WordNet-Guided Domain Data Selection for LLM Continual Pre-training

Anonymous authors
Paper under double-blind review

Abstract

Continual pre-training (CPT) adapts language models to specialized domains, but selecting useful training data from large domain corpora remains difficult. We introduce Frequency-Conceptual Difficulty (FCD), a model-free metric that combines lexical frequency with semantic distance in WordNet to score document difficulty. In the medical domain, FCD-Hard selects a fixed one-third subset and trains on it for three epochs. Under the same total token exposure, this strategy improves domain-average accuracy over full-corpus CPT by 1.36 percentage points and outperforms the compared data-selection baselines. We also examine whether the preferred FCD range is domain dependent. Medical performance is highest when training on documents with rare, WordNet-distant terminology, whereas FCD-Easy performs best in our experiment on an English translation of FinEval. Qualitative examples suggest that some high-FCD documents in the financial corpus contain technical vocabulary from other domains. These results support using FCD as a lexical criterion for domain-data analysis while showing that the sampling direction must be selected for the target corpus.

1 Introduction

Large Language Models (LLMs) perform well on many general tasks (Brown et al., 2020; Chowdhery et al., 2023), but adapting them to specialized domains remains challenging. Alongside work on model architectures (Vaswani et al., 2017; Cui & Yao, 2024), recent research has increasingly examined how training data affect model behavior (Zha et al., 2025; Jakubik et al., 2024). This question is particularly relevant in medicine, where general-purpose models can lack specialized knowledge (Guo et al., 2025b). **Continual Pre-training (CPT)** on domain corpora is a common approach to such adaptation (Jin et al., 2022; Qiu et al., 2024).

Selecting data for domain-specific CPT requires deciding which documents are most useful under a fixed training budget. Existing methods provide several signals for this decision: (1) **Rule-based methods** (Penedo et al., 2024; 2025; Li et al., 2025) filter undesirable content using manually designed criteria, but these criteria do not directly measure domain specificity. (2) **Model-based methods**, such as LLM scoring (Marion et al., 2023; Zhuang et al., 2025; Chen et al., 2025; Peng et al., 2025) or Reward Models (Penedo et al., 2024), depend on the auxiliary model and add inference cost at corpus scale. (3) **Domain-similarity approaches** (Xie et al., 2023; 2024) compare candidate data with a target distribution and therefore require a representative seed set.

Computational linguistics offers structured resources that provide a different selection signal. For example, **WordNet** (Miller, 1995) serves as a semantic network that organizes words into sets of synonyms (synsets) linked by conceptual-semantic and lexical relations. Similarly, **Probase** (Wu et al., 2012) provides a massive-scale taxonomy that maps millions of entities to their corresponding concepts. A question is whether these semantic resources can be used to guide the selection of training corpora for LLMs.

We study this question through **lexical difficulty**. Motivated by the account in cognitive psychology that **basic concepts** form the scaffolding for complex expertise (Rosch et al., 1976), we hypothesize that the technical density of a document is reflected by its concentration of vocabulary that deviates from these foundational terms. To implement this, we introduce the **Frequency-Conceptual Difficulty (FCD)** metric. FCD quantifies document difficulty by integrating statistical frequency with the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

conceptual distance from basic concepts in WordNet. It can be computed without model inference or annotated training examples.

The main contributions of our work are summarized as follows:

- We propose the Frequency-Conceptual Difficulty (FCD), a metric for estimating the lexical difficulty of domain corpora. By integrating word frequency with the conceptual distance from basic concepts within the WordNet semantic network, FCD provides a document-level score at low computational overhead (e.g., scoring 6B tokens in only 41.6 CPU-hours).
- We evaluate three CPT strategies: FCD-Hard, which selects the highest-scoring documents; FCD-Soft, which samples in proportion to score; and FCD-CL, which orders training by difficulty. On medical benchmarks, FCD-Hard improves the domain average over full-corpus CPT under the same total training-token budget.
- We compare FCD with perplexity and examine its behavior across domains. High-FCD selection performs best in the medical experiments, while low-FCD selection performs best in the financial experiment, indicating that the preferred score range depends on the target corpus.

2 Related Work

2.1 Data Selection for Training LLMs

Data selection for LLMs generally evaluates three dimensions: quality, diversity, and complexity (Havrilla et al., 2024). Current methodologies range from rule-based filtering and reward-based scoring (Penedo et al., 2024; 2025) to LLM-based selection for instruction tuning (Liu et al., 2024; Du et al., 2023; Cao et al., 2024). For domain-specific scenarios, selection often relies on distributional alignment via N-gram similarity (Xie et al., 2023) or embedding-based metrics (Xie et al., 2024; Xia et al., 2024). Other work has introduced multi-actor frameworks for collaborative optimization of these dimensions (Havrilla et al., 2024). While these methods focus on alignment or general quality, our approach uses corpus frequency and semantic-network distance to measure lexical difficulty.

2.2 Word Difficulty

Word difficulty has been studied in Natural Language Processing and language acquisition, where frequency is a common proxy for domain-specific expertise (Stewart et al., 2022; Zhang et al., 2022a; Desai et al., 2021). Frequency predicts acquisition order in language models (Chang & Bergen, 2022) and has been used as a signal in training and inference (Zhang et al., 2022b; Zhao et al., 2025; Yu et al., 2022). Model performance can also vary across frequency bands, with weaker coverage of rare domain terms (Su et al., 2024). Frequency alone does not represent conceptual structure (Ha et al., 2024; Hashimoto, 2021; Chen et al., 2019; Katsumata et al., 2018; Dascalu et al., 2016). Unlike frequency-only measures, FCD combines lexical rarity with WordNet-based semantic distance.

3 Method

This section defines Frequency-Conceptual Difficulty (FCD). The metric has two components: the frequency dimension (§3.1.1) measures rarity, while the conceptual dimension (§3.1.2) measures semantic distance in WordNet.

We first define separate word-level scores for nouns and non-nouns (§3.1.3), and then aggregate them into a document score (§3.1.4).

Three FCD-based CPT strategies are defined in §3.2.

3.1 Frequency-Conceptual Difficulty

Frequency-Conceptual Difficulty (FCD) characterizes document vocabulary using two dimensions: word frequency and semantic distance from "basic concepts" (Rosch et al., 1976) via WordNet (Miller, 1995).

3.1.1 FREQUENCY DIMENSION

We use word frequency statistics from the Google Books Ngram dataset (Michel et al., 2011) to quantify lexical rarity. Guided by ZIPF'S LAW (Zipf, 1949), we apply a logarithmic transformation to the raw frequencies and employ a Sigmoid function to normalize them into a unified range $[0, 1]$. The transformation is most sensitive to differences in the middle-frequency range. The frequency difficulty score $d_f(w)$ is defined as:

$$d_f(w) = 1 - \sigma(k(\log f(w) - c)), \tag{1}$$

where $\sigma(x) = 1/(1 + e^{-x})$ denotes the Sigmoid function; k controls the slope of the curve and is set to 0.7. Sensitivity results for k and the other parameters are reported in subsection 6.4; c represents the central parameter of the Sigmoid function, defined as the logarithmic word frequency corresponding to the 0.1-quantile of word frequencies in the basic vocabulary set $\mathcal{W}_{\text{basic}}$ (detailed in §3.1.2).

3.1.2 CONCEPTUAL DIMENSION

The conceptual dimension characterizes lexical difficulty by measuring a word's semantic distance from "basic concepts" (Rosch et al., 1976). We implement this using the WordNet (Miller, 1995) noun semantic network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ represent synset nodes and hypernym-hyponym relations. We apply graph-based distance only to nouns because the WordNet noun graph provides the hierarchy used by the metric.

To construct the reference set of basic concepts, we use the Oxford Dictionary to select words at the A1/A2 levels under the CEFR standard (Figueras et al., 2013), as primary school-level vocabulary aligns with the scope of foundational human concepts (Rosch, 1975). These words are mapped to WordNet synsets to form the basic node set: $\mathcal{V}_{\text{basic}} = \{v \in \mathcal{V} \mid \text{name}(v) \in \mathcal{W}_{\text{basic}}\}$.

The conceptual difficulty of any synset v is then defined by its **minimum semantic distance (MSD)** to $\mathcal{V}_{\text{basic}}$, computed via multi-source Breadth-First Search (details in Appendix B):

$$\text{MSD}(v) = \min_{v_b \in \mathcal{V}_{\text{basic}}} \{\text{dist}(v, v_b, \mathcal{G})\}. \tag{2}$$

For corpus-scale scoring, we do not perform word-sense disambiguation and use the most frequent WordNet noun sense for each noun w. Let $\widetilde{\text{MSD}}(w)$ equal the MSD of this synset when one is available and 2 otherwise. The final conceptual difficulty $d_c(w)$ is calculated by normalizing this word-level distance into a $[0, 1]$ range:

$$d_c(w) = \frac{\log(1 + \widetilde{\text{MSD}}(w))}{\log\left(1 + \max_{v_i \in \mathcal{V}}\{\text{MSD}(v_i)\}\right)}. \tag{3}$$

The logarithmic normalization reduces the influence of large distances. A higher $d_c(w)$ indicates greater semantic distance from the basic-concept set.

3.1.3 FCD AT THE WORD LEVEL

While noun synsets in WordNet form a dense hierarchical network suitable for conceptual assessment, non-noun synsets exhibit sparse associations that preclude reliable distance modeling. For nouns, we combine frequency difficulty $d_f(w)$ and conceptual difficulty $d_c(w)$ using a weighted geometric mean; for non-nouns, only $d_f(w)$ is considered:

$$\text{FCD}(w) = \begin{cases} d_f(w)^{w_f} d_c(w)^{1-w_f} & \text{if } w \text{ is a noun,} \\ d_f(w) & \text{if } w \text{ is not a noun.} \end{cases} \tag{4}$$

We set $w_f = 0.5$, which reduces the noun score to $\sqrt{d_f(w) d_c(w)}$.

