ADIS-LAW: Unified Scaling Laws for Annealing-Phase Domain Injection in Large Language Models

Lyuxin Xue Hui Cai Xiaoyun Feng Xuanwei Hu Xin Zhang*
Ant International
{xuelyuxin.xlx, biyu.ch, fengxiaoyun.fxy,
huxuanwei.hxw, evan.zx}@ant-intl.com

Abstract

Continual pre-training (CPT) enables injecting domain-specific knowledge into large language models during the learning rate annealing phase. However, this compute-efficient paradigm forces practitioners to navigate a challenging optimization space: simultaneously choosing the injection fraction λ, the data mixture ratio r, and the re-warmup peak learning rate η. Because existing scaling laws have yet to comprehensively integrate and couple these interacting variables, they cannot reliably forecast the resulting trade-offs and final convergence outcomes across unseen injection configurations. To bridge this gap, we introduce the Annealing-phase Domain Injection Scaling Law (ADIS-LAW), a parametric framework that explicitly models late-stage distribution shifts as physical perturbations applied to the base pre-training trajectory. ADIS-LAW formalizes two opposing macroscopic behaviors: general-domain performance follows a linear forgetting-recovery process, whereas target-domain adaptation adheres to a power-law adaptation trend. Across data-budget and model-scale extrapolation, ADIS-LAW predicts target-domain trajectories with an average Spearman correlation of 0.951 and forecasts tail losses with relative error below 5%. We further show that ADIS-LAW can predict optimal injection strategies at extrapolated scales: it ranks target-scale candidate policies with a mean Spearman correlation of 0.936. By replacing blind grid search over (λ, r, η) with a single fitted law, ADIS-LAW offers a data-driven approach for strategy selection in annealing-phase domain injection.

1 Introduction

Continual pre-training (CPT) on a domain-specific corpus is conventionally applied to fully pre-trained large language models (LLMs) [1]. An alternative, compute-efficient paradigm is to inject domain data earlier—specifically during the *annealing phase* of the initial pre-training run, where the learning rate decays rapidly toward zero. While this avoids the overhead of launching an independent post-training phase, the dynamics of domain injection during a rapidly decaying schedule remain largely unexplored. Consequently, practitioners lack principled guidance: this regime is governed by at least three strongly coupled hyperparameters: the injection fraction λ (the fraction of the total training budget allocated to CPT, dictating when injection begins), the data mixture ratio r, and the re-warmup learning rate η (which has been shown to be critical for overcoming optimization stagnation during continual pre-training [2]). Without theoretical models, these hyperparameters are usually selected by grid search at small scale and copied over to larger runs, which is expensive and provides no guarantee that the chosen configuration remains optimal at the target scale.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

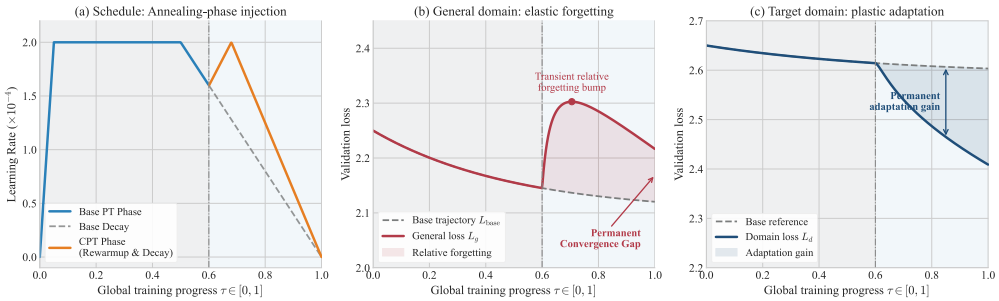

Figure 1: **Three-panel schematic illustrating learning schedules and loss perturbations.** Both the baseline and CPT trajectories show an overall decrease in loss as training progresses. **Background colors indicate different training stages: light gray for the pre-training (PT) phase and light blue for the continual pre-training (CPT) phase.** (a) A typical Learning Rate schedule, where CPT triggers a re-warmup η during the annealing phase at global progress $(1 - \lambda)$. (b) The general loss L_g operates under a forgetting-recovery mechanism. It suffers an initial deficit relative to the baseline governed by the perturbation $\phi(aI - b\tau)$, followed by a partial recovery. Because the training budget ends at global progress $\tau = 1$, a permanent Convergence Gap remains. (c) The target-domain loss L_d exhibits plastic adaptation, monotonically diverging from its respective base reference to achieve a permanent performance gain via a saturating $r^{\mu}\tau^{\nu}$ power law.

A predictive scaling law would remove the need for this grid search by directly forecasting the loss trajectory of an unseen injection strategy at the target scale. Classical pre-training scaling laws [3, 4] strictly govern static base trajectories and cannot account for late-stage distribution shifts. While recent pioneering efforts have attempted to model continual pre-training dynamics [5, 6], they suffer from critical limitations when applied to the annealing regime. These frameworks often assume fixed learning rate schedules and rigid injection timings, or they were not designed to explicitly model the interplay between the data mixture ratio r and the re-warmup peak learning rate η. Since loss dynamics is driven by the interaction of all three variables (λ, r, η), existing models do not naturally extend to provide a single closed-form prediction across varying (λ, r, η).

To bridge this gap, we introduce the **Annealing-phase Domain Injection Scaling Law** (ADIS-LAW), a parametric framework that explicitly models late-stage distribution shifts not as isolated training segments, but as physical perturbations applied to the base pre-training trajectory. Through this framework, we formalize two opposing macroscopic behaviors during domain injection: general-domain performance follows a *linear forgetting-recovery process* primarily governed by injection intensity and duration, whereas target-domain knowledge absorption adheres to a *power-law adaptation trend* constrained by diminishing marginal returns. Crucially, ADIS-LAW accurately forecasts loss dynamics and policy rankings of annealing-phase domain injection, and also generalizes to post-annealing injection setups.

Our contributions are summarized as follows:

- **Parametric Framework for Domain Injection:** We propose ADIS-LAW, the first phenomenological scaling law designed to model annealing-phase domain injection as a trajectory perturbation. By explicitly decoupling the base pre-training dynamics from the distribution shift (governed by injection intensity $I = r\frac{\eta}{\eta_0}$ and a generalized saturating gate ϕ), our formulation provides a consistent prediction across arbitrarily coupled parameters without hyperparameter-specific refitting.
- **Zero-Shot Forecasting and Cross-Scale Extrapolation:** ADIS-LAW demonstrates remarkable predictive robustness. Across a diverse space of unseen injection configurations (λ, r, η), it forecasts full learning curves with both high rank correlations ($\rho > 0.90$) and remarkably low relative errors ($\varepsilon_{\text{rel}} < 5\%$), accurately capturing the overarching perturbation behaviors and the absolute magnitude of the final tail loss. Furthermore, it reliably extrapolates from small proxy runs to macro-scale budgets (from 25B to 100B tokens) and larger models (from 115M to 1.67B parameters), predicting the crucial end-of-training tail loss with relative errors (ε_{rel}) below 3%.
- **Decision-Oriented Policy Selection:** Moving beyond evaluating models purely as curve predictors, we establish ADIS-LAW as a practical policy selector. By fitting the law at a cheap scale, it accurately forecasts the target-scale Pareto landscape, ranking hundreds of candidate injection

policies with Spearman correlations up to 0.99 and replacing costly empirical grid search with a predictive framework.

2 Related work

Scaling Laws and Learning Rate Dynamics. Scaling laws characterize performance across model sizes, data budgets, architectures, and training regimes [3, 4, 7–12]. Learning rate (LR) schedules and batch size also shape training dynamics [13–15]. The LR-aware law of Tissue et al. [16], also known as Momentum Law, models full loss trajectories through cumulative LR and annealing effects and provides our base formulation. Multi-Power Law refines loss curve prediction across LR schedules [17]. NCPL studies configuration-conditioned predictors supporting both final loss and loss curve prediction [18]. Together, these works broaden scaling law prediction beyond model size and token count to account for training configurations and optimization dynamics.

Data Mixtures and Scheduling. BiMix jointly models the effects of domain proportions and data volume on pretraining loss [19]. CAMEL captures interactions between model capacity and mixture composition for data optimization across model scales [20], while Dai and Zheng [21] explain mixture-dependent losses through capacity competition and noise reduction. These works address how data mixture and scheduling shape training outcomes, but do not explicitly model how injection timing, mixture ratio, and re-warmup peak LR jointly shape both general and target domain loss trajectories during annealing. This joint trajectory prediction problem is the central focus of ADIS-LAW.

