfitting and reduce generalization (Cao et al., 2024).

5 Conclusion

The CDF Sampling with Grammatical Complexity (CDF-GC) data selection method proposed in this study extracts data containing deep domain knowledge through multi-dimensional grammatical complexity metrics and balances data complexity and diversity using a CDF-based sampling strategy. The experiments show that, in the scenario of performing CPT on a 1B LLM using financial domain data, under the same selection ratio, this method improves accuracy on the domain-specific QA test set by 2.0% compared to the optimal baseline, and enhances domain-general capabilities by 0.9%. Besides, this method selects only 20% of the data, and compared to full-scale training, domain-specific QA accuracy improves by 1.7% and domain-general capabilities increase by 0.8%.

6 Limitations

Grammatical Complexity Completeness This study quantifies grammatical complexity (GC) using two dimensions: lexical diversity and syntactic complexity.

Although this framework provides a relatively comprehensive evaluation, there is still significant room for improvement. In terms of lexical diversity, further indicators such as the number of token type, token type entropy, type-token ratio, and N-gram types could be introduced to enhance the precision of the evaluation. In terms of syntactic complexity, the verb ratio and other related metrics are also worth considering. Moreover, since GC is a multi-dimensional composite index, it is difficult to quantify using a single dimension, which makes it challenging to accurately measure the independent contribution of each specific metric to the final selection effect through ablation experiments.

Objectivity of Grammatical Complexity This study uses GC as an objective evaluation metric for data quality to reduce dependence on reward models or LLMs. However, the calculation of this metric still relies on traditional natural language processing techniques, including tasks such as tokenization, part-of-speech tagging, and dependency parsing. Although studies have shown that the accuracy of mainstream tokenization and part-of-speech prediction techniques exceeds 95%, and the accuracy of dependency parsing tasks exceeds 90% (Che et al., 2021; Zhang et al., 2020a,b; Sun; Bird, 2006), we must acknowledge that errors in the underlying models may have a slight impact on the calculation of GC. Furthermore, as we predict sentence boundaries based on newline characters and certain punctuation marks, unexpected punctuation usage within the document could lead to deviations in sentence segmentation, potentially introducing errors in dependency parsing predictions. Based on the current maturity of the technology, we believe this metric has a high degree of objectivity, but future research could further explore the cumulative effect of syntactic analysis errors.

Semantic Processing Limitations While our method leverages GC to select information-rich text, and GC does reflect some semantic aspects (e.g., lexical diversity, syntactic complexity), it fundamentally struggles with deeper semantic understanding. Specifically, metrics like content word entropy and syntactic complexity operate at the lexical and syntactic levels, unable to directly assess the true semantic depth. For instance, sentences with similar GC can have vastly different semantic contentone concrete and the other abstract, requiring inference – a distinction our method cannot make. Furthermore, by focusing on isolated sentences, it neglects crucial discourse – level information such as coherence, context dependency, and pragmatics, all vital for comprehensive semantic comprehension. Therefore, despite its strength in selecting syntactically and lexically complex texts, the method is limited in tasks requiring advanced semantic understanding. Future work could integrate semantic analysis techniques into the data selection framework.

Scale Limitation Due to experimental constraints , this study only conducted CPT experiments on LLMs with a parameter size of 1B, using training dataset of 10B tokens. This limitation prevented us from verifying the applicability of the CDF-GC method on larger LLMs or on larger datasets.

Domain Limitations Due to experimental constraints, this study exclusively validated the data selection performance of the CDF-GC method within the financial domain. While we anticipate the applicability of this method to other domains, its efficacy in such contexts necessitates further empirical investigation. Besides, while we aimed to enhance the model's ability to leverage domain knowledge without compromising its general capability, the experimental results in Table 1 indicate that the CDF-GC method did not achieve sufficiently high performance on certain general tasks when filtering domain-specific data.