3.1.4 FCD AT THE DOCUMENT LEVEL

To assign more weight to high-scoring words in the document score, we use **quantile-based filtering** and the **power mean**.

Formally, we first partition all words in a document doc (excluding punctuation and stop words) into two disjoint multisets: a noun multiset M_{noun} (containing all nouns in doc) and a non-noun multiset $M_{\text{non_noun}}$ (containing all non-nouns in doc). We then apply the word-level FCD calculation (Equation 4) to each word in M_{noun} and $M_{\text{non_noun}}$, resulting in two corresponding score multisets $M^{\text{FCD}}_{\text{noun}}$ and $M^{\text{FCD}}_{\text{non_noun}}$. The overall FCD score of doc is defined as a weighted sum of the aggregated scores of these two sets:

$$\begin{aligned}\text{FCD}(doc) = \beta_{\text{noun}} \times \text{agg}(M^{\text{FCD}}_{\text{noun}}, \tau, w, \alpha) \\ +(1-\beta_{\text{noun}}) \times \text{agg}(M^{\text{FCD}}_{\text{non_noun}}, \tau, w, \alpha),\end{aligned} \tag{5}$$

where $\beta_{\text{noun}} \in [0,1]$ is the weight coefficient for the noun multiset, fixed at $\beta_{\text{noun}} = 0.7$; $\text{agg}(\cdot)$ is a multiset-level score aggregation function that integrates FCD scores of all elements in a score multiset. Its calculation relies on defining $M_{\text{high}}(M,\tau)$ (high-difficulty subset, determined by τ) and $M_{\text{ord}}(M,\tau)$ (ordinary-difficulty subset) via the quantile function:

$$\begin{aligned}M_{\text{high}}(M,\tau) &= \{m \in M \mid m \geq \mathcal{Q}(M,\tau)\}, \\ M_{\text{ord}}(M,\tau) &= M \setminus M_{\text{high}}(M,\tau).\end{aligned}$$

Here, $\tau \in (0,1]$ (fixed at $\tau = 0.9$) is the quantile threshold used to isolate the top 10% of word scores, and $\mathcal{Q}(M,\tau)$ denotes the τ-th quantile of M. Accordingly, $M_{\text{high}}(M,\tau)$ contains the elements in M that fall within the top $(1-\tau)$ proportion, representing high-difficulty word scores, while $M_{\text{ord}}(M,\tau)$ contains the remaining ordinary-difficulty scores. The aggregation formula is:

$$\begin{aligned}\text{agg}(M,\tau,w,\alpha) = w \times \text{power_mean}(M_{\text{high}}(M,\tau),\alpha) \\ +(1-w) \times \text{power_mean}(M_{\text{ord}}(M,\tau),\alpha),\end{aligned}$$

where $w \in [0,1]$ is the weight coefficient for $M_{\text{high}}(M,\tau)$ (fixed at $w = 0.7$); the power mean function is a generalized mean that adjusts the sensitivity to different values by varying α, calculated as:

$$\text{power_mean}(M,\alpha) = \left(\frac{1}{|M|}\sum_{m\in M} m^{\alpha}\right)^{1/\alpha}.$$

When $\alpha > 1$, larger values receive more weight. We set $\alpha = 1.5$ in all experiments.

3.2 FCD-Based Continual Pre-training Strategies

We define three CPT strategies based on the FCD score:

• **FCD-Hard:** This strategy performs **hard sampling** (Xie et al., 2024) to directly select a subset of documents with the highest FCD scores (as defined in Equation 5).

• **FCD-Soft:** Following the CDF-based balanced sampling approach (Ying et al., 2025), this strategy adopts **soft sampling** by assigning higher sampling probabilities to documents with higher FCD scores while retaining documents from lower score ranges.

• **FCD-CL:** This strategy implements **curriculum learning** by sequencing the training data in ascending order of FCD difficulty (Naïr et al., 2024). The model progresses from low-difficulty (easy) to medium and finally high-difficulty (hard) samples.

The experiments compare their domain and general-task accuracy.

4 Understanding FCD Scores

4.1 Component Score Distributions

Before evaluating data selection, we examine how the two word-level components behave over corpus statistics. For frequency difficulty d_f (Equation 1), we group words by log frequency and compute their weighted cumulative distribution. For conceptual difficulty d_c (Equation 3), we group nouns by their minimum semantic distance (MSD; Equation 2) to the basic vocabulary and weight each

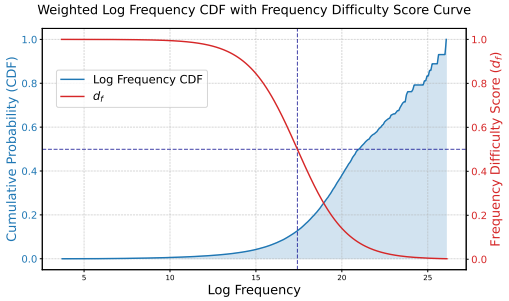


Figure 1: Weighted CDF of $\log f(w)$ and its transformation into d_f.

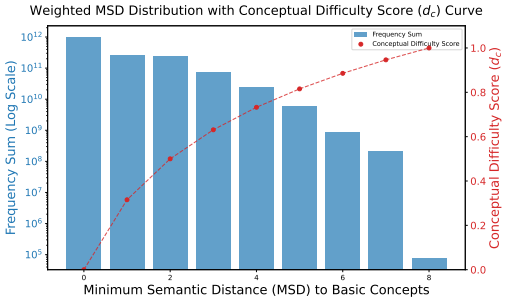


Figure 2: Weighted distribution of MSD and its transformation into d_c.

group by corpus frequency. Figure 1 and Figure 2 show these empirical distributions together with the corresponding score transformations.

Most token mass lies in the high-frequency region where d_f is near zero; the transition around c assigns larger scores to less frequent words while keeping scores for common words small. The MSD distribution is likewise concentrated near the basic vocabulary. Logarithmic normalization gives greater resolution to small distances and limits the influence of large ones. These plots characterize the two score mappings; their utility for data selection is evaluated through the downstream experiments in Table 1.

4.2 FCD Distribution Across Domains

We compare the FCD distributions of five corpora: **Medical** (MMedC, 2024), **General Web** (FineWeb, 2024), **Finance** (YiZhao-FinDataSet, 2024), **STEM** (arXiv, 2024), and **Philosophy** (SEP, 2024). Figure 3 illustrates the probability density functions derived from the binned FCD scores for each domain. Their mean scores follow the ordering Philosophy $<$ STEM $<$ Finance $<$ General Web $<$ Medical.

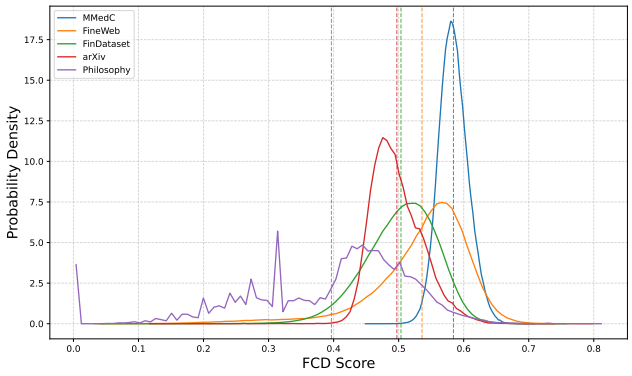


Figure 3: Probability density functions of FCD scores across five domains. The vertical dashed lines represent the mean FCD scores.

STEM and **Philosophy** have lower FCD distributions than **General Web** and **Medical**, and the mean for STEM is close to that of **Finance**. This ordering should not be read as a ranking of overall domain difficulty. FCD measures lexical rarity and WordNet distance, but does not represent syntactic complexity or logical reasoning.

The higher scores for **General Web** and **Medical** are consistent with a greater presence of long-tail vocabulary in these corpora. In the medical data, examples include rare diseases and chemical compounds. These distributions clarify that FCD is primarily a measure of terminological properties; whether a score range is useful for CPT must be determined by downstream evaluation.

5 EXPERIMENTS

Table 1: CPT experiments on Qwen2.5-0.5B comparing established baselines with FCD variants. Values represent Accuracy (%), with **best** and second best results highlighted (excluding Base Model/CPT). Superscripts indicate the gap relative to Base CPT. Averages are computed from unrounded per-task scores; displayed entries are rounded to two decimals. While all competitive baselines underperform the full-corpus Base CPT, our **FCD-Hard** improves the domain average by **1.36 percentage points**.

