Perplexity-Aware Data Scaling Law: Perplexity Landscapes Predict Performance for Continual Pre-training

Lei Liu[1,2], Hao Zhu[2], Xiaoyan Yang[2], Yue Shen[2], Jian Wang[2], Jinjie Gu[†,2], Zhixuan Chu[†,1], Kui Ren[1]

[1]The State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]Ant Healthcare, Ant Group

Contact: liulei1497@gmail.com zhixuanchu@zju.edu.cn

Abstract

Continual Pre-training (CPT) serves as a fundamental approach for adapting foundation models to domain-specific applications. Scaling laws for pre-training define a power-law relationship between dataset size and the test loss of an LLM. However, the marginal gains from simply increasing data for CPT diminish rapidly, yielding suboptimal data utilization and inefficient training. To address this challenge, we propose a novel perplexity-aware data scaling law to establish a predictive relationship between the perplexity landscape of domain-specific data and the test loss. Our approach leverages the pre-trained model's own perplexity on domain data as a proxy for estimating the knowledge gap, effectively quantifying the informational perplexity landscape of candidate training samples. By fitting this scaling law across diverse perplexity regimes, we enable adaptive selection of high-utility data subsets, prioritizing content that maximizes knowledge absorption while minimizing redundancy and noise. Extensive experiments on both medical and general-domain benchmarks demonstrate that our method consistently identifies near-optimal training subsets, achieving superior performance with significantly reduced data consumption.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across a wide range of domains (Liu et al., 2024). However, their general-purpose pre-training objectives often leave them ill-suited for specialized applications such as healthcare, where domain-specific knowledge, precise terminology, and structured reasoning are critical. To bridge this gap, Continual Pre-Training (CPT) has emerged as a dominant paradigm (Que et al., 2024): by further pre-training a general-purpose LLM on domain-specific corpora, models can internalize nuanced medical concepts, factual knowledge, and domain-typical reasoning patterns, thereby improving performance on downstream tasks.

Despite its success, CPT remains largely guided by heuristic data practices, with limited understanding of how data characteristics influence learning dynamics (Wang et al., 2025; Chen et al., 2025b). A key challenge is the inefficiency of scaling data under continual pre-training. Classical scaling laws exhibit power-law predictive relationship between loss and dataset size, but they assume that each token contributes equally to learning (Hoffmann et al., 2022). In practice, domain-specific corpora exhibit high levels of redundancy and noise, and vary significantly in conceptual density. For instance, biomedical literature may contain long passages that restate known facts, while clinical notes often include unstructured or repetitive entries. As a result, simply increasing the amount of training data leads to sharply diminishing returns. This observation underscores the need for data-informed strategies that move beyond raw data volume and instead emphasize the quality, diversity, and informational value of training samples.

This breakdown calls for a shift from purely quantity-driven paradigms to data-centric strategies that explicitly account for sample effectiveness in CPT (Yu et al., 2024; Engstrom et al., 2024). Rather than treating all domain texts equally, we argue that one should prioritize instances that most effectively reduce the model's uncertainty, particularly those that target its most salient knowledge gaps. A natural question then arises: how can we quantify such gaps in a manner that is both computationally efficient and strongly correlated with downstream performance improvements?

In this work, we propose that the answer lies in the model's own uncertainty signal: *perplexity* (Ankner et al., 2024). We introduce the concept of *perplexity landscapes*, fine-grained distributions of model perplexity over streaming domain data, as

[†]Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21889–21901

a powerful diagnostic tool for characterizing the knowledge frontier between general and domain-specific expertise. Crucially, we observe that the shape of these landscapes at early stages of CPT strongly correlates with eventual fine-tuning performance, suggesting that initial perplexity encodes actionable information about data utility.

Building on this insight, we derive a novel *perplexity-aware data scaling law* that establishes a predictive functional relationship between statistics of the initial perplexity distribution (e.g., mean, variance, tail mass) and final task performance after CPT. Unlike traditional scaling laws based solely on data volume, our law incorporates intrinsic model responses to individual data points, enabling adaptive selection of training subsets that maximize knowledge absorption while filtering out redundant, overly difficult, or noisy samples.

Our method requires only a single forward pass over unlabeled domain data using the frozen initial model, making it efficient and scalable. By fitting the scaling law on small pilot batches, we can estimate the expected return of larger data subsets and select those predicted to yield optimal performance. This enables principled, model-informed data curation without requiring labeled examples or expensive retraining loops. We validate our approach across medical and general benchmarks. Results show that perplexity landscapes consistently identify near-optimal data subsets, achieving a superior improvement with perplexity as a proxy for knowledge acquisition. Our contributions are threefold:

- We introduce perplexity landscapes as a diagnostic tool for CPT, where perplexity distributions are strongly predictive of downstream performance, providing a window into the evolving knowledge frontier during specialization.
- We propose a novel perplexity-aware data scaling law to link statistics of data perplexity to final CPT performance, which moves beyond only scaling data volume.
- We develop an efficient model-aware data selection framework to identify high-value training samples. Our method enables scalable data curation for achieving superior performance with less data and demonstrates consistent gains across medical and general benchmarks.

2 Related Work

Scaling Law A growing body of research has established that the performance of large language models (LLMs) follows predictable scaling laws with respect to key resources such as model size, training compute, and dataset size (Kaplan et al., 2020; Hoffmann et al., 2022). In particular, numerous studies have demonstrated that model performance improves according to a power-law relationship as the number of parameters or the volume of training data increases (Kaplan et al., 2020; Hoffmann et al., 2022), enabling principled extrapolation from small-scale experiments to large-scale deployments. These scaling laws provide a theoretical foundation for optimizing training efficiency and guide decisions in model design and data budget allocation. More recently, scaling laws have been extended beyond parameter and data size to encompass more nuanced factors, such as data mixture proportions. For instance, Que et al. (2024) proposed a data-mixing scaling law that predicts performance based on the composition of multi-domain training sets, offering guidance for curriculum and domain-adaptive pre-training.