Data Diversity Limitation Although experiments demonstrate that our Balanced Sampling method can enhance overall GC and shift the distribution curve toward the higher score range, while ensuring the balance of GC and its individual dimension distributions, this balanced distribution of quantitative metrics does not necessarily reflect the diversity of data in terms of knowledge coverage, topic coverage, and other aspects.

References

Amittai Axelrod. 2017. Cynical selection of language model training data. *CoRR*, abs/1709.02279.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Steven Bird. 2006. NLTK: the natural language toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2024. Instruction mining: Instruction data selection for tuning large language models. *Preprint*, arXiv:2307.06290.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: An open-source neural language technology platform for Chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wentong Chen, Yankai Lin, ZhenHao Zhou, HongYun Huang, Yantao Jia, Zhao Cao, and Ji-Rong Wen. 2025. Icleval: Evaluating in-context learning ability of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10398–10422. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Ian Connick Covert, Wenlong Ji, Tatsunori Hashimoto, and James Zou. 2024. Scaling laws for the value of individual data points in machine learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Seamus Donnelly, Evan Kidd, Jay Verkuilen, and Caroline Rowland. 2025. The separability of early vocabulary and grammar knowledge. *Journal of Memory and Language*, 141:104586.

Chengyu Du, Jinyi Han, Yizhou Ying, Aili Chen, Qianyu He, Haokun Zhao, Sirui Xia, Haoran Guo, Jiaqing Liang, Zulong Chen, Liangyue Li, and Yanghua Xiao. 2024. Think thrice before you act: Progressive thought refinement in large language models. *Preprint*, arXiv:2410.13413.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Duxiaoman-DI. 2023. Fincorpus.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, Qianyu He, Rui Xu, Wenhao Huang, Jingping Liu, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, and Yanghua Xiao. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18099–18107. AAAI Press.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, Duy Phung, Maia Iyer, Dakota Mahan, Chase Blagden, Srishti Gureja, Mohammed Hamdy, Wen-Ding Li, Giovanni Paolini, Pawan Sasanka Ammanamanchi, and Elliot Meyerson. 2024. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *CoRR*, abs/2412.02980.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 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, Toronto, Canada, July 9-14, 2023*, pages 8003–8017. Association for Computational Linguistics.

Shadi Iskander, Sofia Tolmach, Ori Shapira, Nachshon Cohen, and Zohar Karnin. 2024. Quality matters: Evaluating synthetic data for tool-using llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4958–4976. Association for Computational Linguistics.

Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2024. Data-centric artificial intelligence. *Bus. Inf. Syst. Eng.*, 66(4):507–515.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024a. CMMLU: measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11260–11285. Association for Computational Linguistics.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024b. 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7602–7635. Association for Computational Linguistics.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3245–3276. Association for Computational Linguistics.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025. URSA: understanding and verifying chain-of-thought reasoning in multimodal mathematics. *CoRR*, abs/2501.04686.

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. *CoRR*, abs/2309.04564.

Robert C. Moore and William D. Lewis. 2010. Intelligent selection of language model training data. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, Short Papers*, pages 220–224. The Association for Computer Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9844–9855. Association for Computational Linguistics.

Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, Yunhe Wang, and Dacheng Tao. 2024. Cfinbench: A comprehensive chinese financial benchmark for large language models. *CoRR*, abs/2407.02301.

John A. O'Leary and Rasmus Steinkrauss. 2022. Syntactic and lexical complexity in l2 english academic writing: Development and competition. *Ampersand*, 9:100096.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024a. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíek, Vinko Sabolec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024b. Fineweb2: A sparkling update with 1000s of languages.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. 2024. Recursive introspection: Teaching language model agents how to self-improve. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Tim Rocktäschel, and Roberta Raileanu. 2024. Rainbow teaming: Open-ended generation of diverse adversarial prompts. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Junyi Sun. Jieba.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. 2024. Improving pretraining data using perplexity correlations. *CoRR*, abs/2409.05816.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the curriculum with bayesian optimization for task-specific word representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Y. Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *CoRR*, abs/2211.14275.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *Preprint*, arXiv:2411.12372.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *Preprint*, arXiv:1911.00359.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *CoRR*, abs/2304.14454.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 10184–10201. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhao Xue, Hanyu Zhao, Sha Yuan, and Yequan Wang. 2022. WuDaoCorpora Text.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yuqing Yang, Yan Ma, and Pengfei Liu. 2024b. Weak-to-strong reasoning. *Preprint*, arXiv:2407.13647.