		Domain Tasks				General Tasks			
Category	Method	MedMCQA	MedQA	MMedBench	**Domain Avg**	MMLU	ARC-E	ARC-C	**Avg**
Original	Base Model	38.06	34.17	34.88	35.70	47.60	67.80	33.96	42.75
	Base CPT	38.35	35.90	36.37	36.87	44.71	66.20	30.63	42.03
Baselines	DSIR	$37.10^{\downarrow 1.24}$	$35.74^{\downarrow 0.16}$	$35.35^{\downarrow 1.02}$	$36.06^{\downarrow 0.81}$	$44.78^{\uparrow 0.07}$	$66.46^{\uparrow 0.25}$	$30.29^{\downarrow 0.34}$	$41.62^{\downarrow 0.41}$
	ETA-DACP	$36.48^{\downarrow 1.86}$	$35.43^{\downarrow 0.47}$	$36.14^{\downarrow 0.24}$	$36.01^{\downarrow 0.86}$	$44.96^{\uparrow 0.25}$	$66.29^{\uparrow 0.08}$	$31.14^{\uparrow 0.51}$	$41.74^{\downarrow 0.29}$
	Fineweb-Edu	$37.56^{\downarrow 0.79}$	$35.43^{\downarrow 0.47}$	$36.21^{\downarrow 0.16}$	$36.40^{\downarrow 0.47}$	$44.54^{\downarrow 0.16}$	$66.37^{\uparrow 0.17}$	$\mathbf{31.91}^{\uparrow \mathbf{1.28}}$	$42.00^{\downarrow 0.02}$
	PPL	$36.00^{\downarrow 2.34}$	$35.98^{\uparrow 0.08}$	$35.66^{\downarrow 0.71}$	$35.88^{\downarrow 0.99}$	$44.94^{\uparrow 0.23}$	$65.95^{\downarrow 0.25}$	$31.14^{\uparrow 0.51}$	$41.61^{\downarrow 0.41}$
Ours	FCD-CL	$38.20^{\downarrow 0.14}$	$36.21^{\uparrow 0.31}$	$35.82^{\downarrow 0.55}$	$36.75^{\downarrow 0.13}$	$44.67^{\downarrow 0.04}$	$66.41^{\uparrow 0.21}$	$31.31^{\uparrow 0.68}$	$42.11^{\uparrow 0.08}$
	FCD-Soft	$37.75^{\downarrow 0.60}$	$36.76^{\uparrow 0.86}$	$36.84^{\uparrow 0.47}$	$37.12^{\uparrow 0.25}$	$\mathbf{45.09}^{\uparrow \mathbf{0.38}}$	$66.50^{\uparrow 0.29}$	$31.06^{\uparrow 0.43}$	$42.33^{\uparrow 0.31}$
	FCD-Hard	$\mathbf{39.76}^{\uparrow \mathbf{1.41}}$	$\mathbf{37.94}^{\uparrow \mathbf{2.04}}$	$\mathbf{37.00}^{\uparrow \mathbf{0.63}}$	$\mathbf{38.23}^{\uparrow \mathbf{1.36}}$	$44.73^{\uparrow 0.02}$	$\mathbf{66.71}^{\uparrow \mathbf{0.51}}$	$31.31^{\uparrow 0.68}$	$\mathbf{42.91}^{\uparrow \mathbf{0.88}}$

5.1 EXPERIMENTAL SETUP

Base Model, Training Data, and Training Budget The base model for continual pre-training (CPT) is **Qwen2.5-0.5B** (Yang et al., 2024). The training corpus is the English subset of the medical dataset **MMedC** (Qiu et al., 2024), totaling 6B tokens. All experiments use a CPT budget of 6B tokens under one of two protocols: **Full CPT** (6B tokens $\times$ 1 epoch) and **Sampled CPT** (a 1/3 subset, $\approx$ 2B tokens, $\times$ 3 epochs). Hyperparameters and dataset details are in Appendix D.1 and Appendix D.2, respectively.

Evaluation Benchmarks Model performance is evaluated across **Domain QA** and **General QA**. Domain QA is assessed using MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2020) and MMed-Bench (Qiu et al., 2024). General QA is tested on MMLU (Hendrycks et al., 2021) and ARC (Clark et al., 2018). We use **5-shot** multiple-choice evaluation based on answer probabilities in the *lm-evaluation-harness* (Liang et al., 2023) and report **Accuracy (ACC)**. We use MedText (BI55, 2024) as the validation set and track **Perplexity (PPL)** during training. Full protocol details are in Appendix D.3.

Baselines and FCD Variants We compare the FCD strategies with a non-CPT **Base Model**, an unfiltered **Base CPT** model trained under the Full CPT Protocol, and four data-selection methods all implemented using the Sampled CPT Protocol: **DSIR** (Xie et al., 2023) for n-gram similarity, **ETA-DACP** (Xie et al., 2024) for lexical diversity-based domain selection, **Fineweb-Edu** (Penedo et al., 2024) for classifier-based educational-quality filtering, and **PPL** (Ankner et al., 2025) for perplexity-based selection.

Our proposed approaches include **FCD-Hard** and **FCD-Soft**, both following the Sampled CPT Protocol. We evaluate the sequence-dependent **FCD-CL** strategy under the Full CPT Protocol. Implementation details for all baselines and variants are provided in Appendix D.4.

5.2 COMPARISON OF CPT STRATEGIES

We compare three FCD-based strategies — **FCD-Hard**, **FCD-Soft**, and **FCD-CL** — with four data-selection baselines. The comparison tests whether frequency and conceptual distance provide a useful selection signal for domain adaptation.

As shown in Table 1, DSIR, ETA-DACP, Fineweb-Edu, and PPL all score below the unfiltered Base CPT on the domain average. **FCD-Hard** improves over Base CPT on MedMCQA, MedQA, and MMedBench by 1.41, 2.04, and 0.63 points, respectively. Its domain-average gain is **1.36 percentage points** under the same total training-token budget.

The appendices report results on Llama-3.2-1B (Appendix E), performance by test-item difficulty (Appendix F), a comparison across sampling ratios (Appendix H), and a discussion of the baseline results (Appendix I).

5.3 Validation Perplexity During CPT

To compare training dynamics across CPT strategies, we track perplexity (PPL) on the MedText validation set (BI55, 2024). As shown in Figure 4, **FCD-Hard** decreases fastest and reaches the lowest final PPL among the plotted runs. It is trained for three epochs on a fixed subset of approximately 2B unique tokens, whereas Base CPT uses the full 6B-token corpus for one epoch. Both process approximately 6B tokens in total. The curve shows better validation-set fit under this protocol; downstream accuracy is reported separately in Table 1.

6 Analysis and Discussion

6.1 Ablation Study of FCD Components

To evaluate the contributions of frequency (d_f, Equation 1) and conceptual (d_c, Equation 3) dimensions, we compare **FCD-Hard** against three variants under the same **Sampled CPT** protocol: (1) **FCD**$_{w/o\ d_c}$ (relying solely on d_f); (2) **FCD**$_{w/o\ d_f}$ (using only d_c); (3) **FCD-Easy** (selecting low-difficulty samples).

Table 2: Ablation study on medical benchmarks. Superscripts indicate absolute gaps relative to FCD-Hard.

Method	MedMCQA	MedQA	MMedBench	Avg
FCD-Hard	39.76	37.94	37.00	38.23
FCD$_{w/o\ d_c}$	39.11$^{\downarrow 0.65}$	37.23$^{\downarrow 0.71}$	36.61$^{\downarrow 0.39}$	37.65$^{\downarrow 0.58}$
FCD$_{w/o\ d_f}$	38.37$^{\downarrow 1.39}$	37.00$^{\downarrow 0.94}$	36.84$^{\downarrow 0.16}$	37.40$^{\downarrow 0.83}$
FCD-Easy	37.96$^{\downarrow 1.80}$	35.43$^{\downarrow 2.51}$	35.04$^{\downarrow 1.96}$	36.14$^{\downarrow 2.09}$

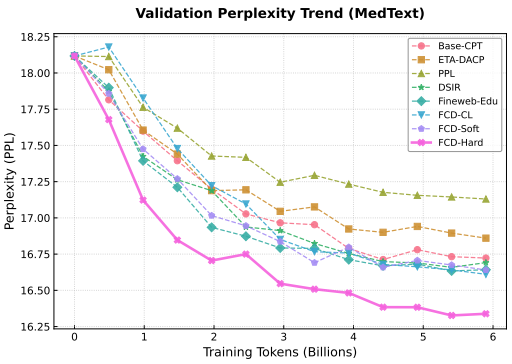


Figure 4: PPL convergence on MedText. FCD-Hard uses three epochs over one-third of the unique data; each protocol processes approximately 6B tokens.

Results in Table 2 show that removing d_c or d_f reduces the average accuracy across the three medical benchmarks by 0.58 and 0.83 points, respectively. Thus, both components contribute to the combined score in this setting. **FCD-Easy** is 2.09 points below FCD-Hard on average, consistent with the main medical-domain result that favors high-FCD documents. The aggregation comparison in Appendix J further shows that the peak-oriented aggregation used by FCD-Hard performs better than simple averaging in this experiment.

6.2 Relationship Between FCD and Perplexity

To measure the overlap between FCD and a model-based signal, we analyze its correlation with **Perplexity (PPL)** using a Qwen2.5-0.5B model. As shown in Figure 5, the Pearson correlation coefficient is $r = 0.0512$. Thus, the two scores have little linear association on the medical corpus under this model.

The metrics are computed from different information: PPL measures how unexpected a sequence is to a particular model, whereas FCD uses corpus frequency and WordNet distance without model inference. The low correlation therefore indicates that FCD is not a proxy for PPL; it does not by

itself establish which signal better measures data quality. Their downstream comparison is reported in Table 1.

Figure 5, Table 3, and Table 4 summarize the correlation pattern, financial-domain comparison, and parameter stability analyzed below.

Table 3: CPT accuracy on the translated FinEval benchmark.

Strategy	**Accuracy (%)**
Base Model	36.17
Random Sample	<u>41.84</u>
ETA-DACP	40.43
FCD-Hard	38.30
FCD-Easy	**43.26**

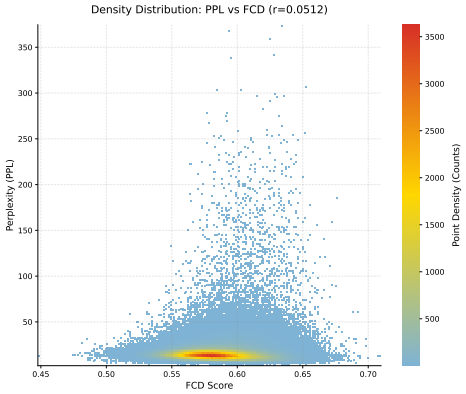


Figure 5: Density heatmap of PPL vs. FCD scores on the medical corpus.

Table 4: Parameter stability. ρ denotes Spearman correlation.