Continual Pretraining. Continual Pretraining (CPT) adapts pretrained models to domain-specific corpora [1, 22]. The domain-specialization literature includes language models for finance [23, 24], code [25], and biomedicine [26]. Adaptation can degrade general-domain performance [27, 28], motivating studies of representation changes [29] and mitigation through regularization and replay [30, 31]. Our focus is on predicting the observable general and target domain loss trajectories induced by domain injection during annealing.

Predictive Laws for CPT. D-CPT predicts general and domain-specific endpoint losses as functions of model size, data size, and mixture ratio [5]. Liew and Kato [32] model how the benefits of second-stage training depend on the amount of prior pretraining, capturing saturation when reusing overtrained models. At the trajectory level, CPT Law [6] combines an LR-dependent base trajectory with a saturating distribution shift term and includes extensions for model size and replay ratio. ADIS-LAW jointly conditions general and domain predictions on model scale, injection timing λ, mixture ratio r, and re-warmup LR η, with the aim of selecting policies at larger model and data scales.

3 Methodology

We formulate the loss trajectory of a domain-injection run (whether during the annealing phase or as a traditional post-training step) as the sum of a base trajectory L_{base} and a bounded perturbation gated by a saturating activation. We first state our notation and modelling assumptions, then write down the proposed scaling law, and finally interpret its qualitative behaviour. We treat the law as phenomenological: it is motivated by an elasto-plastic analogy and a small set of modelling assumptions, and is justified empirically in Section 4.

3.1 Setup and Notation

We consider a CPT run defined by: model parameter count N; total token budget D for the pre-training schedule; learning-rate (LR) schedule η_t over training steps $t = 1, \ldots, T$; injection fraction $\lambda \in (0, 1]$, defined as the fraction of the total training budget assigned to the domain-injection CPT phase; mixture ratio $r \in (0, 1]$ of domain-vs.-general data after injection; and re-warmup peak learning rate η at the moment of injection (relative to a reference η_0). We use cumulative learning

quantities to characterize training progress, following Tissue et al. [16]:

$$
\begin{aligned}
S_1 &= \sum_{i=1}^{t} \eta_i, \quad S_1^{\text{cpt}} = \sum_{i=(1-\lambda)T}^{t} \eta_i, \\
S_2^{\text{pt}} &= \sum_{i=1}^{\min(t,(1-\lambda)T)} \sum_{k=1}^{i} (\eta_{k-1} - \eta_k)\gamma^{i-k}, \\
S_2^{\text{cpt}} &= \sum_{i=(1-\lambda)T}^{t} \sum_{k=(1-\lambda)T}^{i} (\eta_{k-1} - \eta_k)\gamma^{i-k}.
\end{aligned} \tag{1}
$$

Here S_2^{pt} and S_2^{cpt} separate the contributions from the pre-injection and post-injection portions of the schedule using an exponentially weighted moving average kernel with a decay factor γ (typically set to 0.999), consistent with Tissue et al. [16]. We further define a global normalized training progress variable $\tau = t/T \in [0, 1]$, which includes both the base pre-training portion and the post-injection CPT portion. Since λ is the injection fraction, injection begins at global progress $\tau = 1 - \lambda$, and training ends at $\tau = 1$. In the following, we show that the post-injection quantity S_1^{cpt} is responsible for switching on the distribution shift, while τ tracks where that shift occurs within the full training trajectory.

3.2 Quantifying Distribution Shift in Domain Injection

When injecting domain data during the annealing phase, the underlying pre-training distribution shifts drastically, perturbing the system from its predictable base trajectory. Following Tissue et al. [16], we independently model the unperturbed trajectory L_{base} using the cumulative learning rate:

$$
L_{\text{base}}(N, S_1, S_2^{\text{pt}}, S_2^{\text{cpt}}) \;=\; \frac{A}{N^{\alpha}} + \frac{B}{S_1^{\beta}} - C_1\, S_2^{\text{pt}} - C_2\, S_2^{\text{cpt}} + L_{\infty}. \tag{2}
$$

Prior CPT scaling law work does not fully couple the interacting drivers of this shift: when the mixture ratio r, injection fraction λ, and active learning rate η change together, their transient effect cannot be reduced to a static offset. We therefore propose that the actual impact force of domain data depends synergistically on both the sheer ratio of domain data $r \in (0, 1]$ and the prevailing energy level of the target LLM—its active learning rate η. We define this universal perturbation strength as the *Injection Intensity* I:

$$
I \;=\; r \cdot \frac{\eta}{\eta_0}, \tag{3}
$$

normalized to a reference learning rate η_0. The system's response to this intensity is neither instantaneous nor purely linear. Drawing from activation phenomenologies, the trajectory perturbation requires a cumulative warmup duration to fully saturate. We model this via a *generalized saturating gate* ϕ:

$$
\phi(S_1^{\text{cpt}}) \;=\; 1 - \left(1 + E \cdot S_1^{\text{cpt}}\right)^{-\beta_\phi}, \tag{4}
$$

where S_1^{cpt} represents the cumulative learning momentum acquired exclusively since the domain-injection moment. The factor $\phi \in [0, 1]$ provides a continuous bridging operator: at $S_1^{\text{cpt}} = 0$, $\phi = 0$ producing no shift; as the cumulative learning rate $S_1^{\text{cpt}} \to \infty$, the shift becomes fully active ($\phi \to 1$). The coefficient $E > 0$ sets the LR-mass scale at which the shift becomes visible, while $\beta_\phi > 0$ controls the curvature of the saturation (note that when $\beta_\phi = 1$, it degenerates into a Michaelis–Menten-like normalized form $x/(K + x)$); the subscript distinguishes it from the base-trajectory exponent β in Eq. (2).

3.3 Linear Forgetting Dynamics of General Knowledge

When integrating the activation multiplier ϕ, we must account for the primary destructive cost to the model's foundational base: catastrophic forgetting of general capabilities. Existing hypotheses conjecture a static degradation or assume an insurmountable dependency between forgetting and specific model capacities. Instead, we mathematically formalize the deviation of general loss, ΔL_g,

as a competitive linear dynamic between the destructive power of the injection intensity ($a \cdot I$) and the temporal elastic self-recovery of the model acting across global training progress τ:

$$L_g(N, S_1, S_2^{\text{pt}}, S_2^{\text{cpt}}, \tau, S_1^{\text{cpt}}, I) \;=\; L_{\text{base}} \;+\; \phi(S_1^{\text{cpt}}) \cdot \Big(\underbrace{a \cdot I}_{\text{Destruction}} \;-\; \underbrace{b \cdot \tau}_{\text{Recovery}} \Big). \tag{5}$$

Through the rigorous decoupling in Equation 5, ADIS-LAW demonstrates that general domain loss evolution can be structurally decoupled from the base trajectory, rendering our phenomenological parameters transferable across different model scales.

This formulation yields a non-monotonic L_g after injection: near $1 - \lambda$, the destruction term $\phi \cdot aI$ dominates, spiking the loss. At larger global progress τ, the recovery term $-\phi \cdot b\tau$ overtakes it. The observed peak of the full loss occurs where the perturbation's upward slope exactly balances the base trajectory's descent (detailed in Appendix B).

Beyond the initial disruption peak, the recovery term $-\phi \cdot b\tau$ gradually reduces the general loss perturbation. However, because the strictly bounded training budget terminates optimization precisely at $\tau = 1$, the compute budget is exhausted before the model can fully reconverge to its unperturbed baseline. Evaluating this deficit at the end of training yields a permanent *Convergence Gap* ΔL_{gap}. This budget-limited decomposition simplifies the learning dynamics into a lightweight, predictive relationship enabling ADIS-LAW to accurately forecast the trade-offs of late-stage domain injection.