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

Junjie Oscar Yin and Alexander M. Rush. 2024. Compute-constrained data selection. *CoRR*, abs/2410.16208.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Ben Hu. 2025. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.*, 57(5):129:1–129:42.

Yu Zhang, Zhenghua Li, and Zhang Min. 2020a. Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of ACL*, pages 3295–3305.

Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020b. Fast and accurate neural CRF constituency parsing. In *Proceedings of IJCAI*, pages 4046–4053.

Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. 2024. Continual learning with pre-trained models: A survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 8363–8371. ijcai.org.

A Sampling Budget Controlling

The parameter r is used to control the sampling budget. When $r = 1$, the expected number of sampled tokens is:

$$E(T) = \sum_{i=1}^{n} \mathrm{CDF}(\mathrm{GC}(x_i))\mathrm{TC}(x_i).$$

Here, $\mathrm{TC}(\cdot)$ represents the token count of a sample or dataset. When all data samples have the same token count, the expected value of $E(T)$ is calculated as follows:

$$\begin{aligned} E(T) &= \sum_{i=1}^{n} \frac{i}{n} \cdot \frac{\mathrm{TC}(\mathcal{D}')}{n} \\ &= \frac{\mathrm{TC}(\mathcal{D}')}{n^2} \sum_{i=1}^{n} i \\ &= \frac{\mathrm{TC}(\mathcal{D}')}{n^2} \cdot \frac{n(n+1)}{2} \\ &= \frac{n+1}{2n} \mathrm{TC}(\mathcal{D}'). \end{aligned}$$

Let the sampling token budget be T_{cdf}. When $T_{\text{cdf}} \leq E(T)$, we set $r = T_{\text{cdf}}/E(T)$. In this case:

$$\begin{aligned} P(x_i) &= \min(r \times \text{CDF}(\text{GC}(x_i)), 1) \\ &= r \times \text{CDF}(\text{GC}(x_i)), \\ E_r(T) &= \sum_{i=1}^{n} r \times \text{CDF}(\text{GC}(x_i))\text{TC}(x_i) \\ &= r \times E(T) \\ &= T_{\text{cdf}}. \end{aligned}$$

When $T_{\text{cdf}} > E(T)$, we also set $r = T_{\text{cdf}}/E(T)$. In this case:

$$P(x_i) \leq r \times \text{CDF}(\text{GC}(x_i)), E_r(T) \leq T_{\text{cdf}}.$$

Based on the above derivation, in general, $E(T) \approx \text{TC}(\mathcal{D}')/2$. When the sampling ratio is below 50% (e.g., 10%, 20%), this method effectively controls the sampling budget. However, when the sampling ratio exceeds 50%, $r > 1$ may result in an undersampling. Although this method does not precisely control the number of sampled tokens, in large-scale training scenarios, small deviations in the sampling volume have minimal impact on the training results. The final sampling volume can be further calibrated through uniform random sampling.

B Experiment Setup Details

B.1 Continual Pre-training Configurations

Our computational setup consists of 4 GPUs, and we utilize the **deepspeed** distributed training framework. We perform continual pretraining on both the *Llama-3.2-1B* and *Qwen2.5-0.5B* models, with each data sample having a length of 4096 tokens. The main training parameters are provided in Table 2.