Active Learning Active learning focuses on data efficiency (selecting the best data) and scaling laws focusing on compute/data efficiency (predicting performance based on quantity). They are increasingly seen as complementary forces to optimize the training trajectory of LLMs. The core distinction from active learning lies in problem setting and feedback mechanisms (Yu et al., 2024; Wang et al., 2024). While AL relies on iterative human-in-the-loop querying via uncertainty sampling, our work leverages pre-trained model perplexity as a static proxy for knowledge gap to perform one-time data selection. AL focuses on model uncertainty (e.g., entropy) (Xia et al., 2025), whereas we establish a predictive scaling law between perplexity landscapes and performance, explicitly modeling how perplexity contributes to test loss reduction. Given that CPT operates at billion-token scale where iterative retraining is prohibitively expensive, our approach offers superior computational efficiency and theoretical interpretability.

3 Data-centric Scaling Law by Perplexity

Unlike pretraining, the primary challenge in CPT lies in balancing knowledge retention (the preservation of the model's previously acquired capabilities)

with knowledge acquisition (effective adaptation to and integration of new information). Consequently, data selection becomes even more critical in CPT.

3.1 Motivation

Motivation-1: Marginal Gain of Scaling Data Diminishes During CPT. Classical scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) works with a potential assumption, *uniformly informative data*. For example, empirical scaling law (Hoffmann et al., 2022) fits a parametric loss function, building the systematic, predictable connections among model size, training dataset size, and the model's final performance.

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} \tag{1}$$

where N denotes the parameters and D represents the number of training tokens. α is parameter scaling exponent and β is dataset scaling exponent. E corresponds to the entropy of natural text. These laws quantify how performance improves as the attributes are "scaled up", providing a foundational framework for guiding LLM development.

However, during CPT, the marginal gain from simply increasing dataset size diminishes, where data quantity becomes inadequate predictor for model performance due to the uncertainty from the base model and significantly varying data distribution. This suggests that conventional scaling laws, which focus solely on quantity, may be insufficient to capture the dynamics of model refinement in later training stages. Noticing that factors related to data effectiveness (*e.g.*, knowledge density, topical relevance, and factual accuracy) emerge as primary drivers of further performance improvement, there is a compelling need to extend the scaling law framework to incorporate a dataset importance weighting dimension, potentially yielding a two-axis formulation that jointly models the effects of dataset size and dataset importance on loss reduction. This motivates the question:

Given a pre-trained model, under the condition of a fixed number of training texts, how to determine the optimal training data subset for continual pre-training?

Motivation-2: Perplexity Landscapes Predict CPT Performance. To select the effective data, perplexity provides a natural metric (Ankner et al., 2024): sequences with low PPL are redundant, while those with high PPL are likely noisy or incomprehensible, both yield diminishing learning returns. The most effective data lie in a "sweet spot" of moderate perplexity, which formalizes the intuition that the most valuable data is neither too similar nor too diverse to the model.

Starting with trained models under different data subsets of varying PPL distributions, experimental results are used to fit an empirical estimator for determining the optimal perplexity range. This formulation extends scaling laws by incorporating perplexity statistics, *i.e.*, the mean and variance. Given the fixed tokenizer and corpus, such statistics become dimensionless across model scales.

3.2 Perplexity-Aware Data Scaling Law

3.2.1 Scaling Law Formula

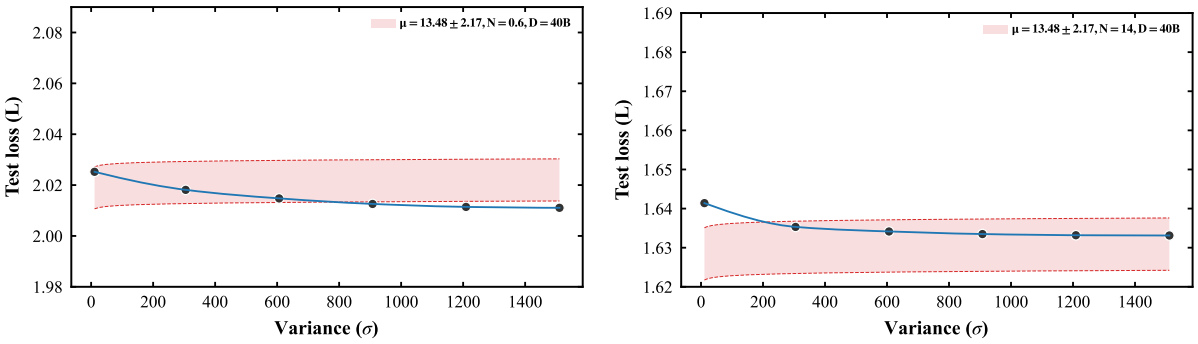
Focusing on the data-centric term in (Kaplan et al., 2020; Hoffmann et al., 2022), we utilize the following functional form over the both dataset size and importance:

$$\hat{L}(Q, D) \triangleq E + \frac{D_c}{Q * D^{\alpha_D}}, \tag{2}$$

where D_c and α_D are hyper-parameters to be fitted. E corresponds to the entropy of natural text. Here, we omit the starting status of the pre-training model because the following perplexity descriptor contains such information. The above formulation typically assumes homogeneous data distributions, treating Q as a constant or implicit factor. Here, we consider parameterize the dataset importance term using basic informativeness measurement.

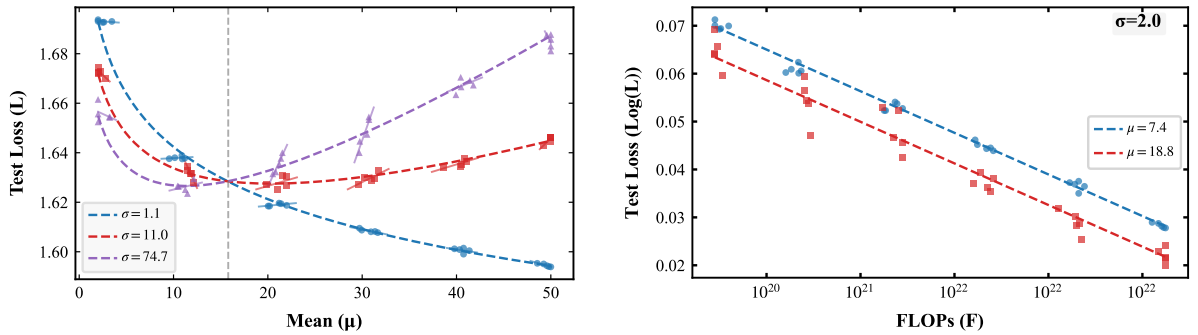
Perplexity (PPL), as a token-level likelihood measure under a reference model, provides a fine-grained proxy for sample informativeness (Brown et al., 2020; Touvron et al., 2023). We argue that summarizing the PPL distribution across a dataset via its statistic distribution (mean and variance) offers a principled and scalable importance indicator. In detail, the mean captures the average sample difficulty, while the variance reflects diversity, both of which can influence learning dynamics and generalization (Zhang et al., 2025). Incorporating these statistics allows for a more nuanced understanding of how data modulates model performance, enabling better predictions and resource allocation in large-scale training regimes.