Default	**Setting**	**Overlap @ Top**			**Corr.**
		10%	20%	30%	ρ
$\tau = 0.9$	$\tau \leftarrow 0.8$	0.87	0.90	0.91	0.9798
	$\tau \leftarrow 0.85$	0.93	0.94	0.95	0.9914
$w = 0.7$	$w \leftarrow 0.6$	0.95	0.96	0.97	0.9979
	$w \leftarrow 0.8$	0.95	0.96	0.97	0.9978
$k = 0.7$	$k \leftarrow 0.6$	0.98	0.98	0.98	0.9993
	$k \leftarrow 0.8$	0.98	0.98	0.99	0.9994
$\beta_n = 0.7$	$\beta_n \leftarrow 0.5$	0.87	0.90	0.92	0.9783
	$\beta_n \leftarrow 0.6$	0.93	0.95	0.96	0.9938
$\alpha = 1.5$	$\alpha \leftarrow 1.0$	0.98	0.99	0.99	0.9997
	$\alpha \leftarrow 2.0$	0.99	0.99	0.99	0.9998
$w_f = 0.5$	$w_f \leftarrow 0.4$	0.94	0.95	0.96	0.9950
	$w_f \leftarrow 0.6$	0.95	0.96	0.96	0.9959

6.3 Results on the Financial Corpus

To examine FCD outside the medical domain, we apply it to a financial corpus. We use the YiZhao-FinDataSet (CMB-AILab, 2024), a large-scale English financial corpus (20B tokens). Each method selects a 30% subset (approximately 6B tokens), which is used to train **Qwen2.5-0.5B** for one epoch. We compare **Random Sample**, **ETA-DACP**, **FCD-Hard**, and **FCD-Easy** under this fixed 6B-token exposure. Evaluation uses **FinEval** (Guo et al., 2025a), translated into English by `Qwen/Qwen2.5-32B-Instruct` with a direct translation instruction.

Our corpus analysis shows a lower FCD distribution for finance than for medicine (subsection 4.2). Based on this observation, we evaluate FCD-Easy as a sampling strategy for the financial corpus, alongside FCD-Hard and the baselines.

As shown in Table 3, **FCD-Easy** achieves the highest accuracy in this experiment, exceeding FCD-Hard and **ETA-DACP** (Xie et al., 2024).

This ordering differs from the medical results and is consistent with the direction suggested by the preceding corpus analysis. To inspect the selected documents, we include a qualitative case study in Appendix K. In those examples, low-FCD documents use common words with financial meanings, such as "grants," "loan," and "repayment," whereas several high-FCD documents contain terminology from biology or communication engineering.

Together, the corpus-level distribution, FinEval result, and case examples show how FCD analysis can inform the choice of candidate score ranges for data selection across domains.

6.4 Parameter Sensitivity Analysis

We analyze the sensitivity of FCD rankings to the quantile threshold τ, high-difficulty weight w, frequency scaling factor k, noun weight β_{noun} (abbreviated as β_n), power mean order α, and frequency weight w_f. The latter controls the relative contribution of d_f and d_c, with the word-level score computed as $d_f(w)^{w_f} \cdot d_c(w)^{1-w_f}$. We fix all parameters to their defaults and perturb one at a time.

For the local perturbations around the defaults shown in Table 4, the top-30% overlap is at least **0.91** and exceeds **0.95** for most settings; all rank correlations exceed **0.97**. The rankings are therefore stable within these local ranges. Results over wider parameter ranges are reported in Appendix C.

7 Conclusion

We introduced Frequency-Conceptual Difficulty (FCD), which combines corpus frequency and WordNet distance into a document-level lexical difficulty score. In the medical experiments, FCD-Hard trains for three epochs on a fixed one-third subset and improves domain-average accuracy over full-corpus Base CPT by 1.36 percentage points under the same total training-token budget; all four compared data-selection baselines score below Base CPT.

The preferred score range differs by corpus: high-FCD selection performs best on the medical data, whereas FCD-Easy performs best on the financial data evaluated with translated FinEval. FCD should therefore be treated as a model-free lexical signal whose sampling direction must be validated for the target domain, rather than as a universal ranking of data quality.

Limitations

Coverage and Uncertainty: We evaluate Qwen2.5-0.5B and Llama-3.2-1B on medical and financial corpora; these experiments do not establish behavior at larger scales or in other domains. Most configurations were run once, and only Base CPT was repeated three times. The differences are therefore point estimates; repeated runs and paired item-level analysis are needed for stronger uncertainty estimates.

Evaluation Scope: MMedBench incorporates questions from MedQA, so these two medical benchmarks are not independent sources of evidence.

Linguistic and Structural Scope: FCD relies on WordNet's noun hierarchy, with incomplete coverage of non-nouns, named entities, and domain-specific senses. FCD measures lexical rarity and WordNet distance, not complete domain knowledge, syntax, formal structure, symbolic reasoning, or algorithmic complexity.

Ethics Statement

The models and datasets used in our experiments, including Qwen2.5-0.5B, Google Books Ngram, MMedC, YiZhao-FinDataSet, and the evaluation benchmarks, are publicly available and were used under their respective licenses. We plan to release the code and trained models under an open-source license. The study did not involve manual data annotation; data selection and difficulty scores were computed automatically.

Reproducibility Statement

The models and datasets used in this study (Qwen2.5-0.5B, Llama-3.2-1B, MMedC, and YiZhao-FinDataSet) are publicly available. Appendix D.1 reports the CPT hyperparameters, Appendix D.3 describes the evaluation protocol, and Appendix D.4 documents the baseline implementations. FCD uses WordNet, Google Books Ngram, and Oxford CEFR word lists. We will release the source code upon acceptance.

USE OF AI ASSISTANCE

We used AI-based tools (e.g., Gemini) for literature discovery, brainstorming, code implementation, and language editing. The authors verified the cited sources, code, analyses, and manuscript, and take responsibility for the work.

REFERENCES

AiresPucrs. Stanford encyclopedia of philosophy dataset, 2024. URL https://hf-mirror.com/datasets/AiresPucrs/stanford-encyclopedia-philosophy. Dataset based on the Stanford Encyclopedia of Philosophy, hosted on Hugging Face.

Muhammet Hatipoglu Alican Acar, Alara Dirik. Arxiver. https://huggingface.co/datasets/neuralwork/arxiver, 2024.

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max M Marion, Matthew Leavitt, and Mansheej Paul. Perplexed by perplexity: Perplexity-based data pruning with small reference models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Representation Learning*, volume 2025, pp. 100615–100634, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fa30825df7ceaba452d5533538ea29c2-Paper-Conference.pdf.

BI55. Medtext, 2024. URL https://huggingface.co/datasets/BI55/MedText.

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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. Instruction mining: Instruction data selection for tuning large language models, 2024. URL https://arxiv.org/abs/2307.06290.

Tyler A Chang and Benjamin K Bergen. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022. URL https://aclanthology.org/2022.tacl-1.1/.

Jie Chen, Zhipeng Chen, Jiapeng Wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Wayne Xin Zhao, Zhicheng Dou, Jiaxin Mao, Yankai Lin, Ruihua Song, Jun Xu, Xu Chen, Rui Yan, Zhewei Wei, Di Hu, Wenbing Huang, and Ji-Rong Wen. Towards effective and efficient continual pre-training of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5779–5795, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.289. URL https://aclanthology.org/2025.acl-long.289/.

Wenhu Chen, Yu Su, Yilin Shen, Zhiyu Chen, Xifeng Yan, and William Yang Wang. How large a vocabulary does text classification need? A variational approach to vocabulary selection. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 3487–3497. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1352. URL https://doi.org/10.18653/v1/n19-1352.

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, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018. URL http://arxiv.org/abs/1803.05457.

CMB-AILab. Yizhao-findataset, 2024. URL https://www.modelscope.cn/datasets/CMB_AILab/YiZhao-FinDataSet.

Yiming Cui and Xin Yao. Rethinking LLM language adaptation: A case study on chinese mixtral. *CoRR*, abs/2403.01851, 2024. doi: 10.48550/ARXIV.2403.01851. URL https://doi.org/10.48550/arXiv.2403.01851.

Mihai Dascalu, Danielle S. McNamara, Scott A. Crossley, and Stefan Trausan-Matu. Age of exposure: A model of word learning. In Dale Schuurmans and Michael P. Wellman (eds.), *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 2928–2934. AAAI Press, 2016. doi: 10.1609/AAAI.V30I1.10372. URL https://doi.org/10.1609/aaai.v30i1.10372.

Abhinandan Tejalkumar Desai, Kai North, Marcos Zampieri, and Christopher Homan. LCP-RIT at semeval-2021 task 1: Exploring linguistic features for lexical complexity prediction. In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurélie Herbelot, and Xiaodan Zhu (eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, pp. 548–553. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.SEMEVAL-1.67. URL https://doi.org/10.18653/v1/2021.semeval-1.67.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. Mods: Model-oriented data selection for instruction tuning. *CoRR*, abs/2311.15653, 2023. doi: 10.48550/ARXIV.2311.15653. URL https://doi.org/10.48550/arXiv.2311.15653.

Neus Figueras, Felianka Kaftandjieva, and Sauli Takala. Relating a reading comprehension test to the cefr levels: A case of standard setting in practice with focus on judges and items. *Canadian modern language review*, 69(4):359–385, 2013. URL https://utppublishing.com/doi/abs/10.3138/cmlr.1723.359.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. FinEval: A Chinese financial domain knowledge evaluation benchmark for large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6258–6292, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.318. URL https://aclanthology.org/2025.naacl-long.318/.

Yiduo Guo, Jie Fu, Huishuai Zhang, and Dongyan Zhao. Efficient domain continual pretraining by mitigating the stability gap. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pp. 32850–32870, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1578. URL https://aclanthology.org/2025.acl-long.1578/.

Hung Tan Ha, Duyen Thi Bich Nguyen, and Tim Stoeckel. What is the best predictor of word difficulty? a case of data mining using random forest. *Language Testing*, 41(4):828–844, 2024. URL https://journals.sagepub.com/doi/full/10.1177/02655322241263628.