3.4 Power-Law Scaling in Domain Adaptation

Conversely, the model does not acquire specialized domain knowledge in a linear fashion. Large language models absorbing vertical domain concepts exhibit *diminishing marginal returns*; doubling the domain ratio r or extending the normalized time τ results computationally in diminishing efficiency returns. Unlike the linear interplay of the general loss, we map the trajectory of the newly injected domain loss L_d via a power-law adaptation scaling format:

$$L_d(N, S_1, S_2^{\text{pt}}, S_2^{\text{cpt}}, \tau, S_1^{\text{cpt}}, r) \;=\; L_{\text{base}} \;-\; \phi(S_1^{\text{cpt}}) \cdot \Psi \cdot r^{\mu} \cdot \tau^{\nu}, \qquad \text{where } \mu, \nu \in (0, 1]. \tag{6}$$

Here, the exponentiated terms μ and ν constrain the maximum rate of continuous knowledge absorption. Our integrated ADIS-LAW formulation streamlines the prior practice of fitting separate parameters for distinct combinations of domain ratio and learning rates, providing a single, generalized equation capable of modeling diverse annealing-phase dataset compositions.

Equation 6 is monotonically decreasing and implies that additional post-injection training progress yields diminishing returns on the domain loss. We motivate this phenomenological form by a strain-hardening analogy: the gradient field becomes locally stiffer as more domain mass is absorbed, shrinking the marginal loss drop. We present a stylized ODE consistent with this picture in Appendix C.

3.5 Cross-Scale Transferability and Boundary Limits

Scale transferability. The perturbation ΔL does not contain N explicitly in a multiplicative factor. Empirically we find that most fitted parameters $\theta = \{a, b, \Psi, \mu, \nu, E, \beta_\phi\}$ transfer robustly across $N \in \{115\text{M}, 366\text{M}, 895\text{M}, 1.67\text{B}\}$, though the ratio a/b systematically decays as model capacity grows. Nevertheless, this parameter transferability ensures that fitting at small scales can zero-shot predict tail outcomes at larger scale. Furthermore, Equation 5 does not utilize an explicit scale dilution term $1/N^{\gamma}$; we validate that this factor remains empirically unnecessary within our tested scale range in Section 4.5.

Limits and failure modes. We expect ADIS-LAW to be highly accurate when the perturbation is bounded relative to the base trajectory ($\phi \cdot aI$ small compared with L_{base}), the mixture ratio r stays in the moderate range explored at fitting time, and the injection target is broadly in-distribution with the pre-training corpus. ADIS-LAW may fail under: (i) extreme intensities $aI \gg b$, where the linear forgetting term predicts an unbounded peak; (ii) late-decay regimes where S_1^{cpt} is too small for ϕ to saturate before training ends; or (iii) heavily out-of-distribution target corpora.

4 Experiments

We design our empirical evaluation to rigorously test ADIS-LAW moving from full-trajectory generalization through decision-making optimization, culminating in mechanistic and structural stress tests:

1. **Generalization via leave-one-out cross-validation(LOO-CV)**: Can a single parameterization predict continuous loss contours for unseen injection policies across diverse model and data scales?
2. **Cross-Scale Extrapolation**: Can the law, fitted on proxy regimes, reliably forecast complete training trajectories and terminal losses at target macro-scales?
3. **Decision-Oriented Strategy Selection**: Can these proxy predictors act as optimal policy selectors—successfully ranking pairwise choices and revealing Pareto frontiers?
4. **Mechanistic Ablations**: Which explicit architectural constraints (e.g., the saturation gate ϕ, dose intensity I) are structurally required to prevent predictive collapse?
5. **Extension to Post-Annealing Injection**: Does the underlying perturbation geometry naturally generalize to continual pre-training launched strictly *after* base annealing concludes?[2]

4.1 Experimental Setup

Models and Pre-training Framework. We pre-train decoder-only SmolLM3 architectures [33] across four parameter scales ($N \in \{115\text{M}, 366\text{M}, 895\text{M}, 1.67\text{B}\}$) using the FineWeb-Edu [34] base corpus. Training adheres to a Warmup-Stable-Decay (WSD) scheduling strategy across data budgets ranging from 25B to 100B tokens. Fully detailed hyperparameters and configurations are provided in Appendix D.

Domain Injection Configurations. We inject data from four orthogonal domains: Mathematics (InfiMM-WebMath-40B [35]), Code (StackEdu [36]), BioMedicine (Biomed-Enriched [37]) and Finance (constructed from internal financial corpus). During the decay phase, sweeping injection fractions (λ), mixture ratios (r), and re-warmup peak learning rates (η) yields hundreds of structural combinations per domain to rigorously stress-test the scaling law.

Evaluation Protocols and Metrics. The predictive power of ADIS-LAW is evaluated using four stringent held-out protocols. For in-distribution interpolation, we use data-budget and model-scale leave-one-out cross-validation (**D-LOO** and **M-LOO**). For zero-shot extrapolation, we use target-scale predictions trained strictly on cheaper proxy models: **D-Up** (predicting 100B token runs using ≤ 50B data) and **M-Up** (predicting 1.67B models using ≤ 895M model fits). We compare against the baseline CPT Scaling Law [6] utilizing two key metric dimensions: (i) **Spearman Rank Correlation (ρ):** Evaluates whether the predicted trajectory preserves the correct geometric ordering. (ii) **Relative Error (ε_{rel}):** Measures the absolute trajectory calibration percentage.[3]

4.2 Leave-One-Out Generalization

We evaluate generalization to unseen (λ, r, η) configurations using Data-budget (D-LOO) and Model-scale (M-LOO) leave-one-out protocols. As visually demonstrated in Figure 2, ADIS-LAW precisely tracks individual trajectory dynamics. Furthermore, the base formulation achieves near-perfect fidelity on the PT-phase loss ($\rho = 0.98$, RelErr $< 1.2\%$), confirming its robust capacity scaling without re-parameterization.

CPT-phase generalization. Table 1 summarizes CPT-phase held-out performance across M-LOO and D-LOO tests. Many existing scaling laws mentioned in Section 2 address only subsets of our prediction problem, which requires predicting both general and target domain loss trajectories across varying scales and (λ, r, η). Thus, rather than presenting non-equivalent numerical comparisons, we compare ADIS-LAW against the published **CPT Law** [6] and its ratio-aware extension (+ **Ratio**), two closely related CPT formulations that allow direct comparison.

[2]Due to space constraints, investigations into Post-Annealing Domain Injection (addressing Question 5) are fully detailed in Appendix I.

[3]We avoid R^2 because general loss fluctuates within exceptionally narrow quantitative bounds during continual pre-training; R^2 fails to objectively reflect fitting quality on microscopic amplitude scales [38, 39].

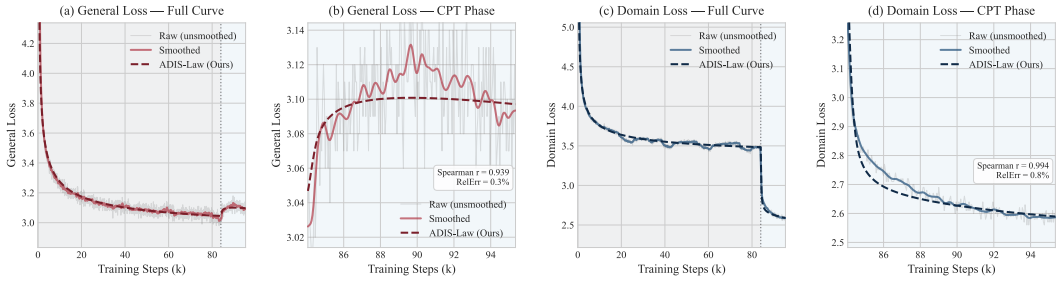

Figure 2: **ADIS-LAW fit on a representative experiment** (`BioMed`, $N = 115$M, $D = 50$B, $r = 0.80$, $\lambda = 0.12$). **Solid curves**: observed loss (Gaussian-smoothed, $\sigma = 2$); **dashed curves**: ADIS-LAW prediction. **Background colors** represent different training stages: light gray for the pre-training (PT) phase, and light blue for the continual pre-training (CPT) phase. **(a)** General loss (**red tones**), full training curve with PT→CPT transition marked. **(b)** General loss CPT-phase zoom, annotated with Spearman ρ and RelErr. **(c–d)** Corresponding panels for domain loss (**blue tones**). The tight overlay across both loss types and both training phases confirms that the functional form of Eqs. (5)–(6) captures the underlying loss dynamics at the individual-experiment level, complementing the aggregate LOO-CV metrics below.

Table 1: CPT-phase leave-one-out generalization. Each cell reports Spearman $\rho \uparrow$ and RelErr% $\downarrow$ averaged over 4 domains. D-LOO and M-LOO perform LOO-CV across data budgets at $N = 115$M and model scales at $D = 50$B, respectively.