Accordingly, let μ and σ indicate the mean and variance of PPL distribution respectively. We instantiate the dataset importance term as $Q(\mu, \sigma) = \mu^{\alpha_\mu} * \sigma^{\alpha_\sigma}$, which models the joint influence of mean and variance in a scale-invariant manner.



(a) Qwen3-0.6B-Base (b) Qwen3-14B-Base

Figure 1: Interdependence between loss and μ (σ).



(a) Test Loss *VS.* PPL Distribution (b) Test Loss (Log) *VS.* FLOPs

Figure 2: Fitted curves for scaling law based on Qwen3-14B-Base.

Here, α_μ and α_σ are hyper-parameters. By enriching the scaling law with interpretable, data-driven signals, we have:

$$\hat{L}(\mu, \sigma, D) \triangleq E + \frac{D_c}{(\mu^{\alpha_\mu} * \sigma^{\alpha_\sigma}) * D^{\alpha_D}}. \quad (3)$$

Empirical experiments (Figure 1) indicate that the relation between loss and μ (σ) is not strictly monotonic, while these two variables exhibit a measurable interdependence. Thus, we introduce the minimal multiplicative interaction to extend Eq. 3 as follows:

$$\alpha_\mu(\sigma) = \alpha_0 + \alpha_1\sigma, \quad \alpha_\sigma(\mu) = \beta_0 + \beta_1\mu, \quad (4)$$

where α_1 and β_1 are relationship variables between μ and σ. Such transformation results in the following format of perplexity-aware data scaling law:

$$\hat{L}(\mu, \sigma, D) \triangleq E + \frac{D_c}{(\mu^{\alpha_\mu(\sigma)} * \sigma^{\alpha_\sigma(\mu)}) * D^{\alpha_D}}, \quad (5)$$

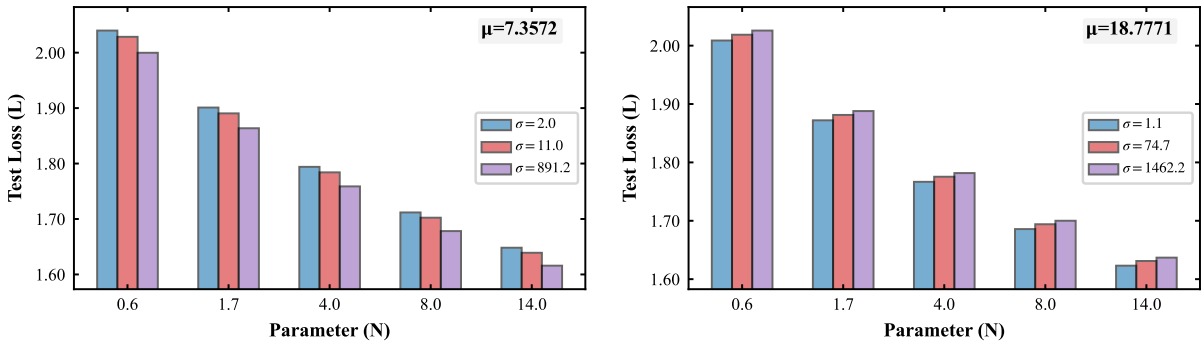
which preserves the interpretable power-law structure. Besides, it incorporates the relationship decomposition between mean and variance of perplexity distribution, *i.e.*, interdependence and independence decomposition:

$$L(\mu, \sigma, D) \triangleq E + \frac{D_c}{\underbrace{\mu^{\alpha_0}\sigma^{\beta_0}}_{\text{Independence}} * \underbrace{\mu^{\alpha_1\sigma}\sigma^{\beta_1\mu}}_{\text{Interdependence}} * D^{\alpha_D}}. \quad (6)$$

This format is strongly favored because it (i) conserves the power-law structure, (ii) keeps interactions simple, and (iii) collapses to the basic model when variables are independent.

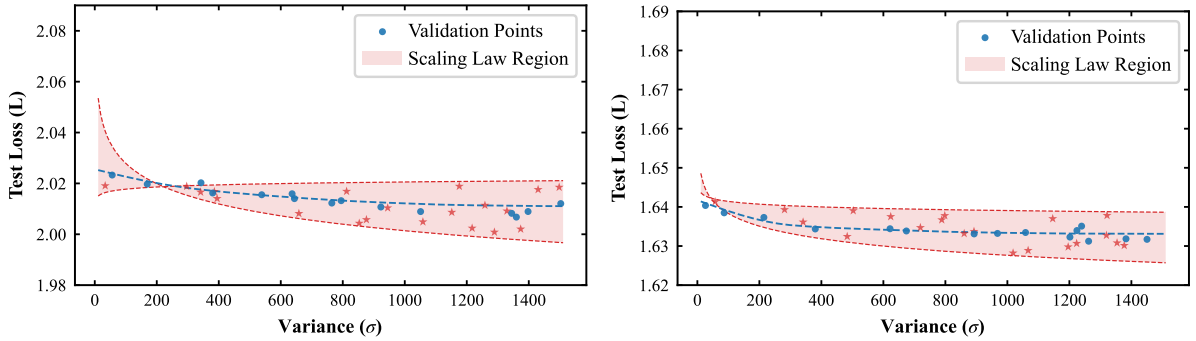
3.3 Scaling Law Fitting

Setting We select medical domain as the target. Given the dataset from PubMed corpus, we firstly perform multiple bootstrap sampling to generate different subsets with varying μ and σ. Among these training subsets, 90% subsets are used for fitting the scaling curve and 10% subsets are for validation. The final fitting curves with sampled fitting points are shown in Figure 2. The validation results are shown in Figure 4. The red region is obtained according to the fitted scaling law with $\mu = 13.48 \pm 2.17$ and σ varying from 25 to 1600



(a) Test Loss *VS.* Parameters (Low PPL Range) (b) Test Loss *VS.* Parameters (High PPL Range)

Figure 3: Performance changes with varying PPL distributions and Parameters.