Brett James Hashimoto. Is frequency enough?: The frequency model in vocabulary size testing. *Language Assessment Quarterly*, 18(2):171–187, 2021. URL https://www.tandfonline.com/doi/abs/10.1080/15434303.2020.1860058.

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. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *CoRR*, abs/2412.02980, 2024. doi: 10.48550/ARXIV.2412.02980. URL https://doi.org/10.48550/arXiv.2412.02980.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. Data-centric artificial intelligence. *Bus. Inf. Syst. Eng.*, 66(4):507–515, 2024. doi: 10.1007/S12599-024-00857-8. URL https://doi.org/10.1007/s12599-024-00857-8.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081, 2020. URL https://arxiv.org/abs/2009.13081.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. Lifelong pretraining: Continually adapting language models to emerging corpora. In Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé (eds.), *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 1–16, virtual+Dublin, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.1. URL https://aclanthology.org/2022.bigscience-1.1/.

Satoru Katsumata, Yukio Matsumura, Hayahide Yamagishi, and Mamoru Komachi. Graph-based filtering of out-of-vocabulary words for encoder-decoder models. In Vered Shwartz, Jeniya Tabassum, Rob Voigt, Wanxiang Che, Marie-Catherine de Marneffe, and Malvina Nissim (eds.), *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, Student Research Workshop*, pp. 112–119. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-3016. URL https://aclanthology.org/P18-3016/.

Xiaomin Li, Mingye Gao, Zhiwei Zhang, Chang Yue, and Hong Hu. Rule-based rating and selection of llm training data. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025. URL https://openreview.net/forum?id=AKoLN21VSu.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 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, 2024. URL https://openreview.net/forum?id=BTKAeLqLMw.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale. *CoRR*, abs/2309.04564, 2023. doi: 10.48550/ARXIV.2309.04564. URL https://doi.org/10.48550/arXiv.2309.04564.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011. URL https://www.science.org/doi/10.1126/science.1199644.

George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

Marwa Naïr, Kamel Yamani, Lynda Lhadj, and Riyadh Baghdadi. Curriculum learning for small code language models. In Xiyan Fu and Eve Fleisig (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 390–401, Bangkok, Thailand, August 2024. Association for Computational Linguistics. ISBN 979-8-89176-097-4. doi: 10.18653/v1/2024.acl-srw.44. URL https://aclanthology.org/2024.acl-srw.44/.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849. Curran Associates, Inc., 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/370df50ccfdf8bde18f8f9c2d9151bda-Paper-Datasets_and_Benchmarks_Track.pdf.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*, 2025. URL https://arxiv.org/abs/2506.20920.

Ru Peng, Kexin Yang, Yawen Zeng, Junyang Lin, Dayiheng Liu, and Junbo Zhao. Dataman: Data manager for pre-training large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=eNbA8Fqir4.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine. *CoRR*, abs/2402.13963, 2024. doi: 10.48550/ARXIV.2402.13963. URL https://doi.org/10.48550/arXiv.2402.13963.

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975. URL https://psycnet.apa.org/record/1976-00172-001.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(76)90013-X. URL https://www.sciencedirect.com/science/article/pii/001002857690013X.

Jeffrey Stewart, Joseph P Vitta, Christopher Nicklin, Stuart McLean, Geoffrey G Pinchbeck, and Brandon Kramer. The relationship between word difficulty and frequency: A response to hashimoto (2021). *Language Assessment Quarterly*, 19(1):90–101, 2022. URL https://www.tandfonline.com/doi/abs/10.1080/15434303.2021.1992629.

Zhenpeng Su, Zijia Lin, Baixue Baixue, Hui Chen, Songlin Hu, Wei Zhou, Guiguang Ding, and Xing W. MiLe loss: a new loss for mitigating the bias of learning difficulties in generative language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 250–262, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.18. URL https://aclanthology.org/2024.findings-naacl.18/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. Probase: a probabilistic taxonomy for text understanding. In K. Selçuk Candan, Yi Chen, Richard T. Snodgrass, Luis Gravano, and Ariel Fuxman (eds.), *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pp. 481–492. ACM, 2012. doi: 10.1145/2213836.2213891. URL https://doi.org/10.1145/2213836.2213891.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 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, 2024. URL https://openreview.net/forum?id=PG5fV50maR.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6b9aa8f418bde2840d5f4ab7a02f663b-Abstract-Conference.html.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. Efficient continual pre-training for building domain specific large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 10184–10201. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.606. URL https://doi.org/10.18653/v1/2024.findings-acl.606.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Junjie Oscar Yin and Alexander M. Rush. Compute-constrained data selection. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=4es2oO9tw1.

Yizhou Ying, Geng Zhang, Cui Danxin, Chengyu Du, Guanglei Yue, Sihang Jiang, Jiaqing Liang, Yifei Fu, Hailin Hu, and Yanghua Xiao. Data-efficient selection via grammatical complexity in continual pre-training of domain-specific LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22066–22080, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1121. URL https://aclanthology.org/2025.emnlp-main.1121/.

Wenhao Yu, Chenguang Zhu, Yuwei Fang, Donghan Yu, Shuohang Wang, Yichong Xu, Michael Zeng, and Meng Jiang. Dict-bert: Enhancing language model pre-training with dictionary. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 1907–1918. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.150. URL https://doi.org/10.18653/v1/2022.findings-acl.150.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Ben Hu. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.*, 57(5): 129:1–129:42, 2025. doi: 10.1145/3711118. URL https://doi.org/10.1145/3711118.

Haomin Zhang, Yuting Han, Xing Zhang, and Liuran Cui. Frequency, dispersion and abstractness in the lexical sophistication analysis of a learner-based word bank: Dimensionality reduction and identification. *Journal of Quantitative Linguistics*, 29(2):195–211, 2022a. URL https://www.tandfonline.com/doi/abs/10.1080/09296174.2020.1782716.

Tong Zhang, Wei Ye, Baosong Yang, Long Zhang, Xingzhang Ren, Dayiheng Liu, Jinan Sun, Shikun Zhang, Haibo Zhang, and Wen Zhao. Frequency-aware contrastive learning for neural machine translation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 11712–11720. AAAI Press, 2022b. doi: 10.1609/AAAI.V36I10.21426. URL https://doi.org/10.1609/aaai.v36i10.21426.

Weilin Zhao, Tengyu Pan, Xu Han, Yudi Zhang, Sun Ao, Yuxiang Huang, Kaihuo Zhang, Weilun Zhao, Yuxuan Li, Jie Zhou, Hao Zhou, Jianyong Wang, Maosong Sun, and Zhiyuan Liu. Fr-spec: Accelerating large-vocabulary language models via frequency-ranked speculative sampling. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 3909–3921. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.198/.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. Meta-rater: A multi-dimensional data selection method for pre-training language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10856–10896, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.533. URL https://aclanthology.org/2025.acl-long.533/.

George Kingsley Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Cambridge, Mass., 1949. Illustrated. $6.50; 110(2868), 669–669.

A Additional Related Work

A.1 Supplementary Studies on Data Selection

In addition to the methods discussed in the main text, diversity can be estimated through "self-similarity" and "coverage" within datasets (Havrilla et al., 2024). Techniques using document embeddings have been widely used for sampling across the data distribution (Xie et al., 2024; Liu et al., 2024; Du et al., 2023). Selection methods range from heuristic filters to model-based evaluators, with different computational costs and selection criteria.

A.2 Alternative Perspectives on Difficulty

Beyond lexical frequency, researchers have explored various dimensions of linguistic difficulty. For instance, Dascalu et al. (2016) explored human-centered readability scores, while others investigated syntactic complexity as a measure of professional text difficulty (Katsumata et al., 2018). Studies of long-tail distributions further distinguish lexical rarity from structural or conceptual complexity.

B Precomputing Semantic Distance in WordNet

After constructing the WordNet noun semantic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and defining the basic vocabulary node set $\mathcal{V}_{\text{basic}}$ (as detailed in §3.1.2), to preprocess the minimum semantic distance for all synset nodes $v \in \mathcal{V}$ (i.e., the shortest path from each node to the nearest node in $\mathcal{V}_{\text{basic}}$), we adopt multi-source Breadth-First-Search (BFS) starting from $\mathcal{V}_{\text{basic}}$. This algorithm computes the distance with a time complexity of $O(|\mathcal{V}| + |\mathcal{E}|)$.

Algorithm 1 Multi-source BFS for calculating MSD to Basic Concepts

Require: Semantic graph (WordNet) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, basic vocabulary node set $\mathcal{V}_{\text{basic}} \subseteq \mathcal{V}$
Ensure: MSD : $\mathcal{V} \to \mathbb{N}$ (mapping from synset nodes to MSD)
Initialize $\text{MSD}(v) = \infty$ for all $v \in \mathcal{V}$
Initialize a queue Q and enqueue all $v_b \in \mathcal{V}_{\text{basic}}$
Initialize a set $S = \emptyset$ to track processed nodes (ensuring each node updates neighbors only once)
For each $v_b \in \mathcal{V}_{\text{basic}}$, set $\text{MSD}(v_b) = 0$
while Q is not empty **do**
    $u \leftarrow \text{Dequeue}(Q)$
    if $u \in S$ **then**
        continue
    end if
    $S \leftarrow S \cup \{u\}$
    for each neighbor v of u in $\mathcal{G}$ (i.e., $(u, v) \in \mathcal{E}$ or $(v, u) \in \mathcal{E}$) **do**
        if $\text{MSD}(v) > \text{MSD}(u) + 1$ **then**
            $\text{MSD}(v) \leftarrow \text{MSD}(u) + 1$
            $\text{Enqueue}(Q, v)$
        end if
    end for
end while
return msd

Note: The algorithm traverses the undirected version of $\mathcal{G}$ (considering both hypernym and hyponym edges) to ensure all nodes are reachable from $\mathcal{V}_{\text{basic}}$. For WordNet 3.0's noun network, this guarantees complete coverage of all 82,115 nodes with a maximum MSD of 8.