B.2 Training Data Sources

Fincorpus (Duxiaoman-DI, 2023): A financial domain-specific corpus (including financial reports, news, and research papers) with an original size of 15B tokens. After preprocessing (removing texts longer than 8192 tokens and performing data cleaning), we retain 10B tokens of high-quality data as the source of financial domain corpora.

B.3 Evaluation Benchmarks

The following provides a detailed introduction to the test sets:

- **CFinBench** (Nie et al., 2024): This is a benchmark dataset in the financial domain, specifically designed to evaluate models' performance on financial text understanding, classification, and question answering tasks. CFinBench includes various finance-related tasks, such as financial report analysis and financial news comprehension.

- **MMLU** (Hendrycks et al., 2021): MMLU (Massive Multitask Language Understanding) is a benchmark test for evaluating the multitask understanding capability of large language models (LLMs) in English. It covers 57 tasks across a wide range of domains, including basic mathematics, American history, computer science, law, and more.

- **CMMLU** (Li et al., 2024a): CMMLU (Chinese Massive Multitask Language Understanding) is a comprehensive benchmark dataset for evaluating the multitask language understanding capabilities of LLMs in Chinese. It spans various fields, including natural sciences, social sciences, engineering, and humanities, with the aim of providing a holistic measure of LLMs' performance across different subjects and scenarios.

- **Xiezhi** (Gu et al., 2024): Xiezhi is a comprehensive domain knowledge evaluation benchmark containing 249,587 multiple-choice questions across 516 subjects. It is used to assess the knowledge mastery of large language models in various fields. The language is in both Chinese and English. For evaluation, we use a manually curated subset (40,000 questions).

- **ICLEval** (Chen et al., 2025): ICLEval is a benchmark dataset designed to evaluate LLMs' In-Context Learning (ICL) capabilities. It includes 12 tasks and 2,040 test samples, primarily evaluating LLMs' ICL abilities in terms of exact copying and rule learning.

B.4 Baselines

The following provides a detailed introduction to the Baseline methods: **Base Model** refers to the base model that has not undergone Continual Pretraining (CPT).

Full-scale applies CPT using the full set of data sources for the model.

Random involves uniformly randomly sampling a fixed amount of data from the original dataset.

DSIR (Xie et al., 2023) (Data Selection via Importance Resampling) is a domain adaptation data selection method that requires a target dataset. It calculates the Hashed N-gram features of the data and uses importance resampling to select data from the source dataset that is similar in distribution to the target dataset. In this case, we use the training set of the CFinBench financial domain benchmark data as the target dataset.

ETA-DACP (Xie et al., 2024) (Efficient Task-Agnostic Domain-adaptive Continual Pre-training) is a method that uses part-of-speech entropy as a reference for data diversity, combined with either hard or soft sampling to efficiently select domain-adaptive data. In this study, we use the part-of-speech entropy combined with hard sampling method, which showed the best per-

LR	SeqLen	Batch Size	Optimizer	Scheduler	Warmup
5e-5	4096	1.6M (tokens)	AdamW	Cosine Annealing (min ratio = 0.1)	0.1

Table 2: Continual Pre-training Configurations

formance in domain knowledge question answering in the original paper, as a comparison baseline.

PPL (Marion et al., 2023; Yin and Rush, 2024) (Perplexity) refers to the method of calculating the perplexity for each data sample using the model to be trained, and then selecting the samples with the highest perplexity for continual pretraining. This method automatically identifies the data most valuable for training a specific model.

C Generalization Verification of CDF-GC

To verify the generalizability of our proposed method, we conducted comparative experiments on different data selection methods under the identical experimental setup as subsection 3.2, with only the base model replaced by *Qwen2.5-0.5B*. The results are presented in Table 3. It is shown that our method CDF-GC outperforms all baseline methods in both domain-specific tasks and comprehensive capabilities. Notably, due to the Qwen series models already possessing strong Chinese processing capabilities and the constraints imposed by the quality of the original training dataset, the performance improvement of the model through continual pre-training is relatively limited.