(a) Qwen3-0.6B-Base (b) Qwen3-14B-Base

Figure 4: Validation curves of scaling law.

with $\mu = 13.48 \pm 2.17$, which is the real PPL range of the dataset. It is observed that the fitted scaling law closely matches the validation points (blue) across different variances and maintains a consistent trend. This strong agreement confirms that the formula accurately captures the relationship between dataset importance, and performance.

Why Consider Interaction Term in Scaling Law? Eq. 3 assumes a monotonic relationship between the loss and each variable, while experimental results indicate that the relationship between the loss and the mean μ or standard deviation σ is not purely monotonic. As shown in Figure 1, fitting Eq. 3, the real samples (black points) are not strictly fall in the region we fitted (red area). Furthermore, since both μ and σ are derived from the same PPL distribution, an intrinsic correlation between them may exist. Therefore, it is necessary to introduce interaction terms into the model to better capture their interdependent effects.

Why Perplexity Distributions Follow Power-Law Forms in Scaling Laws? Power-law behavior arises from language data's inherent hierarchy. Natural language data exhibits scale-invariant hierarchical structure: linguistic units (tokens, phrases, documents, domains) follow a power-law distribution in their frequency and complexity. For example, Zipf's law (Zipf, 2013) describes how word frequencies scale as $f \propto r^{-s}$ (where r is rank and $s \approx 1$), and document complexity (measured by syntactic depth or semantic ambiguity) similarly follows a power-law, where a small fraction of "high-complexity" documents drive most variation in PPL (Wold et al., 2024). This hierarchy directly shapes μ and σ:

- Mean μ: As incremental data volume (D_{new}) scales, μ converges to a domain-specific limit (μ_0) because larger datasets increasingly sample the full range of linguistic complexity. The convergence rate follows a power law ($\mu - \mu_0 \propto D_{\text{new}}^{-\alpha}$, $\alpha > 0$) because the remaining "unseen" complexity (driving deviations from μ_0) is dominated by low-frequency, high-complexity data, whose contribution decays as a power of D_{new} (consistent with (Cagnetta et al., 2025)).
- Variance σ: Variance quantifies diversity in data

quality/complexity, which is inherently tied to the number of distinct subdomains (C) in D_{new}. Each subdomain contributes a unique PPL sub-distribution. The total variance scales as $\sigma \propto C^{\gamma}$ ($\gamma > 0$), *i.e.*, a power law because subdomain complexity itself follows a power-law hierarchy.

How Perplexity Landscape Affects Performance? From the data-centric perspective, the test loss of a LLM model can be predicted under the condition of data information, *i.e.*, PPL mean (μ), variance (σ), and dataset size (D). There are some key observations from Figure 2 and 3.

- In Figure 2(a), perplexity distribution has a non-monotonic effect. There exists an optimal point for μ and σ. Higher σ initially decreases loss by introducing useful diversity or hardness. However, beyond an optimal point μ would harm performance. Low-μ data benefits more from higher σ, while high-μ data suffers at very high σ.
- In Figure 2(b), with a moderate variance, lower mean yields worse convergence performance.
- In Figure 3, test loss decreases with model size (N), but the magnitude depends on data characteristics. In a low PPL range (Figure 3(a)), higher σ results in lower test loss than lower σ for the same number of parameters. In a high PPL range (in Figure 3(b)), lower σ enables significantly better scaling, whereas higher σ limits.

Could Perplexity-Aware Scaling Law Generalize to New Points? To rigorously assess the generalization of the fitted scaling law, we evaluate its predictions on an independent validation set that is entirely disjoint from the data used for fitting. The validation configurations are sampled from the same underlying data distribution while exhibiting different perplexity (PPL) profiles, ensuring that the evaluation tests extrapolation to unseen points rather than simple interpolation.

As shown in Figure 4, the validation points (blue dots) closely follow the predicted trend and remain largely within the shaded scaling-law region for both Qwen3-0.6B-Base and Qwen3-14B-Base. Across the full range of variance, the empirical test losses align well with the power-law curve, and deviations are small and symmetric, indicating no systematic bias. This consistency between the model's predictions and the independently obtained validation measurements demonstrates that the derived power-law relationship is both robust and reliable, and that it captures the dominant dependence of test loss on variance for these models.

PPL Landscapes Visualization Figure 5 visualizes how the test loss varies as a function of the mean and standard deviation of the PPL distribution, and how this relates to our scaling law.

In Figure 5(a), the level sets of test loss form smooth, roughly elliptical basins in the (mean, std) plane. The three descent paths start from different initial PPL configurations but all move along the gradient of the loss surface toward the same low-loss region, highlighted by the red star. This indicates that the loss is not determined by mean or variance alone. It depends on their joint configuration, and different PPL profiles can converge to a common optimum predicted by the scaling law.

The 3D surface (Figure Figure 5(b)) makes this relationship explicit: test loss forms a bowl-shaped surface over (mean, std), with a single, well-defined minimum. Near this minimum, loss changes smoothly and approximately follows our power-law scaling with respect to the PPL variance (for a given mean). Moving away from the optimum in either direction, by increasing or decreasing the mean or the variance, monotonically increases the loss, consistent with the scaling-law behavior observed in our 1D slices.

Thus, the PPL landscape analysis provides a geometric interpretation of the scaling law: the power-law relationship describes how test loss scales along the main descent directions in the (mean, std) space of perplexity.