	D-LOO		M-LOO	
Method	Gen. $\rho\uparrow$ / Err%$\downarrow$	Dom. $\rho\uparrow$ / Err%$\downarrow$	Gen. $\rho\uparrow$ / Err%$\downarrow$	Dom. $\rho\uparrow$ / Err%$\downarrow$
CPT Law	0.500 / 1.51%	0.545 / 47.63%	0.674 / 11.20%	0.785 / 48.50%
+ Ratio	0.524 / 1.76%	0.775 / 34.06%	–	–
ADIS-LAW (ours)	**0.913 / 0.56%**	**0.965 / 2.75%**	**0.924 / 1.78%**	**0.966 / 2.47%**

ADIS-LAW drastically outperforms existing baselines on both objectives. On general loss, it elevates Spearman ρ from 0.50–0.67 to over 0.91, driven by the linear forgetting-recovery mechanism term $\phi \cdot (aI - b\tau)$ that explicitly responds to varying injection intensities $I = r \cdot \eta / \eta_0$. On the target-domain loss, ADIS-LAW achieves $\rho \approx 0.96$ and limits relative error to under 3% (compared to $> 34\%$ for the + Ratio baseline). Unlike the baselines, our geometric formulation correctly models the diminishing marginal returns of domain knowledge absorption via the generalized saturation gate ϕ and the scalable power-law parameterization $r^{\mu}\tau^{\nu}$.

4.3 Cross-Scale Extrapolation and Pairwise Strategy Ranking

Whereas LOO-CV tests generalization at observed scales, a practical scaling law must forecast the loss trajectories of larger runs using only smaller, cheaper proxy runs. We evaluate this through two zero-shot extrapolation protocols: **D-Up** (fits on $D = 25$B/50B at $N = 115$M to predict 100B) and **M-Up** (fits on $N \leq 895$M at $D = 50$B to predict 1.67B). Neither target scale is seen during fitting.

Table 2 demonstrates that ADIS-LAW retains strong predictive accuracy across these vast inference gaps (extended visualizations are deferred to Appendix E). For D-Up, the law achieves an average Gen. $\rho = 0.913$ and Dom. $\rho = 0.945$, with a mere 2.78% average relative error in the target domain. For the M-Up protocol, Dom. ρ improves to 0.956. This confirms that the parameterization of the domain-adaptation terms in Eqs. (5)–(6) transfers smoothly across scales. We further assess fitting stability across optimizer initializations and report configuration-level bootstrap confidence intervals for D-Up and M-Up in Appendix E.1.

Pairwise Decision Forecasting. The ultimate goal in domain continual pre-training is to identify optimal hyperparameters without training them. We therefore evaluate ADIS-LAW as a *decision maker*. For any pair of unseen injection strategies, we use the predicted final loss to select the superior configuration. Figure 3 shows how predicted rankings closely mirror actual rankings. Quantitatively, as shown in Table 2, the law achieves up to 95.2% Pairwise Accuracy and 0.991 NDCG at selecting

Table 2: Cross-scale extrapolation and pairwise decision ranking (averaged across domains). Full-trajectory prediction captures the shape of the CPT phase; Pairwise Final-Loss relates to selecting the better injection strategy of the held-out $D = 100\text{B}, N = 1.67\text{B}$ setting. See Appendix E for expanded general-loss visualizations.

	Full-Trajectory Extrapolation				**Pairwise Final-Loss Rank**			
	D-Up		**M-Up**		**D-Up**		**M-Up**	
Metric Type	$\rho \uparrow$	**Err%** $\downarrow$	$\rho \uparrow$	**Err%** $\downarrow$	**Acc** $\uparrow$	**NDCG** $\uparrow$	**Acc** $\uparrow$	**NDCG** $\uparrow$
General Loss	0.913	0.73%	0.904	0.83%	0.857	0.961	0.810	0.954
Domain Loss	0.945	2.78%	0.956	4.31%	0.881	0.991	0.952	0.989

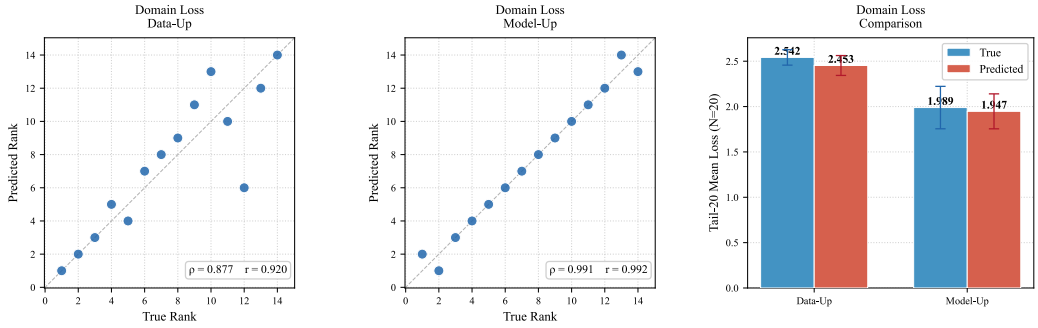

Figure 3: **Final-loss prediction under upward extrapolation. Scatter panels (left and center):** per-experiment true vs. predicted domain rank for D-Up (left) and M-Up (center). Points represent experiments from `Finance` and `BioMed`, annotated with Spearman ρ and Pearson r. **Bar panels (right):** mean tail-20 domain loss across all held-out strategies, with ± 1 std error bars. A comprehensive visualization including general loss predictions is provided in Appendix E.

strategies that achieve the lowest final target-domain loss, while accurately ranking strategies based on general knowledge preservation (Accuracy $\geq 81.0\%$, NDCG ≥ 0.954). By reliably gauging unseen strategy superiority, our framework can systematically extract optimal Pareto frontiers.

4.4 Finding the Optimal Domain Injection Strategy

While pairwise accuracy evaluates binary choices, a predictive law's ultimate utility lies in searching a dense strategy space for the Pareto-optimal frontier *before* executing target-scale runs. Using ADIS-LAW fitted exclusively on proxy scales, we scored a grid of 486 synthetic injection policies on `BioMed` against an *oracle* fitted directly on the target scale (100B for D-Up, 1.67B for M-Up).

Figure 4 demonstrates that the predicted global rankings tightly correlate with the target-scale oracle. Crucially, even where global rank correlation slightly degrades (e.g., General Loss under M-Up), the predicted *Pareto frontier* geometrically matches the oracle's frontier with a Chamfer distance strictly below 0.06 (Figure 5). The result shows that the proxy-fitted law has the capability to select near-perfect hyperparameter configurations (λ, r, η) at scale without prior training.

Actionable Hyperparameter Guidance and ε-constraint. To single out a strategy from the Pareto front, first we define ε as the tolerance degree of general-loss degradation ($L_g \leq L_g^{\min}(1+\varepsilon)$). We apply this ε-constraint scalarization to both frontiers ($\varepsilon \in [0.001, 0.20]$). Across all tolerances and scalarization scales, the proxy predictor identifies the most critical parameter vector—specifically selecting target-domain injection ratio $r \approx 0.90$ identically to the oracle. The mean discrepancies between proxy- and oracle-selected configurations are proved to be close. Detailed comparisons are provided in Appendix F. In short, ADIS-LAW directly supplies mathematically robust, production-ready continual pre-training configurations. To verify that these predictive capabilities reliably transfer across different domains, we additionally supply the corresponding Pareto analysis on the `Finance` domain in Appendix H.

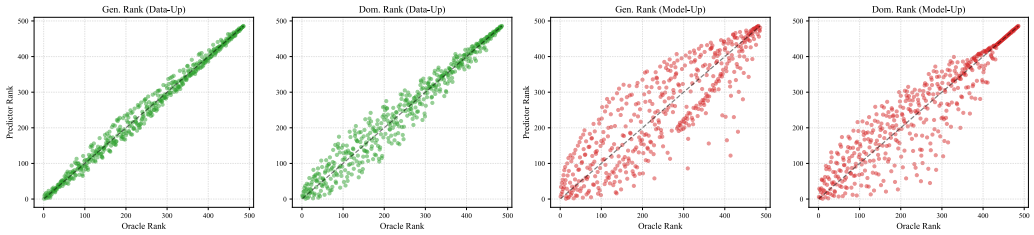

Figure 4: **Predicted vs. oracle rank across 486 synthetic policies on `BioMed`.** Points cluster strongly around the ideal diagonal under D-Up; M-Up general loss ranking exhibits expected scatter, but domain ranking remains tightly correlated.