C Detailed Results of Parameter Sensitivity Analysis

This appendix extends the local perturbations in subsection 6.4 to wider parameter ranges. The full results for τ, w, k, β_n, α, and w_f are presented in Table 5. Larger ranking changes at settings far from the defaults are expected; these broader results delimit the local stability claim made in the main text.

Table 5 reports the overlap rates and rank correlations for each configuration.

D Experimental Setup

D.1 Continual Pre-training Configurations

We implement the CPT framework using PyTorch and DeepSpeed, and the models are optimized with the standard cross-entropy loss on a server equipped with 4 GPUs. The detailed hyperparameter configurations used in our experiments are summarized in Table 6.

Table 5: Stability analysis. Param. (Default) shows the symbol and its default value. ρ is the Spearman correlation.

Default	Setting	Overlap @ Top			Corr.
		10%	20%	30%	ρ
$\tau = 0.9$	$\tau \leftarrow 0.5$	0.77	0.82	0.85	0.9374
	$\tau \leftarrow 0.7$	0.82	0.86	0.88	0.9638
	$\tau \leftarrow 0.8$	0.87	0.90	0.91	0.9798
	$\tau \leftarrow 0.85$	0.93	0.94	0.95	0.9914
	$\tau \leftarrow 0.95$	0.85	0.89	0.91	0.9774
$w = 0.7$	$w \leftarrow 0.5$	0.91	0.93	0.94	0.9916
	$w \leftarrow 0.6$	0.95	0.96	0.97	0.9979
	$w \leftarrow 0.8$	0.95	0.96	0.97	0.9978
	$w \leftarrow 0.9$	0.89	0.93	0.94	0.9911
$k = 0.7$	$k \leftarrow 0.6$	0.98	0.98	0.98	0.9993
	$k \leftarrow 0.8$	0.98	0.98	0.99	0.9994
	$k \leftarrow 1.0$	0.94	0.96	0.96	0.9960
$\beta_{\text{n}} = 0.7$	$\beta_{\text{n}} \leftarrow 0.0$	0.71	0.76	0.79	0.8633
	$\beta_{\text{n}} \leftarrow 0.4$	0.83	0.86	0.88	0.9574
	$\beta_{\text{n}} \leftarrow 0.5$	0.87	0.90	0.92	0.9783
	$\beta_{\text{n}} \leftarrow 0.6$	0.93	0.95	0.96	0.9938
	$\beta_{\text{n}} \leftarrow 0.8$	0.92	0.94	0.95	0.9920
	$\beta_{\text{n}} \leftarrow 0.9$	0.82	0.86	0.88	0.9646
	$\beta_{\text{n}} \leftarrow 1.0$	0.71	0.78	0.81	0.9135
$\alpha = 1.5$	$\alpha \leftarrow 1.0$	0.98	0.99	0.99	0.9997
	$\alpha \leftarrow 2.0$	0.99	0.99	0.99	0.9998
	$\alpha \leftarrow 3.0$	0.97	0.98	0.98	0.9986
	$\alpha \leftarrow 5.0$	0.94	0.95	0.96	0.9951
$w_f = 0.5$	$w_f \leftarrow 0.0$	0.75	0.77	0.79	0.8617
	$w_f \leftarrow 0.3$	0.88	0.90	0.92	0.9790
	$w_f \leftarrow 0.4$	0.94	0.95	0.96	0.9950
	$w_f \leftarrow 0.6$	0.95	0.96	0.96	0.9959
	$w_f \leftarrow 0.7$	0.90	0.92	0.93	0.9854
	$w_f \leftarrow 1.0$	0.78	0.82	0.85	0.9370

Consistent with the main experiments, every comparison uses a total training exposure of 6B tokens. Under the **Full CPT** protocol, the model is trained for one epoch on the full 6B-token corpus. Under the **Sampled CPT** protocol, the model is trained for three epochs on a fixed one-third subset containing approximately 2B unique tokens. Thus, the protocols differ in the amount of unique training data but not in total token exposure.

Each CPT configuration is trained once because of computational constraints. As a reference for run-to-run variability, we trained Base CPT three times; the maximum-minus-minimum accuracy was at most 0.4 points on each domain task. Because the selection methods were not repeated, this range is descriptive and does not provide method-level uncertainty estimates.

Table 6: Hyperparameter settings for CPT.

Hyperparameter	Value
Optimizer	Adam
Learning Rate (LR)	2e-5
LR Scheduler	WarmupCosineLR
Warmup Ratio	0.1
Min LR Ratio	0.1
Sequence Length	4096
Global Batch Size	$\sim$3M tokens

D.2 Training Dataset

MMedC (Qiu et al., 2024) is a multilingual medical corpus containing 25.5B tokens in English, Chinese, Japanese, French, Russian, and Spanish. It combines medical content filtered from general corpora, OCR-processed textbooks, medical websites, and existing medical corpora. We use the English subset, which accounts for 42% of MMedC (approximately 10.7B tokens) and covers 21 medical fields. The CPT corpus contains 6B tokens from this subset.

D.3 Evaluation Benchmarks

All benchmark results use the 5-shot multiple-choice protocol in *lm-evaluation-harness*, with answer options scored by model probability.

D.3.1 MedMCQA

MedMCQA (Pal et al., 2022) is a large-scale multi-subject MCQA dataset derived from Indian postgraduate medical entrance exams (AIIMS PG and NEET PG), containing 193,155 MCQs covering 21 medical subjects and 2,400+ healthcare topics. Each sample includes a question, four options, a correct answer, and a detailed human-annotated explanation. The dataset adopts an exam-based split to ensure generalization: the training set (182,822 questions) consists of mock tests, the development set (4,183 questions) is from NEET PG exams (2001–present), and the test set (6,150 questions) is from AIIMS PG exams (1991–present), with duplicate questions (Levenshtein similarity >0.9) excluded to avoid leakage. It tests over 10 reasoning abilities (e.g., diagnostic reasoning, multi-hop reasoning, mathematical reasoning) and covers diverse answer types (drug names, treatment procedures, disease causes). For this study, we evaluate on the English development set.

D.3.2 MedQA

MedQA (Jin et al., 2020) is the first large-scale open-domain multiple-choice QA dataset for medical scenarios, collected from professional medical licensing exams (USMLE, MCMLE, TWMLE) across three regions, covering English, simplified Chinese, and traditional Chinese with a total of 61,097 MCQA pairs. The dataset is split into 80% training, 10% development, and 10% test sets, with each question accompanied by 4 options (one correct, three distractors). It provides a supporting document collection curated from 18 English medical textbooks (shared by USMLE and TWMLE) and 33 simplified Chinese textbooks (for MCMLE), covering 88.0% to 100.0% of the question knowledge needs. Questions are categorized into Type 1 (single knowledge point queries) and Type 2 (clinical case-based queries requiring multi-hop reasoning), with USMLE dominated by Type 2 questions (98.0%). We use the English USMLE test set.

D.3.3 MMedBench

MMedBench (Qiu et al., 2024) is a multilingual medical benchmark, constructed by aggregating QA pairs from five existing datasets (MedQA, IgakuQA, FrenchMedMCQA, RuMedDaNet, Head-QA) covering six languages. It contains 53,566 multiple-choice QA pairs (45,048 for training and 8,518 for testing) across 21 medical disciplines, with part of the test set supplemented by human-validated rationales generated via GPT-4. For this study, only the English subset of MMedBench is used.

D.3.4 MMLU

The Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021) is a general QA benchmark covering 57 diverse subjects across humanities, social sciences, STEM, and professional fields (e.g., law, medicine, ethics). It consists of 14,079 multiple-choice test questions (each with 4 options and one correct answer), with each subject containing at least 100 test examples, spanning difficulty levels from elementary to advanced professional. It supports zero-shot and few-shot evaluation.

D.3.5 ARC

The AI2 Reasoning Challenge (ARC) dataset (Clark et al., 2018) is a general QA benchmark constructed from natural grade 3–9 science curricula, containing 7,787 multiple-choice questions

(each with 4 candidate options) covering biology, chemistry, physics, earth science, and other natural science topics. It is partitioned into three subsets: ARC-Easy (5,197 questions solvable by simple retrieval/co-occurrence methods), ARC-Challenge (2,590 questions answered incorrectly by both retrieval and word co-occurrence algorithms), and a combined dev/test partition (869/3,548 questions). The dataset emphasizes integration of scientific knowledge and diverse reasoning types (e.g., multihop, counterfactual, analogy), with a supporting ARC Corpus of 14M science-related sentences covering $\sim$95% of Challenge set knowledge needs.

D.4 Baseline CPT Methods

D.4.1 DSIR

DSIR (Xie et al., 2023) is an importance resampling-based data selection framework for matching raw data to a target distribution. In our experiments, the target data is the training set of benchmark tasks. It maps raw data (from MMedC) and benchmark training data into a hashed n-gram feature space, learns their feature distributions via a bag-of-ngrams model, and computes importance weights as the ratio of target to raw feature probabilities. Samples are resampled without replacement (via Gumbel top-k trick) to prioritize those with high n-gram similarity to the benchmark training set, balancing relevance and diversity.

D.4.2 ETA-DACP

ETA-DACP (Xie et al., 2024) is a task-agnostic data selection method for domain-adaptive CPT. It scores novelty (approximated by perplexity via a lightweight surrogate model) and diversity (measured by part-of-speech tag entropy). Supporting both hard (top-k ranking) and soft (metric quantile-based probability) sampling, the original method selects 10% of the domain corpus.

D.4.3 FineWeb-Edu

FineWeb-Edu (Penedo et al., 2024) is a quality-based data selection method that uses a classifier to identify samples with high educational value. Following the referenced literature, we employ the `HuggingFaceFW/fineweb-edu-classifier` , which is trained to predict the educational quality of web pages. We compute the educational scores for all candidate samples in the corpus and select the subset with the highest scores for CPT. This applies a general educational-quality classifier to the domain corpus.