3.4 Distance-to-Optimum Selection

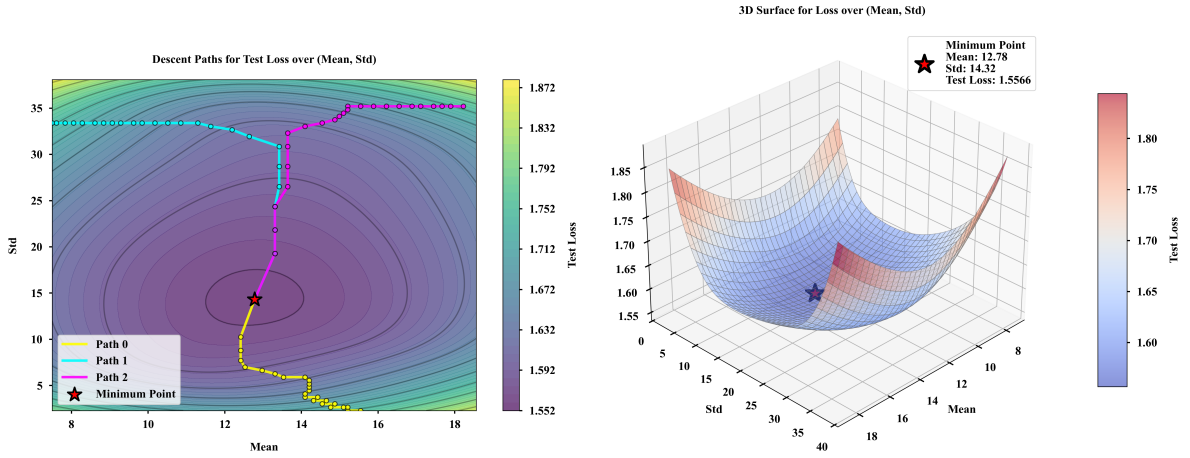
Given a fitted perplexity-aware scaling law, we assume it identifies an optimal region of the data-perplexity distribution characterized by a target mean $\hat{\mu}$ and variance $\hat{\sigma}$. Our goal is to construct a subset whose empirical perplexity statistics approximate $(\hat{\mu}, \hat{\sigma}^2)$ under a fixed token budget.

Distance-to-Optimum Objective For a subset $\mathcal{S}$, we measure the deviation from the optimal distribution by:

$$J(\mathcal{S}) = w_{\mu}\big(\mu(\mathcal{S}) - \hat{\sigma}\big)^2 + w_{\sigma}\big(\sigma^2(\mathcal{S}) - \hat{\sigma}^2\big)^2, \quad (7)$$

where $w_{\mu}, w_{\sigma} > 0$ are weighting coefficients (typically $w_{\mu} = w_{\sigma} = 1$). The Distance-to-Optimum Selection (DOS) problem can be formulated as:

$$\min_{\mathcal{S} \subseteq \mathcal{D}} \quad J(\mathcal{S}) \text{s.t.} \quad \sum_{c_j \in \mathcal{S}} |c_j| \leq T_{\text{budget}}. \quad (8)$$



(a) PPL Descent Paths

(b) Loss-PPL 3D Distribution

Figure 5: Visualizations for perplexity landscape on Qwen3-14B-base model.

Solving this problem exactly is intractable, therefore we adopt a greedy approximation.

Greedy Selection Algorithm Initially, the whole corpus is chunked into N random subsets $\mathcal{D} = \{c_j\}_{j=1}^{N}$. Define the optimal subset as $\mathcal{S} \leftarrow \emptyset$ and data budge as $T \leftarrow 0$. To initialize a non-empty subset, choose the chunk whose perplexity is closest to $\hat{\mu}$, *i.e.*, $j^* = \arg\min_j |p_j - \hat{\mu}|$ and add it to $\mathcal{S}$. The greedy expansion is conducted as following steps while $T < T_{\text{budget}}$:

- For each candidate $c_j \notin \mathcal{S}$ with $T + |c_j| \leq T_{\text{budget}}$, form $\mathcal{S}' = \mathcal{S} \cup \{c_j\}$, update $\hat{\mu}(\mathcal{S}')$ and $\hat{\sigma}^2(\mathcal{S}')$, and compute

$$J_j = w_\mu \big(\mu(\mathcal{S}') - \hat{\mu}\big)^2 + w_\sigma \big(\sigma^2(\mathcal{S}') - \hat{\sigma}^2\big)^2.$$

- Select $j^* = \arg\min_j J_j$, set $\mathcal{S} \leftarrow \mathcal{S}'$, $T \leftarrow T + |c_{j^*}|$, and update $\hat{\mu}(\mathcal{S}), \hat{\sigma}^2(\mathcal{S})$.

4 Experiment

4.1 Experimental Setting

Hyper-Parameters We employ the AdamW optimizer with hyper-parameters set to $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight-decay $= 0.1$. We set the maximum sequence length to 8K during the whole CPT stage. As for the learning rate scheduling, we first use a warming up with peak learning rate of 1×10^{-5}. During the anneal training procedure, we gradually decay the learning rate following a cosine decay curve. The gradient clipping norm is set to 1.0. The base model is Qwen3-14B-Base.

Evaluation Datasets We evaluate the models on leading benchmarks over both medical and general tasks. Medical tasks include MMLU, Diagnosis-Arena (Zhu et al., 2025), MedMCQA (Pal et al., 2022), MedQA-USMLE (Jin et al., 2021), PubMedQA (Jin et al., 2019), MedBullets (Chen et al., 2025a), NEJMQA (Katz et al., 2024), SuperGPQA-Med (Du et al., 2025), GPQA-Med (Rein et al., 2024), and medical subsets of CEVAL (Huang et al., 2023), CMMLU (Li et al., 2023), MMLU (Hendrycks et al., 2020). General tasks include CEVAL, CMMLU, and MMLU.

Baseline Four baselines are: (1) Base model without CPT, (2) RS-CPT with random data sampling, (3) QBS-CPT with quantiles-based sampling, (4) LPS-CPT with low-PPL sampling, and (5) HPS-CPT with high-PPL sampling, *i.e.*, PPL scors are lower/higher than optimal one derived from DOS.

4.2 Main Results

Loss Curve As shown in Figure 6, across all methods, test losses decrease steadily over the first several hundred training steps, but the four strategies exhibit noticeably different convergence behaviors. DOS-CPT shows the most rapid and consistent reduction in test loss, achieving the lowest final loss and demonstrating stable improvement throughout training. In contrast, HPS-CPT converges more slowly and plateaus at a higher loss, exhibiting small fluctuations. LPS-CPT performs similarly to HPS-CPT but ultimately settles at a higher loss. RS-CPT shows the weakest performance, consistently lagging behind all other methods. Overall, the figure highlights that DOS-CPT

Table 1: Effectiveness of Perplexity-Aware Data Scaling Law. All models are evaluated under the same evaluation setting. The highest, the second-best scores are shown in **bold** and underlined, respectively.