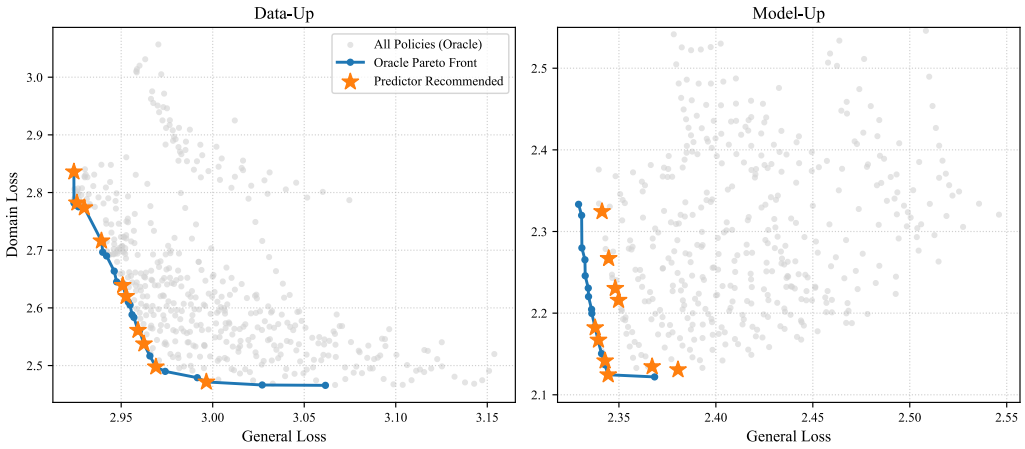

Figure 5: **Pareto frontier alignment.** The predictor-recommended strategies (orange stars) lie close to the true oracle frontier (blue line). This suggests that the predictor can identify strategies with near-optimal trade-offs between general and domain loss, as evaluated by the target-scale oracle.

4.5 Mechanistic Ablation: Scale Transferability

We conduct ablation studies to assess how individual components contribute to the predictive performance of ADIS-LAW. The necessity of the generalized saturating gate and dose-coupled absorption is detailed in Appendix G; here we focus on cross-scale transferability.

To test if the perturbation magnitude depends heavily on model size, we introduce a $1/N^{\gamma}$ multiplicative factor to both the forgetting and absorption terms (+ScaleDilution). The configuration fundamentally rejects the inclusion of parameter count N within our tested constraints: the fitted exponent γ converges to near zero (≈ 0.001). As shown in Table 3, this variant yields negligible metric changes ($|\Delta\rho| \leq 0.002$) relative to the base law.

Table 3: Scale-dilution ablation averaged across all domains and protocols.

Variant	**General Loss**			**Domain Loss**		
	$\rho\uparrow$	Err%$\downarrow$	$\Delta\rho$	$\rho\uparrow$	Err%$\downarrow$	$\Delta\rho$
Ours Full	0.911	0.85%	—	0.954	5.23%	—
+ScaleDilution	0.913	0.89%	+0.002	0.956	5.25%	+0.002

This confirms that an explicit multiplicative scale dilution $1/N^{\gamma}$ factor is unnecessary within the tested scale range. This empirical property represents a practical foundation of ADIS-LAW's cross-scale transferability—it explains why injection hyperparameters and strategy rankings optimized at surrogate scales (115M parameters) reliably preserve their ordering when applied to our larger target architectures (1.67B parameters).

5 Conclusion and limitations

We presented ADIS-LAW, a phenomenological scaling law modeling the loss trajectory of an LLM under annealing-phase domain injection. Using a base trajectory and a bounded, generalized saturating gate perturbation parameterized by an injection intensity ($I = r\frac{\eta}{\eta_0}$), the law characterizes the asymmetric dynamics of adaptation: a linear forgetting-recovery mechanism for general capabilities, and a saturating power-law mechanism for domain acquisition. With seven parameters per domain, it extrapolates zero-shot to 100B-token budgets and 1.67B parameters, supporting policy ranking and target-scale parameter decision making.

Limitations and Future Work. ADIS-LAW is a phenomenological predictive law. Limitations include: (i) Evaluation covers models with up to 1.67B parameters and training budget up to 100B tokens; behavior at substantially larger scale, such as trillions of parameters or tokens, remains unexplored. (ii) The linear forgetting term aI assumes standard intensities; in extreme regimes ($aI \gg b$), an unbounded peak is predicted unless structurally capped. (iii) Underestimation of transient loss peaks: When extrapolating to larger scales, ADIS-LAW accurately predicts final loss but may underestimate the transient increase of general-domain loss following injection. The more fundamental principle underlying this transient bump requires further exploration.

Acknowledgments and Disclosure of Funding

This work is supported internally by Ant International. We have no additional competing interests to disclose.

References

[1] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, et al. Don't stop pretraining: Adapt language models to domains and tasks, 2020. URL https://arxiv.org/abs/2004.10964.

[2] Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, et al. Continual pre-training of large language models: How to (re)warm your model?, 2023. URL https://arxiv.org/abs/2308.04014.

[3] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

[4] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

[5] Haoran Que, Jiaheng Liu, Ge Zhang, et al. D-CPT law: Domain-specific continual pre-training scaling law for large language models, 2024. URL https://arxiv.org/abs/2406.01375.

[6] Xingjin Wang, Howe Tissue, Lu Wang, Linjing Li, and Daniel Dajun Zeng. Learning dynamics in continual pre-training for large language models, 2025. URL https://arxiv.org/abs/2505.07796.

[7] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers, 2022. URL https://arxiv.org/abs/2106.04560.

[8] Aidan Clark, Diego de las Casas, Aurelia Guy, et al. Unified scaling laws for routed language models, 2022. URL https://arxiv.org/abs/2202.01169.

[9] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer, 2021. URL https://arxiv.org/abs/2102.01293.

[10] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training, 2021. URL https://arxiv.org/abs/2110.02095.

[11] Niklas Muennighoff, Alexander M. Rush, Boaz Barak, et al. Scaling data-constrained language models, 2025. URL https://arxiv.org/abs/2305.16264.

[12] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws, 2023. URL https://arxiv.org/abs/2210.14891.

[13] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts, 2017. URL https://arxiv.org/abs/1608.03983.

[14] Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't decay the learning rate, increase the batch size, 2018. URL https://arxiv.org/abs/1711.00489.

[15] Shengding Hu, Yuge Tu, Xu Han, et al. MiniCPM: Unveiling the potential of small language models with scalable training strategies, 2024. URL https://arxiv.org/abs/2404.06395.

[16] Howe Tissue, Venus Wang, and Lu Wang. Scaling law with learning rate annealing, 2024. URL https://arxiv.org/abs/2408.11029.

[17] Kairong Luo, Haodong Wen, Shengding Hu, et al. A multi-power law for loss curve prediction across learning rate schedules, 2025. URL https://arxiv.org/abs/2503.12811.

[18] Huaqing Zhang, Kaiyue Wen, and Tengyu Ma. Configuration-to-performance scaling law with neural ansatz, 2026. URL https://arxiv.org/abs/2602.10300.

[19] Ce Ge, Zhijian Ma, Daoyuan Chen, Yaliang Li, and Bolin Ding. BiMix: A bivariate data mixing law for language model pretraining, 2025. URL https://arxiv.org/abs/2405.14908.

[20] Jingwei Li, Xinran Gu, and Jingzhao Zhang. Capacity-aware mixture law enables efficient LLM data optimization, 2026. URL https://arxiv.org/abs/2603.08022.

[21] Rui Dai and Shuran Zheng. Explaining data mixing scaling laws, 2026. URL https://arxiv.org/abs/2606.08167.

[22] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models, 2023. URL https://arxiv.org/abs/2302.03241.

[23] Shijie Wu, Ozan Irsoy, Steven Lu, et al. BloombergGPT: A large language model for finance, 2023. URL https://arxiv.org/abs/2303.17564.

[24] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models, 2025. URL https://arxiv.org/abs/2306.06031.

[25] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, et al. Code Llama: Open foundation models for code, 2024. URL https://arxiv.org/abs/2308.12950.

[26] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, et al. BioMedLM: A 2.7b parameter language model trained on biomedical text, 2024. URL https://arxiv.org/abs/2403.18421.