D.4.4 PPL

PPL (Yin & Rush, 2025; Marion et al., 2023) is a perplexity-based data selection method, following the paradigm described in the referenced literature. We compute the perplexity (PPL) of each candidate sample using the model to be trained, and select the subset with the highest PPL scores for CPT.

D.4.5 FCD-Soft

FCD-Soft Sample adopts a weighted probabilistic sampling strategy for CPT data selection. Unlike hard filtering that only retains high-difficulty samples, it prioritizes samples with higher FCD scores by assigning them higher sampling probabilities while retaining a non-zero probability for lower-scoring documents. Following the CDF-Based Balanced Sampling strategy in (Ying et al., 2025), we map FCD scores to sampling probabilities via cumulative distribution function (CDF) normalization.

D.4.6 FCD-CL

Curriculum Learning (CL) orders training from simpler to more difficult samples. Following the paradigm in (Naïr et al., 2024), we adopt its Hybrid three-stage CL strategy for CPT: first training on low-FCD samples (easy), then combining top 50% high-FCD easy samples with medium-FCD samples, and finally integrating high-FCD samples from the previous two stages with high-FCD (hard) samples.

E Additional Experiment on Llama-3.2-1B

We repeat the medical experiment with **Llama-3.2-1B** to test FCD-Hard on a second model family and size. As shown in Table 7, FCD-Hard obtains a domain average of 27.65, compared with 27.11 for Base CPT. The four data-selection baselines score between 26.49 and 26.92. This result has the same direction as the primary Qwen2.5-0.5B experiment, although experiments on two small models do not establish a general scaling trend.

Table 7: CPT experiments on **Llama-3.2-1B** comparing established baselines with FCD variants on domain-specific benchmarks. Values represent Accuracy (%), with **best** and second best results highlighted (excluding Base Model/CPT).

Category	Method	MedMCQA	MedQA	MMedBench	Avg
Original	Base Model	27.80	27.34	25.45	26.86
	Base CPT	28.45	27.49	25.37	27.11
Baselines	DSIR	27.80	26.00	**26.94**	26.92
	ETA-DACP	**29.52**	26.00	25.14	26.89
	Fineweb-Edu	29.29	25.37	24.82	26.49
	PPL	29.21	25.92	25.45	26.86
Ours	FCD-Hard	29.38	**27.18**	26.39	**27.65**

F Results by Test-Item Difficulty

We categorize each domain test set by FCD score. Using the $1/3$ and $2/3$ quantiles of training-set FCD scores as thresholds, we partition test samples into **Easy**, **Medium**, and **Hard** groups. Aggregated results across these partitions are provided in Appendix G.

We compare FCD-Hard with Base CPT and the data-selection baselines in Table 8 and Table 9. On the aggregated results, FCD-Hard scores above Base CPT in all three groups: by 1.36 points on Easy, 2.56 on Medium, and 1.20 on Hard. The largest point-estimate gain occurs in the Medium group. These results describe how gains are distributed under the chosen partition; they do not identify why the groups respond differently.

Table 8: Difficulty-level breakdown (Easy, Medium, Hard) across domain tasks. Comparison relative to Base CPT shown in superscripts.

		MedMCQA			MedQA			MMedBench		
Category	Method	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
Original	Base Model	38.15	30.00	39.66	33.72	38.17	33.71	33.72	36.64	35.24
	Base CPT	38.15	34.44	43.04	34.89	43.51	35.10	33.96	39.69	37.20
Baselines	DSIR	$37.01^{\downarrow 1.14}$	$32.22^{\downarrow 2.22}$	$40.51^{\downarrow 2.53}$	$34.19^{\downarrow 0.70}$	$38.17^{\downarrow 5.34}$	$36.22^{\uparrow 1.12}$	$\mathbf{35.36}^{\uparrow 1.41}$	$35.11^{\downarrow 4.58}$	$35.38^{\downarrow 1.82}$
	ETA-DACP	$36.49^{\downarrow 1.66}$	$27.78^{\downarrow 6.67}$	$39.66^{\downarrow 3.38}$	$35.60^{\uparrow 0.70}$	$42.75^{\downarrow 0.76}$	$33.99^{\downarrow 1.12}$	$34.66^{\uparrow 0.70}$	$36.64^{\downarrow 3.05}$	$36.92^{\downarrow 0.28}$
	Fineweb-Edu	$37.37^{\downarrow 0.78}$	$34.44^{\uparrow 0.00}$	$41.77^{\downarrow 1.27}$	$36.07^{\uparrow 1.17}$	$42.75^{\downarrow 0.76}$	$33.71^{\downarrow 1.40}$	$34.19^{\uparrow 0.23}$	$41.22^{\uparrow 1.53}$	$36.50^{\downarrow 0.70}$
	PPL	$35.79^{\downarrow 2.36}$	$30.00^{\downarrow 4.44}$	$41.77^{\downarrow 1.27}$	$\mathbf{37.00}^{\uparrow 2.11}$	$38.17^{\downarrow 5.34}$	$34.97^{\downarrow 0.14}$	$34.89^{\uparrow 0.94}$	$41.22^{\uparrow 1.53}$	$35.10^{\downarrow 2.10}$
Ours	FCD-CL	$37.91^{\downarrow 0.23}$	$34.44^{\uparrow 0.00}$	$\mathbf{44.30}^{\uparrow 1.27}$	$33.96^{\downarrow 0.94}$	$\mathbf{45.80}^{\uparrow 2.29}$	$35.80^{\uparrow 0.70}$	$32.79^{\downarrow 1.17}$	$42.75^{\uparrow 3.05}$	$36.36^{\downarrow 0.84}$
	FCD-Soft	$37.71^{\downarrow 0.44}$	$30.00^{\downarrow 4.44}$	$41.35^{\downarrow 1.69}$	$35.83^{\uparrow 0.94}$	$39.69^{\downarrow 3.82}$	$36.78^{\uparrow 1.68}$	$35.13^{\uparrow 1.17}$	$38.17^{\downarrow 1.53}$	$\mathbf{37.62}^{\uparrow 0.42}$
	FCD-Hard	$\mathbf{39.65}^{\uparrow 1.50}$	$\mathbf{36.67}^{\uparrow 2.22}$	$42.62^{\downarrow 0.42}$	$36.53^{\uparrow 1.64}$	$44.27^{\uparrow 0.76}$	$\mathbf{37.62}^{\uparrow 2.52}$	$33.72^{\downarrow 0.23}$	$\mathbf{44.27}^{\uparrow 4.58}$	$\mathbf{37.62}^{\uparrow 0.42}$

G AGGREGATED RESULTS ACROSS DIFFICULTY LEVELS

Table 9: Weighted aggregated accuracy comparison across FCD difficulty bins (Easy, Medium, Hard).

Category	Method	Easy	Medium	Hard
Original	Base Model	37.35	35.51	35.21
	Base CPT	<u>37.47</u>	39.77	37.13
Baselines	DSIR	36.60	35.51	36.47
	ETA-DACP	36.22	36.65	36.05
	Fineweb-Edu	36.96	40.06	36.05
	PPL	35.82	37.22	35.99
Ours	FCD-CL	37.09	<u>41.76</u>	37.25
	FCD-Soft	37.30	36.65	<u>37.79</u>
	FCD-Hard	**38.83**	**42.33**	**38.33**

H IMPACT OF SELECTION RATIOS ON PERFORMANCE

We evaluate FCD-Hard at different data selection ratios. Specifically, for **FCD-Hard**, we select subsets ranging from 1/12 to 3/4 of the original 6B corpus (corresponding to 0.5B to 4.5B tokens).

The **Base CPT** baseline can be evaluated at different checkpoints during one continuous training run. Because our sampling method is offline, we train a separate model from the same initialization for each selection ratio. Each run consumes its selected subset once.

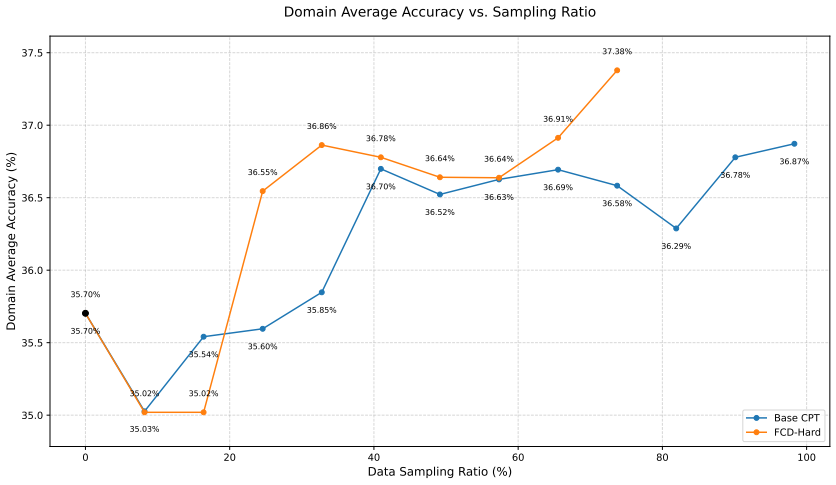


Figure 6: Performance on medical benchmarks across different selection ratios. The x-axis represents the sampling ratio (from 0.5B to 4.5B tokens), and the y-axis shows the average score on Domain QA tasks.

In Figure 6, the **1/3** subset reaches a domain average close to the final Base CPT checkpoint. The FCD-Hard curve changes little from the 1/3 to 3/4 ratios, and the 3/4 run has the highest FCD-Hard score shown. Because each ratio corresponds to one training run, the curve does not establish whether the local fluctuations are statistically reliable.