Task	Benchmark	Base	CPT				
			RS	QBS	LPS	HPS	DOS
Medical	DiagnosisArena	41.30	56.40	49.80	56.40	45.70	61.11
	GPQA-Med	57.89	57.89	53.29	57.89	57.89	63.16
	PubMedQA	76.60	76.70	75.20	76.40	78.60	77.40
	Medbullets	55.52	56.82	57.14	56.82	58.44	57.14
	NEJMQA	64.73	66.06	66.06	65.60	64.79	66.14
	MedMCQA	66.89	67.75	67.75	67.51	67.75	68.28
	MedQA-USMLE	72.58	73.21	71.80	73.21	72.66	73.61
	CEVAL-Med	89.86	90.14	90.44	89.63	91.46	90.44
	CMMLU-Med	86.68	86.78	86.38	86.89	86.58	86.38
	MMLU-Med	81.19	81.62	81.62	81.85	81.68	82.11
	Average	69.32	71.34	69.95	71.22	70.56	**72.48**
General	CEVAL	85.52	85.14	85.44	85.14	85.17	85.44
	CMMLU	84.92	84.50	84.50	84.79	84.53	84.78
	MMLU	82.03	82.23	82.13	82.03	82.11	82.25
	Average	**84.16**	83.94	84.02	83.99	83.94	**84.16**

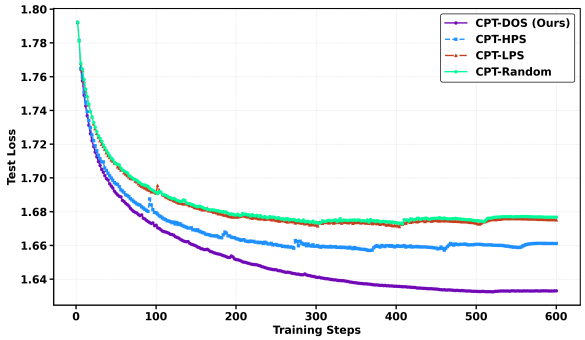


Figure 6: Test loss curves during CPT. CPT with DOS converges more rapidly at a lower loss.

Table 2: Ablation Study on Deepseek-V3-Base.

Benchmark	Base	CPT				
		RS	QBS	LPS	HPS	DOS
DiagnosisArena	50.20	58.77	57.20	48.90	51.70	62.70
GPQA-Med	63.16	68.42	68.42	68.42	57.89	63.16
PubMedQA	74.40	81.00	81.00	76.88	76.00	80.60
MedBullets_4	71.43	69.48	70.45	73.70	73.38	72.08
NEJMQA	75.11	75.00	74.89	72.99	73.92	75.11
MedMCQA	76.86	80.22	80.87	76.67	76.43	84.94
MedQA	83.03	86.80	84.84	85.00	84.29	88.22
Average	70.59	74.24	73.95	71.79	70.52	**75.26**

not only accelerates convergence but also delivers a substantial performance improvement.

Benchmark Performance Table 1 evaluates the impact of our perplexity-aware data scaling law on continual pre-training. On medical benchmarks, DOS-Base achieves the best overall performance, reaching an average score of 72.48, compared to 71.34 for RS-CPT and 71.22 for LPS-CPT. This corresponds to a 3.16 improvement over the base model. The gains are consistent across datasets: DOS-Base attains the highest or second-highest score on almost all medical tasks, with particularly notable improvements on DiagnosisArena and GPQA-Med. On general-domain benchmarks, DOS-Base maintains the par-level performance (84.16), matching the base model and slightly exceeding CPT-Base (83.94), indicating that the DOS could well reduce the forgetting of general knowledge. These results demonstrate that guiding CPT with our perplexity-aware scaling law is more effective than naive CPT: it converts additional training data into substantial, domain-specific gains while preserving strong general performance.

Data Distribution Figure 7 presents a t-SNE projection of sentence embeddings, where each point is colored by its perplexity range. Samples with medium PPL values (roughly the middle bins) are more diffusely scattered and less concentrated in distinct clusters. LPS-CPT or HPS-CPT tends to

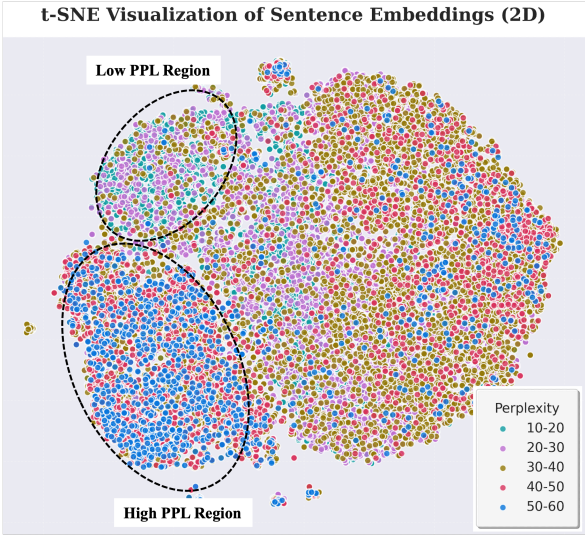


Figure 7: t-SNE visualization for varying perplexity ranges: low-PPL region and a high-PPL region.

focus on extreme-confidence outputs, while underrepresenting moderate-difficulty cases, which are often important for balanced generalization. Notably, applying the DOS strategy improves the coverage of the embedding space by explicitly controlling variance during data selection. As a result, DOS produces samples that span both low and high PPL regions, increasing overall data diversity and encouraging exploration of different linguistic patterns reflected in the two extremes. The distribution remains quality-preserving and balanced, suggesting that DOS can avoid bias toward either easy or overly difficult examples while still expanding the effective coverage of the data manifold.

Varying Foundation Models Table 2 evaluates the performance of our method for continually pretraining Deepseek-V3-Base. On medical benchmarks, DOS-CPT achieves the best overall performance with average score 75.26, compared to 74.24 for RS-CPT and 71.79 for LPS-CPT. Our method also outperforms traditional QBS-CPT. Thus, the gains for DOS are consistent across different datasets and foundation models.

General Corpus Table 3 presents the ablation study results on general domain benchmarks (CEVAL, CMMLU, and MMLU). Overall, the proposed DOS-CPT demonstrates competitive performance relative to other strategies, emerging as the most effective approach. Specifically, DOS-CPT achieves the highest average score of 84.43, surpassing the baseline by 0.27 points. This improvement is consistent across all three datasets, yielding notable gains on CEVAL (+0.16) and CMMLU (+0.17), which suggests that the DOS strategy effectively enhances the model's general knowledge capabilities with less catastrophic forgetting.