[27] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks, 2017. URL https://arxiv.org/abs/1612.00796.

[28] Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. A comprehensive survey of forgetting in deep learning beyond continual learning, 2024. URL https://arxiv.org/abs/2307.09218.

[29] Vinay V. Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics, 2020. URL https://arxiv.org/abs/2007.07400.

[30] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence, 2017. URL https://arxiv.org/abs/1703.04200.

[31] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning, 2019. URL https://arxiv.org/abs/1811.11682.

[32] Seng Pei Liew and Takuya Kato. Reusing overtrained language models saturates scaling, 2026. URL https://arxiv.org/abs/2510.06548.

[33] Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, et al. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

[34] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, et al. The FineWeb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

[35] Xiaotian Han, Yiren Jian, Xuefeng Hu, et al. InfiMM-WebMath-40B: Advancing multimodal pre-training for enhanced mathematical reasoning, 2024. URL https://arxiv.org/abs/2409.12568.

[36] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, et al. SmolLM2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

[37] Rian Touchent, Nathan Godey, and Eric de la Clergerie. Biomed-Enriched: A biomedical dataset enriched with LLMs for pretraining and extracting rare and hidden content, 2025. URL https://arxiv.org/abs/2506.20331.

[38] Tarald O. Kvålseth. Cautionary note about R^2. *The American Statistician*, 39:279–285, 1985. URL https://api.semanticscholar.org/CorpusID:120946184.

[39] Andrej-Nikolai Spiess and Natalie Neumeyer. An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a monte carlo approach. *BMC Pharmacology*, 10:6 – 6, 2010. URL https://api.semanticscholar.org/CorpusID:12765399.

[40] Anton Lozhkov, Raymond Li, Loubna Ben Allal, et al. StarCoder 2 and The Stack v2: The next generation, 2024. URL https://arxiv.org/abs/2402.19173.

A Elasto-plastic motivation

We motivate the asymmetric general/domain forms of ADIS-LAW via an elasto-plastic analogy in parameter space. We treat this as motivation, not as a structural proof from the dynamics of a neural network; the law's empirical justification is the ablation in Section 4.5.

When domain data is injected during the annealing phase, at global progress $1-\lambda = t_{\text{inj}}/T$ (where t_{inj} is the injection step), the parameters experience a perturbation whose amplitude scales with the injection intensity $I = r\frac{\eta}{\eta_0}$. We split the response into two qualitatively different components. **Elastic component (general loss).** The pre-trained representations sit in wide, flat minima. A bounded perturbation displaces the parameters but the optimization dynamics exert a restorative pull as global training progress advances after injection, so the general loss first rises and then recovers. We model this as a linear displacement–recovery dynamic, $\phi \cdot (aI - b\tau)$, where τ is the global normalized progress, which produces the observed bump. **Plastic component (domain loss).** The network simultaneously acquires target-domain statistics. The acquired information is retained, and (by analogy with strain hardening in metallurgy) further acquisition becomes harder as more domain mass is absorbed. We model this as a saturating power law $\phi \cdot \Psi r^{\mu}\tau^{\nu}$ with $\nu \in (0, 1]$, which is monotonically decreasing in τ.

We emphasize that this analogy is local—it describes a bounded perturbation around the base trajectory, not a global theory of LLM dynamics. The two derivations below should be read as "consistency derivations under stylized assumptions" that are designed to recover our functional form, not as structural proofs.

B Bump-peak derivation

The main text defines τ as global training progress, with domain injection beginning at $1-\lambda = t_{\text{inj}}/T$. The observed bump peak is the maximum of the *full* general loss $L_g = L_{\text{base}} + \Delta L_g$, not merely the maximum of the perturbation ΔL_g. This distinction matters because L_{base} is still improving during annealing and therefore contributes a negative slope. Let

$$\Delta L_g(\tau) \;=\; \phi(\tau)\,\big(aI - b\,\tau\big), \qquad g_{\text{base}}(\tau) \;\equiv\; -\frac{\mathrm{d}L_{\text{base}}}{\mathrm{d}\tau} \;\geq\; 0. \tag{7}$$

where $\phi(\tau) \equiv \phi(S_1^{\text{cpt}}(\tau))$ is zero before injection and then increases with the post-injection LR mass. The actual peak τ_{peak} satisfies

$$\frac{\mathrm{d}L_g}{\mathrm{d}\tau} = 0 \quad\Longleftrightarrow\quad \phi'(\tau)\big(aI - b\tau\big) - b\phi(\tau) \;=\; g_{\text{base}}(\tau). \tag{8}$$

Thus the peak occurs when the instantaneous upward slope created by the distribution-shift perturbation exactly equals the ordinary downward slope of the base trajectory.

For completeness, if the gate is written in terms of the cumulative post-injection LR mass, with $S_1^{\text{cpt}}(\tau) = 0$ for $\tau \leq (1-\lambda)$, the exact stationarity condition for $\tau > (1-\lambda)$ is

$$\frac{\mathrm{d}L_{\text{base}}}{\mathrm{d}\tau} + \phi'\big(S_1^{\text{cpt}}(\tau)\big)\frac{\mathrm{d}S_1^{\text{cpt}}}{\mathrm{d}\tau}\big(aI - b\tau\big) - b\phi\big(S_1^{\text{cpt}}(\tau)\big) \;=\; 0, \tag{9}$$

where $\phi'(S) = \beta_\phi E(1+ES)^{-\beta_\phi - 1}$. In the experiments, this one-dimensional condition can be evaluated directly along the recorded schedule.

Local closed form. To obtain intuition, approximate the base improvement rate as locally constant, $g_{\text{base}}(\tau) \approx g_0$, and write $x = \tau - (1-\lambda) \geq 0$ for elapsed progress since injection. Under the local generalized saturating gate $\phi(x) = 1 - (1+cx)^{-1}$, where c collects the local rate at which the gate saturates near the injection point, the full-loss peak solves

$$c(1+cx)^{-2}\big(aI - b((1-\lambda)+x)\big) - b\big[1-(1+cx)^{-1}\big] = g_0. \tag{10}$$

Multiplying by $(1+cx)^2$ gives

$$c^2(b+g_0)x^2 + 2c(b+g_0)x + \big(g_0 - c(aI - b(1-\lambda))\big) = 0. \tag{11}$$

Therefore, when the injected perturbation is strong enough to overcome the initial base improvement rate ($c(aI - b(1-\lambda)) > g_0$), the positive root is

$$\tau_{\text{peak}}^{\text{full}} \;=\; (1-\lambda) + \frac{-1 + \sqrt{\frac{b + c(aI - b(1-\lambda))}{b + g_0}}}{c}. \tag{12}$$

If $c(aI - b(1-\lambda)) \le g_0$, the full general loss is already decreasing at the injection point and there is no positive interior bump peak under this local approximation.

Relation to the perturbation-only peak. Setting $g_0 = 0$ in Eq. (12) recovers the auxiliary peak of the perturbation alone,

$$\tau_{\Delta} \;=\; (1-\lambda) + \frac{-1 + \sqrt{1 + \frac{c}{b}(aI - b(1-\lambda))}}{c}. \tag{13}$$

Since $g_0 \ge 0$ increases the denominator inside the square root, $\tau_{\text{peak}}^{\text{full}} \le \tau_{\Delta}$. The descending base trajectory therefore shifts the observed bump earlier than the peak of the perturbation term alone. Finally, $\tau_{\text{root}} = aI/b$ remains only the zero of the global recovery factor $aI - b\tau$; it is not the observed bump peak, and a positive bump requires injection to begin before this root under the local approximation.

C Domain power-law derivation

We give a stylized derivation that recovers $\Delta L_d \propto r^{\mu}\tau^{\nu}$ from a hardening-style ODE on the absorbed knowledge $K(\tau)$.