D Training loss comparison between CDF-GC and Full-scale

During CPT with data selected by CDF-GC, we observed elevated training loss values compared to full-scale. To systematically analyze this phenomenon, we present comparative training loss curves between CDF-GC and full-scale training in Figure 7.

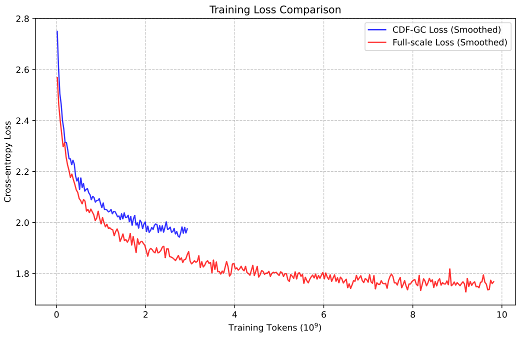

Figure 7: Training loss curves (cross-entropy) comparing CDF-GC with full-scale baseline. All trajectories are smoothed to reveal underlying trends.

The training curves reveal that CDF-GC-selected data exhibits slower loss convergence compared to full-dataset training during CPT, which we attribute to the inherent complexity of high-GC samples containing more sophisticated linguistic patterns and presenting greater learning difficulty. Notably, despite showing a final loss gap, CDF-GC achieves improvement on domain-specific QA benchmarks (CFinBench), suggesting that partial acquisition of complex features suffices for downstream performance gains and that the method's selective pressure effectively identifies pedagogically valuable samples even without complete convergence.

Considering that cross-entropy loss is related to perplexity ($PPL = exp(loss)$), which is also an effective data complexity metric, we conjecture that the high-GC characteristic of data may be associated with high PPL. Therefore, we sampled 1,000 data points from the `fin_articles` subset of *Fincorpus*, plotted a scatter diagram of GC versus PPL, and calculated the correlation coefficient. However, the results reveal only weak correlation between GC and PPL ($r = 0.0479$), suggesting that GC score distinct from perplexity. The experimental results are shown in Figure 8:

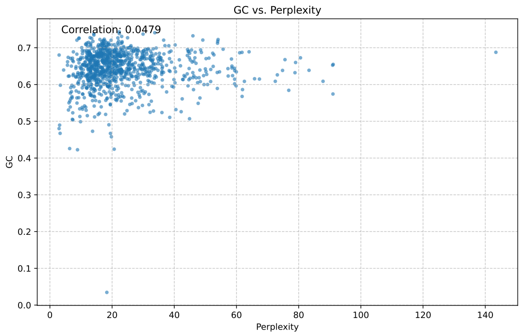

Figure 8: Correlation analysis between grammatical complexity (GC) and perplexity (PPL) for 1,000 randomly sampled documents from the `fin_articles` subset.

Model	Domain	General Benchmarks				Avg.
	CFinBench	CMMLU	MMLU	Xiezhi	ICLEval	
Base Model	38.34	49.47	44.74	54.22	39.85	45.32
Random	36.71	<u>45.84</u>	41.38	50.37	37.99	42.46
DSIR	37.54	45.25	41.79	49.74	38.53	42.57
ETA-DACP	37.46	45.53	**42.41**	**50.74**	**39.07**	<u>43.04</u>
PPL	<u>38.08</u>	45.69	42.00	49.85	<u>38.77</u>	42.88
CDF-GC (Ours)	**38.52**	**45.86**	<u>42.31</u>	<u>50.65</u>	38.53	**43.17**

Table 3: Performance comparison of data selection methods for continual pre-training in the financial domain. This table presents a comparative evaluation of CDF-GC against four baseline data selection strategies – Random, DSIR, ETA-DACP, and PPL – when applied to the *Fincorpus* financial dataset. All experiments were conducted using the *Qwen2.5-0.5B* model as the base and employed identical configurations in CPT and evaluation. The best and second-best performance metrics achieved by each method are highlighted in bold and underlined text, respectively.