Table 3: Ablation Study on General Corpus.

Benchmark	Base	CPT				
		RS	QBS	LPS	HPS	DOS
CEVAL	85.52	85.35	84.86	85.54	85.52	85.68
CMMLU	84.92	84.72	84.77	84.95	84.32	85.09
MMLU	82.03	82.37	82.38	82.38	82.23	82.52
Average	84.16	84.14	84.03	84.29	84.02	**84.43**

Table 4: Performance Comparison under Different Token Budgets (Average Score on Medical Domain).

Budget	CPT				
	RS	QBS	LPS	HPS	DOS
10B	70.12	69.35	69.88	70.05	71.15
50B	70.89	69.78	70.75	70.21	71.92
100B	71.34	69.95	71.22	70.56	72.48

Varying Token Budgets Table 4 illustrates the scaling behavior across varying token budgets within the medical domain. A clear positive correlation emerges between the training budget and performance for all methods, confirming that increased data exposure generally enhances domain-specific proficiency. However, the scaling behavior varies significantly across different sampling strategy. DOS-CPT method consistently outperforms all baselines across every budget scale, achieving a peak score of 72.48 at 100B tokens.

5 Conclusion

We rethink the prevailing assumption that CPT performance scales monotonically with dataset size, demonstrating that naive data scaling yields diminishing returns. To address this inefficiency, we introduce a perplexity-aware data scaling law that exploits a model's initial perplexity landscape over domain corpora as a proxy for knowledge gaps. By modeling the relationship between perplexity statistics and model performance, our method adaptively selects high-utility training subsets while discarding redundant or noisy examples. Empirical results on medical and general-domain benchmarks show that this approach consistently identifies near-optimal data subsets, achieving superior convergence performance with significantly less data than conventional CPT.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (62502435), the Zhejiang Provincial Natural Science Foundation (LQN26F020002), the Open Research Fund of State Key Laboratory of Internet Architecture under Grant HLW2025ZD16, and Ant Group through CCF-Ant Research Fund.

Limitations

Our proposed perplexity-aware data scaling law introduces a principled approach to data selection by leveraging the model's own perplexity on domain-specific text as a proxy for knowledge gap estimation. By modeling the relationship between perplexity landscapes and downstream performance, our method enables the identification of high-utility training samples that maximize knowledge absorption while minimizing redundant or noisy data. However, the effectiveness of this approach depends on the pre-trained model's calibration and sensitivity to domain-specific patterns. Additionally, the current formulation operates at the token or sequence level and does not explicitly account for broader document structure, topic distribution, or semantic diversity, which can influence the true informational value of training data.

References

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, and Mansheej Paul. 2024. Perplexed by perplexity: Perplexity-based data pruning with small reference models. *arXiv preprint arXiv:2405.20541*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Francesco Cagnetta, Hyunmo Kang, and Matthieu Wyart. 2025. Learning curves theory for hierarchically compositional data with power-law distributed features. *arXiv preprint arXiv:2505.07067*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

Jie Chen, Zhipeng Chen, Jiapeng Wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Wayne Xin Zhao, Zhicheng Dou, Jiaxin Mao, and 1 others. 2025b. Towards effective and efficient continual pre-training of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5779–5795.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. 2024. Dsdm: Model-aware dataset selection with datamodels. *arXiv preprint arXiv:2401.12926*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Uriel Katz, Eran Cohen, Eliya Shachar, Jonathan Somer, Adam Fink, Eli Morse, Beki Shreiber, and Ido Wolf. 2024. Gpt versus resident physicians—a benchmark based on official board scores. *Nejm Ai*, 1(5):AIdbp2300192.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, and 1 others. 2024. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *Advances in Neural Information Processing Systems*, 37:90318–90354.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Xingjin Wang, Howe Tissue, Lu Wang, Linjing Li, and Daniel Dajun Zeng. 2025. Learning dynamics in continual pre-training for large language models. *arXiv preprint arXiv:2505.07796*.

Sondre Wold, Petter Mæhlum, and Oddbjørn Hove. 2024. Estimating lexical complexity from document-level distributions. *arXiv preprint arXiv:2404.01196*.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, and 1 others. 2025. From selection to generation: A survey of llm-based active learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14552–14569.

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

Xuemiao Zhang, Feiyu Duan, Liangyu Xu, Yongwei Zhou, Sirui Wang, Rongxiang Weng, Jingang Wang, and Xunliang Cai. 2025. Frame: Boosting llms with a four-quadrant multi-stage pretraining strategy. *arXiv preprint arXiv:2502.05551*.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. Diagnosisarena: Benchmarking diagnostic reasoning for large language models. *arXiv preprint arXiv:2505.14107*.

George Kingsley Zipf. 2013. *The psycho-biology of language: An introduction to dynamic philology*. Routledge.

6 Appendices

6.1 Perplexity-Aware Data Scaling Law

Why Perplexity Distributions Follow Power-Law Forms in Scaling Laws? The adherence of perplexity distribution statistics (*i.e.*, specifically mean (μ) and variance (σ)) to power-law forms in scaling laws is rooted in the hierarchical structure of language data.

Power-law behavior arises from language data's inherent hierarchy. Natural language data exhibits scale-invariant hierarchical structure: linguistic units (tokens, phrases, documents, domains) follow a power-law distribution in their frequency and complexity. For example, Zipf's law (Zipf, 2013) describes how word frequencies scale as $f \propto r^{-s}$ (where r is rank and $s \approx 1$), and document complexity (measured by syntactic depth or semantic ambiguity) similarly follows a power-law, where a small fraction of "high-complexity" documents drive most variation in PPL (Wold et al., 2024). This hierarchy directly shapes μ and σ:

- Mean μ: As incremental data volume (D_{new}) scales, μ converges to a domain-specific limit (μ_0) because larger datasets increasingly sample the full range of linguistic complexity. The convergence rate follows a power law ($\mu - \mu_0 \propto D_{\text{new}}^{-\alpha}$, $\alpha > 0$) because the remaining "unseen" complexity (driving deviations from μ_0) is dominated by low-frequency, high-complexity data, whose contribution decays as a power of D_{new} (consistent with Zipfian sampling) (Cagnetta et al., 2025).