Setup. Let $K(\tau)$ denote the (notional) target-domain absorption potential indexed by the global progress variable τ. The actual loss change is still multiplied by $\phi(\tau)$, so the adaptation term is zero before the post-injection gate turns on. We posit that the marginal absorption rate is driven by the available domain signal $f(r)$ and damped by the amount already absorbed:

$$\frac{\mathrm{d}K}{\mathrm{d}\tau} \;=\; \kappa\,\frac{f(r)}{K^{m}}, \qquad m > 0. \tag{14}$$

Power-law absorption derivation. Separating variables in Eq. (14) and integrating from 0 to τ yields

$$K(\tau) \;=\; \big[(m+1)\,\kappa\, f(r)\,\tau\big]^{1/(m+1)}. \tag{15}$$

Setting $\nu = 1/(m+1) \in (0,1]$ and $f(r) \propto r^{k}$ gives $K(\tau) \propto r^{\mu}\,\tau^{\nu}$ with $\mu = k\,\nu$. Absorbing the resulting multiplicative constant into Ψ and applying the gate $\phi(\tau)$ recovers the perturbation in Eq. (6). □

Caveats. This derivation is stylized: it treats $K(\tau)$ as a scalar "knowledge" variable and assumes a specific hardening form. We do not claim it is structurally derived from gradient-descent dynamics on a neural network. Its role is to show that a saturating power-law form is consistent with a simple mechanistic picture; the ablation DL-2 vs A1 in Appendix G, Table 7, provides the actual empirical evidence that the power-law form is preferred over a linear one for calibration of L_d.

D Extended Experimental Setup and Dataset Details

To ensure complete reproducibility of the ADIS-LAW analysis and findings, this section provides comprehensive details regarding the model architecture, pre-training corpus, injection strategies, and evaluation protocols.

D.1 Model Architectures

We employ decoder-only Transformer architectures adopting the SmolLM3 structural design [33]. All models are trained from scratch with random initialization. Across all scales, we maintain a fixed vocabulary size of $128,256$, a context length of 4096 tokens, a head dimension of 128, and an approximate Grouped-Query Attention (GQA) ratio of 4:1. The exact hyperparameter configurations for the four model scales evaluated (115M, 366M, 895M, and 1.67B) are detailed in Table 4.

Table 4: Architectural hyperparameter configurations for the four model scales evaluated. N refers to the approximate parameter count.

Model Alias	N	Layers	Hidden Size	Attn Heads	KV Heads	Intermediate Size
Tiny	115M	10	512	4	1	2752
Mini	366M	12	1024	8	2	5504
Middle	895M	16	1536	12	3	8192
Large	1.67B	18	2048	16	4	11008

D.2 Base Pre-training Config & Data Budgets

The general-domain base pre-training utilizes the **FineWeb-Edu** dataset [34]. Models are optimized using a Warmup-Stable-Decay (WSD) learning rate scheduling strategy. Across all data budgets D_{total}, the warmup phase is fixed at $2,000$ steps, and the learning rate decays completely to zero during the final 20% of the training steps (the decay phase size is denoted as D_{decay}). The global batch size is maintained at 128, achieving $128 \times 4096 \approx 0.5$M tokens per step. The base pre-training data budget configurations are summarized in Table 5.

Table 5: Base pre-training configurations establishing the base trajectories before domain injection. D_{decay} represents the number of tokens in the 20% decay phase.

Total Budget (D_{total})	**Peak LR**	**Total Steps**	**Warmup Steps**	**Decay Steps**	**Decay Tokens** (D_{decay})
25B	3×10^{-4}	$47,684$	$2,000$	$9,537$	5B
50B	3×10^{-4}	$95,368$	$2,000$	$19,074$	10B
100B	3×10^{-4}	$190,735$	$2,000$	$38,147$	20B

D.3 Dataset Descriptions

To rigorously evaluate ADIS-LAW under varied distribution shifts, we utilized combinations of publicly available and internal proprietary corpora.

- **General Base Corpus:** We utilize the **FineWeb-Edu** dataset [34], a corpus of educational web text, for foundational pre-training.
- **Mathematics:** The mathematics domain injection relies on **InfiWebMath** (**InfiMM-WebMath-40B**) [35] (a curated subset of mathematical web data).
- **Code:** The programming domain utilizes **StackEdu** [36], a corpus of educational code filtered from The Stack v2[40].
- **BioMedicine:** The biomedical corpus (`BioMed`, **Biomed-Enriched**) consists of publicly available **PubMed** abstracts and full-text archives [37].
- **Finance:** The financial corpus (`Finance`) utilizes an internal, proprietary text collection (consisting of anonymized market reports and financial filings).

D.4 Domain Datasets and Injection Strategy Grids

During the continual pre-training (CPT) phase within the learning rate decay region, we inject data from four specialized domains to form the post-annealing distributions.

Our sweeping strategy explicitly controls three coupled variables governing the injection dynamics:

- **Decay-phase injection fraction** ($\lambda_{\text{decay}} \in \{0.25, 0.50, 0.80\}$): The proportion of the decay phase budget assigned to CPT. Higher λ_{decay} corresponds to an earlier start and a longer duration. This dictates the global injection fraction via $\lambda = (D_{\text{decay}}/D_{\text{total}}) \cdot \lambda_{\text{decay}}$.
- **Mixture ratio** ($r \in \{0.25, 0.50, 0.80\}$): The fraction of tokens within the CPT mixture drawn purely from the domain corpus d_o.
- **Re-warmup peak LR** ($\eta \in \{5 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4}\}$): The reset learning rate amplitude upon domain injection.

By varying the baseline scale (total budget D or parameter count N) and the CPT configuration ($(\lambda_{\text{decay}}, r, \eta)$), we populate a dense experimental grid. Table 6 summarizes the archetypal injection strategies deployed.

Table 6: Representative configuration sweep mapping the $(\lambda_{\text{decay}}, r, \eta)$ hyperparameter space to strategy archetypes (e.g., late vs. early injection, high vs. low domain saturation).

Decay Injection Fraction (λ_{decay})	**Mixture Ratio** r	**Re-warmup LR** (η)	**Strategy Archetype**
25%	50%	3×10^{-4}	Late injection, High LR
50%	25%	3×10^{-4}	Mid injection, Low ratio, High LR
50%	50%	3×10^{-4}	Mid injection, Mid ratio, High LR
50%	50%	1×10^{-4}	Mid injection, Mid ratio, Mid LR
50%	50%	5×10^{-5}	Mid injection, Mid ratio, Low LR
50%	80%	3×10^{-4}	Mid injection, High ratio, High LR
80%	50%	3×10^{-4}	Early injection, Mid ratio, High LR

This design inherently produces comprehensive experimental grids for validation:

1. **Data-Scaleup Sets (84 configurations):** Fixing the model size at 115M and iterating through three varying total pre-training data budgets $D_{\text{total}} \in \{25\text{B}, 50\text{B}, 100\text{B}\}$ across the chosen configurations and domains.
2. **Model-Scaleup Sets (56 configurations):** Fixing the pre-training data volume at 50B and scaling through four model sizes up to 1.67B parameters.

D.5 Evaluation Configurations & Metric Details

To systematically evaluate the capability of ADIS-LAW to generalize and forecast, we split the experimental combinations into fitting datasets and purely held-out evaluation sets using four distinct protocols. Specifically:

- **D-LOO & M-LOO (Interpolation):** Leave-One-Out validation. We iteratively exclude one data scale (or one model scale) from the training pool to serve as the evaluation target. This measures the law's ability to bridge the gap between known smaller and larger scales.
- **D-Up & M-Up (Extrapolation):** Strict upwards extrapolation. For example, **M-Up** utilizes solely the trajectories of 115M, 366M, and 895M models to zero-shot predict the convergence dynamics and tail losses of the unseen 1.67B scale model. No information from the target scale is leaked into the optimizer.

In all fitting routines, general loss (L_g) and domain loss (L_d) are resolved independently applying L-BFGS-B optimization against the robust Huber Loss (with $\delta = 0.05$) across 10 independent random restarts to circumvent local minima. The scaling parameter α_N governing compute-optimality is frozen at ≈ 0.2, exhibiting stable transferability across domains.

D.6 Compute Resources

All models, spanning from 115M to 1.67B parameters with pre-training token budgets up to 100B, were trained using a high-performance compute cluster equipped with 64 NVIDIA B200 GPUs. The entire experimental pipeline—encompassing base pre-training, extensive continual domain injection sweeping across hundreds of structural combinations, and rigorous held-out evaluations—required approximately one month of continuous wall-clock time on this hardware configuration.

E Extrapolation Details

In this section, we provide extended visualizations for the cross-scale extrapolation experiments discussed in Section 4.3. Figure 6 presents the complete final-loss prediction analyses, including both general loss (demonstrating the model's preservation of general knowledge) and domain loss. As shown, ADIS-LAW maintains strong predictive correlations across both objectives.