I Discussion of Baseline Performance

Table 1 shows that DSIR, ETA-DACP, Fineweb-Edu, and PPL score below the unfiltered **Base CPT** on the domain average under the same token budget. These methods optimize different proxies: POS-tag entropy for ETA-DACP, n-gram similarity for DSIR, general educational quality for Fineweb-Edu, and model surprise for PPL. A mismatch between these proxies and the target medical benchmarks is one possible explanation for the result. However, the current experiments do not isolate the contribution of each proxy, so we do not attribute the performance differences to a single cause.

FCD-Hard and FCD-Soft exceed Base CPT by 1.36 and 0.25 points on the domain average, respectively. This comparison supports frequency and WordNet distance as useful selection signals for this corpus, without implying that the same ordering will hold in other domains.

J Aggregation Strategy: Peak vs. Mean Difficulty

We compare the quantile-based aggregation in **FCD-Hard** with a simple average of word-level scores, denoted as **FCD-Hard (avg)**. Table 10 shows domain averages of 38.23 and 35.78, respectively; the mean variant also scores below **Base CPT**. In this setting, assigning more weight to high-scoring words produces a more useful document ranking than averaging all word scores.

Table 10: Ablation of FCD aggregation strategies on medical benchmarks.

Method	MedMCQA	MedQA	MMedBench	Avg
Base CPT	38.35	35.90	36.37	36.87
FCD-Hard	**39.76**	**37.94**	**37.00**	**38.23**
FCD-Hard (avg)	36.86	35.51	34.96	35.78

K Case Study on Financial Corpus

To illustrate which words contribute to Frequency-Concept Difficulty (FCD), we sample three document snippets each from the top 10% and bottom 10% of the FCD score distribution in the finance domain. Specifically, we **boldface the top 30% of nouns** and underline the top 10% of non-nouns ranked by FCD scores, with illegal tokens and stop words excluded from the selection process. Part-of-speech tagging is implemented using NLTK.

In these examples, several high-FCD financial-corpus documents contain vocabulary from other domains. Biological terms such as *vaquita*, *GMO*, and *Bioengineered*, as well as communication terms including *bandwidth* and *broadband*, occur in the selected high-FCD snippets. The low-FCD snippets include financial terms such as *grants*, *scholarship*, and *repayment*. These six examples illustrate the lexical patterns discussed in subsection 6.3, but are not a corpus-level estimate.

Category: High FCD (Top 10%)

Sample 1: Background: Minimizing **fishery bycatch** threats might involve trade-offs between maintaining viable populations and economic benefits. Understanding these trade-offs can help managers reconcile conflicting goals. An example is a set of **bycatch** reduction measures for the **Critically Endangered** vaquita **porpoise** (**Phocoena sinus**), in the Northern Gulf of California, Mexico. The **vaquita** is an endemic species threatened with **extinction** by artisanal net **bycatch** within its limited range; in this area fisheries are the chief source of economic **productivity**.
Methodology/Principal Findings: We analyze trade-offs between **conservation** of the **vaquita** and fisheries, using an end-to-end **Atlantis ecosystem** model for the Northern Gulf of California. **Atlantis** is a spatially-explicit model intended as a strategic tool to test alternative management strategies. We simulated increasingly restrictive fisheries regulations contained in the vaquita **conservation** plan: **implementing** progressively larger spatial management areas that exclude **gillnets**, shrimp **driftnets** and introduce a fishing gear that has no vaquita **bycatch**. We found that only the most extensive spatial

management **scenarios** recovered the vaquita population above the threshold necessary to downlist the species from **Critically Endangered**. The scenario that excludes existing net gear from the 2008 area of vaquita distribution led to moderate decrease in net present value (US$ 42 million) relative to the best-performing scenario and a two-fold increase in the abundance of adult **vaquita** over the course of 30 years.

Sample 2: In 2016, various states considered legislation to require the **labeling** of products made with genetically modified organisms (**GMOs**), which would have created a **patchwork** of different state laws. And in June of that year, **Vermont** began **implementation** of their **GMO labeling** law. **CRA** undertook an economic impact analysis of the **Vermont** law and demonstrated that the cost of **implementation** would lead to an average increase in annual food cost of $1,050 per American family. The analysis also found that lower income families would have been especially burdened, as the increased costs could account for nearly 2.5 percent of the median income for the poorest fifth of the population. To avoid the confusion and expense of conflicting state laws, Congress passed the National **Bioengineered Labeling** Standard in July 2016. The law mandates that the U.S. Department of Agriculture (**USDA**) finalize a **GMO** labeling standard by July 2018. While **USDA** has missed this deadline, the Fall 2018 **Unified** Regulatory Agenda notes that the final rule should be available in November 2018. The standard set by Congress is solely intended to address **labeling** and marketing and is neither a health nor safety standard. Congress also directed the **USDA** to minimize the law's impact on the domestic and international value chain. The **USDA** is expected to mandate that firms choose at least one of the following three options for **GMO** disclosure: a **QR** code, an on-pack label, or on-pack text to be determined by the **USDA** Secretary. **CRA** believes the **USDA**'s proposal should be consistent with other global standards and limit **burdensome** and costly analyses. Further, the law notes that bioengineered food includes modified genetic material. Refined **ingredients** sourced from bioengineered **crops**, such as refined corn products, are processed as to no longer contain modified genetic material at measurable levels. As a result, **CRA** believes that the threshold for disclosure should be practical and that refined **ingredients** sourced from bioengineered **crops** do not warrant disclosure. This would allow for **flexibility** of the rule, avoid consumer **misunderstanding**, reduce the potential for negative trade **implications**, and reflect the broad scientific **consensus** behind the safety of bioengineered **ingredients**.

Sample 3: With **loads** of **fanfare**, **excitement**, and speculation, 5G is slowly creating the **transition** from a general plan to network **implementation**. Network operators are already planning 5G network trials and check beds to sort out how the 5G **vision** will be realized. 5G networks promise to support new services, more video, and cloud **connectivity**. There are 3 main case uses driving the 5G revolution:

- **Enhanced** mobile **broadband**. With the promise of 10 **Gbps connectivity** and **latency** of but 5 **milliseconds**, it's no surprise the continuing **surge** in demand for mobile **connectivity** can accelerate dramatically. The industry estimates this enhanced speed can end in a 10- to 100-fold increase within the variety of 5G-connected **devices** over the amount of 4G **devices**.
- The Internet of Things (**IoT**). because of 5G's virtualized, radio technology-agnostic **core**, revealed **predictions** estimate as several as twenty billion **IoT** connections by 2020—connections that will drive smart buildings and smart cities. **CommScope** anticipates 5G can provide 1,000 times the **bandwidth** of 4G, and up to 5 times the **density**, creating space for all those "things" on the network.
- High-reliability, low-latency networks. Beyond simply doing what 4G does better and faster, 5G speeds open new doors to permit driverless cars to coordinate over the network, enable augmented reality and virtual reality, and expand the **horizons** of remote **surgery** and alternative applications which will fulfill their promise solely on a network with such ultra-low **latency** times as 5G's five-millisecond **threshold**.

Category: Low FCD (Bottom 10%)

Sample 1: Types of financial aid available There are a few kinds of financial aid which college students can receive. They are **grants** and **scholarships**, student **loans**, and part-time jobs through the college. **Grants** and **scholarships** are financial help that college students receive which they do not need to pay back. In other **words**, the student has to qualify for this financial aid and once he or she qualifies for it he or she does not have to pay it back. Government **grants** and **scholarships** are free money, in that sense, the student receives from either the government, financial **institutions**, or the college itself. **Scholarships** are sometimes termed as college financial aid. Student **loans** are provided by financial **institutions** to students at low-interest **rates**. Although interest **rates** may be lower than the usual interest **rates**, these

loans still have to be repaid. **Repayment** can be a very difficult task if the student does not find a proper job after his education. These may be the last option for students looking to find financial aid for college since they have to be repaid. In some cases students are offered part-time jobs in the colleges to help them pay the **fees**. This is often called a work study program. This is a very good opportunity for the student to gain some experience at work and also be able to earn some money to pay the **fees** as well as to take care of their personal needs. This is a job, so students do not have to repay the money earned from work study programs. Though it may not be very easy task to find government financial aid for college, it is well worth the time invested. If you are looking to continue **schooling** and go for higher education, then it may be well worth your time and to try to find government financial aid.

Sample 2: A **Value Curve** is a **diagram** which can be used to show instantly where **value** is created within an organisation's products or **services**, showing graphically how a company can create new market spaces and provide **value** to customers against competition. The **Value Curve**, from Blue Ocean Strategy, has two purposes. Firstly, it shows the current competitive **landscape** in the marketplace as it exists today. This information allows you to see where competitors are heavily investing and how different products are being positioned to customers. Secondly, the **Value Curve** allows you to use this information to find market space, which does not compete with existing competitors. When drawing a **Value Curve** the vertical **axis** represents the products available in the marketplace, or the offering available to the **consumer**. The horizontal **axis** represents the factors the industry competes on. You can also consider the **horizontal axis** to be the range of factors the industry invests in. A higher score represents the customer being offered more, **hence** we can say that the company invests more in this area. We can now draw the **value curve** by plotting the relative **values** of the different offerings for each of the factors on the horizontal **axis**.

Sample 3: An economy **grows** and reaches new stages of development, so the share of national **output** from each of the main production **sectors** changes. The standard **hypothesis** is that the long road to economic **maturity** brings a **switch** from **manufacturing** and construction **towards** service industries in the tertiary and quaternary **sectors**. This is because the income **elasticity** of demand for services is higher than for primary and manufactured products. As average living **standards** grow, the pattern of **consumer** demand **tilts** away from goods **towards** services. Increased government **spending** on services (such as education and health) adds to this comparative advantage between countries changes over time, we expect to see nations switching their resources to industries where they can exploit a new advantage. In **Britain**'s case, it is widely held that financial services is a **sector** where we hold a significant advantage over other nations.