- Variance σ: Variance quantifies diversity in data quality/complexity, which is inherently tied to the number of distinct subdomains (C) in D_{new}. Each subdomain contributes a unique PPL sub-distribution. The total variance scales as $\sigma \propto C^{\gamma}$ ($\gamma > 0$), *i.e.*, a power law because subdomain complexity itself follows a power-law hierarchy.

Why Consider Interaction Term in Scaling Law? In order to establish the scaling law between performance loss (PPL) and data distribution parameters, we first consider the following baseline power-law formulation:

$$L(\mu, \sigma, D) \triangleq E + \frac{D_c}{(\mu^{\alpha_0} * \sigma^{\beta_0}) * D^{\alpha_D}}. \quad (9)$$

However, this model assumes a monotonic relationship between the loss and each variable, while experimental results indicate that the relationship between the loss and the mean μ or standard deviation σ is not purely monotonic. As shown in Figure 1, the real samples (black points) are not strictly fall in the formula we fitted (red area). Furthermore, since both μ and σ are derived from the same PPL distribution, an intrinsic correlation between them may exist. Therefore, it is necessary to introduce interaction terms into the model to better capture their interdependent effects.

When extending the model, we adhere to the following principles: preserve the original power-law structure; keep the interaction form as simple as possible to avoid over-parameterization; and ensure the model can revert to the original form if the interaction effect is insignificant. Accordingly, we introduce a linear interaction term related to μ into the exponent of σ, yielding:

$$L(\mu, \sigma, D) \triangleq E + \frac{D_c}{(\mu^{\alpha_0} * \sigma^{\beta_0}) * (\sigma^{\beta_1 \mu}) * D^{\alpha_D}}. \quad (10)$$

Under this specification, if the fitted value of β_1 is zero, the model automatically reduces to the baseline formulation (Eq. 9), indicating no significant interaction between μ and σ. To further examine the symmetry of the interaction structure, we also consider introducing a linear term dependent on σ into the exponent of μ, leading to a symmetric formulation:

$$L(\mu, \sigma, D) \triangleq E + \frac{D_c}{\underbrace{(\mu^{\alpha_0} * \sigma^{\beta_0})}_{Independence} * \underbrace{(\mu^{\alpha_1 \sigma} * \sigma^{\beta_1 \mu})}_{Interdependence} * D^{\alpha_D}}. \quad (11)$$

Empirical fitting results show that the parameter α_1 is extremely small (typically less than 5×10^{-9}) and statistically insignificant, thus it can be neglected. Ultimately, we retain β_1 as the sole interaction parameter, which satisfies the modeling requirements while adhering to the principle of parsimony.

In summary, by incorporating a linear interaction term, we are able to flexibly capture the potential relationship between μ and σ while maintaining the power-law scaling property. This extended model naturally reduces to the baseline model when the interaction effect is negligible, demonstrating its soundness and robustness.

6.2 Mean/Variance as Sufficient Descriptors

Theoretical View Our choice of mean and variance as primary descriptors is not arbitrary but de-

Table 5: Comparisons to Alternative Descriptors.

Descriptor	Explanation	Relationship	Failure case	Strategy
Mean/Variance	Captures both data difficulty and diversity	Sufficient for most distributions	Suitable for most distributions	DOS (Ours)
Quantiles	Robust to outliers, pinpoints specific percentiles	Approximated by $\mu + z_p\sigma$ for near-Gaussian data	Drops more useful information over plateaued PPL distribution	QBS
Tail Mass	Focuses on extreme values (rare events)	Controlled by variance σ^2	Introduces data noise from tail of data distribution	LPS/HPS

rives from the exponential family of distributions, which includes Gaussian, Poisson, and Gamma.

By the Factorization Theorem (Fisher-Neyman), for any distribution in the exponential family with density:

$$f(x \mid \theta) = h(x) \exp\big(\eta(\theta)^T T(x) - A(\theta)\big)\,, \quad (12)$$

the statistic $T(X)$ is sufficient for θ. For the Gaussian distribution specifically: The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sufficient for μ. The pair $\big(\bar{X}, S^2\big)$ where $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is jointly sufficient for (μ, σ^2).

Quantiles or tail measures does not provide additional information about (μ, σ^2) given these sufficient statistics. This indicates mean and variance exhaustively capture all parameter-relevant information in the data, a property that quantiles fundamentally lack.

Relationship to Alternative Descriptors Quantiles and tail mass are useful alternatives, especially when dealing with heavy-tailed or skewed data. Our framework treats them as special cases that can be derived from mean and variance. The comparisons are shown in Table 5.

For bell-shaped (Gaussian-like) distributions, any percentile can be computed directly:

$$Q_p = \mu + z_p\sigma.$$

Instead of estimating quantiles separately, we use mean-variance interaction terms to account for non-Gaussian quirks like skewness. Extreme value theory indicates that tail heaviness is primarily a variance phenomenon (with corrections from higher moments), reinforcing our use of variance and interaction terms as sufficient descriptors.

6.3 Knowledge Gap and Informativeness

Our framework builds on the established observation that perplexity on pre-training data correlates with final task performance.

Perplexity and Knowledge Gap In-Domain perplexity (PPL_{CPT}) measures how well the model predicts domain-specific text of CPT. Lower values indicates the model already possesses relevant domain knowledge. Here, we define the knowledge gap using the perplexity differential:

$$\Delta PPL = PPL_{\mathrm{CPT}} - PPL_{\mathrm{PT}} \quad (13)$$

where PPL_{CPT} is perplexity on target CPT domain text and PPL_{PT} is perplexity on source pre-training domain text. This differential quantifies the relative unfamiliarity of the model with the target domain. Positive values indicate the domain contains patterns not well-captured by the model's current parameters.

Perplexity and Informativeness Documents with perplexity values in specific ranges (typically moderate-to-high relative to the model's current capability) yield greater performance gains when included in training. Perplexity is defined as:

$$PPL(x) = \exp\big(H(P_{\mathrm{data}}, P_{\mathrm{model}})\big), \quad (14)$$

which is the exponential of cross-entropy. It measures the compression inefficiency of the model on the data. Higher perplexity denotes that more bits needed to encode the sequence, which implies the data contains statistical patterns absent from the model's current representation.