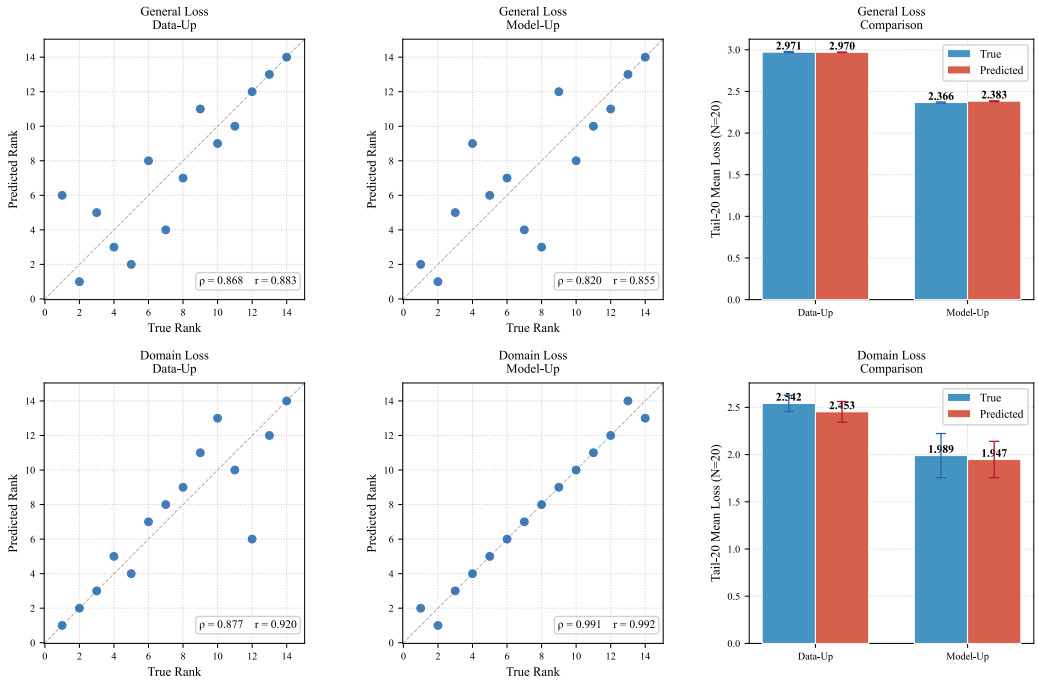

Figure 6: **Comprehensive final-loss prediction under upward extrapolation. Top row:** general loss predictive performance. **Bottom row:** domain loss predictive performance. **Scatter panels (left and center):** per-experiment true vs. predicted rank for D-Up (left) and M-Up (center). Points pool `Finance` and `BioMed`, annotated with Spearman ρ and Pearson r. **Bar panels (right):** mean tail-20 loss across all held-out strategies, with ± 1 std error bars.

E.1 Fitting stability and confidence intervals

We quantify two distinct sources of predictive uncertainty.

First, we repeated fitting across 100 distinct optimizer initializations and random seeds. We report the **mean $\pm$ standard deviation** and **95% Student-t confidence interval of the mean**:

Setting	Branch	Metric	Mean $\pm$ SD	95% CI
D-Up	General	Spearman	0.9131 ± 0.0082	$[0.9115, 0.9147]$
D-Up	General	Relative error	$0.758\% \pm 0.032\%$	$[0.752\%, 0.764\%]$
D-Up	Domain	Spearman	0.9603 ± 0.0046	$[0.9594, 0.9612]$
D-Up	Domain	Relative error	$2.833\% \pm 0.176\%$	$[2.798\%, 2.868\%]$
M-Up	General	Spearman	0.8990 ± 0.0051	$[0.8980, 0.9000]$
M-Up	General	Relative error	$0.829\% \pm 0.011\%$	$[0.827\%, 0.831\%]$
M-Up	Domain	Spearman	0.9580 ± 0.0055	$[0.9569, 0.9591]$
M-Up	Domain	Relative error	$4.196\% \pm 0.047\%$	$[4.187\%, 4.205\%]$

Second, we performed a configuration-level refit bootstrap. Each replicate resamples complete trajectories, rather than correlated checkpoints, and refits ADIS-Law before evaluation. The resulting 95% bootstrap percentile intervals are:

Setting	Branch	Metric	95% CI
D-Up	General	Spearman	$[0.834, 0.932]$
D-Up	General	Relative error	$[0.641\%, 0.961\%]$
D-Up	Domain	Spearman	$[0.899, 0.965]$
D-Up	Domain	Relative error	$[2.118\%, 3.899\%]$
M-Up	General	Spearman	$[0.863, 0.916]$
M-Up	General	Relative error	$[0.722\%, 0.953\%]$
M-Up	Domain	Spearman	$[0.916, 0.968]$
M-Up	Domain	Relative error	$[3.870\%, 4.761\%]$

These results demonstrate stability across optimizer initializations and resampled configurations.

F Detailed Strategy Selection under ε-Constraint

The Pareto frontier contains multiple mutually non-dominating strategies; deploying one in practice requires committing to a single configuration. A natural approach is ε-constraint scalarization: select the strategy that minimizes domain loss subject to $L_g \leq L_g^{\min} \cdot (1 + \varepsilon)$, where ε encodes the practitioner's tolerance for general-loss degradation. This formulation exploits the fact that ADIS-LAW predicts domain loss more robustly than general loss across scales (Dom. $\rho > 0.90$ in all settings), using the general-loss constraint only to exclude the region where prediction is least reliable.

We apply the same ε-constraint to both the predictor's and oracle's Pareto frontiers, varying ε from 0.001 to 0.20, and compare the selected strategies. Under D-Up, for $\varepsilon \geq 0.02$, the predictor and oracle select identical (λ_{decay}, r) combinations (0.50, 0.90), disagreeing only on η (1×10^{-4} vs. 1.5×10^{-4}); the exact match rate is 25%, with mean $\Delta\lambda_{\text{decay}} = 0.011$ and $\Delta r = 0.007$. Under M-Up, the predictor consistently recommends a slightly higher injection duration fraction than the oracle ($\lambda_{\text{decay}} = 0.60$ vs. 0.50), but r is matched closely (0.90) across all ε; the exact match rate is 10.7%, with $\Delta\lambda_{\text{decay}} = 0.086$ and $\Delta r = 0.014$. In both protocols, the parameter differences are small and concentrated in η and λ_{decay}, while r—the knob with the shortest causal path through the ADIS-Law formula—is the most reliably predicted.

G Detailed Ablation Study Metrics

We ablate the architectural choices in ADIS-LAW to determine which components are strictly necessary. Each variant changes exactly one component relative to the full model; all metrics are averaged across four domains and four evaluation protocols (D-LOO, M-LOO, D-Up, M-Up).

Table 7: Mechanistic ablation across four domains and four evaluation protocols. $\Delta\rho$ is the absolute change relative to ADIS-LAW full.

ID	Variant	General			Domain		
		ρ	Err%	$\Delta\rho$	ρ	Err%	$\Delta\rho$
A1	Ours Full	**0.911**	**0.85%**	—	**0.954**	**5.23%**	—
General-loss perturbation (GL)							
GL-2	Ratio-Only	0.820	1.11%	$\mathbf{-0.091}$	0.959	5.38%	$+0.005$
GL-3	Power-Law	0.921	0.83%	$+0.010$	0.958	5.20%	$+0.004$
Domain-loss absorption (DL)							
DL-2	Linear Adapt	0.910	0.86%	-0.001	0.707	22.63%	$\mathbf{-0.247}$
DL-3	Dose Coupled	0.911	0.87%	$+0.000$	0.886	5.65%	$\mathbf{-0.068}$
Saturation gate (ϕ)							
PHI-1	Exponential	0.915	0.84%	$+0.004$	0.678	6.33%	$\mathbf{-0.276}$
PHI-2	Step-wise	0.910	0.85%	-0.001	0.644	6.98%	$\mathbf{-0.310}$

(i) The generalized saturating gate ϕ is the single most critical component. Replacing ϕ with an exponential gate (PHI-1) or a step function (PHI-2) leaves Gen. ρ virtually unchanged but collapses

Dom. ρ from 0.954 to 0.678 and 0.644, respectively ($\Delta = -0.28$ and -0.31). Exponential saturation is too abrupt; step activation lacks the smooth transition needed for the network's gradual acceptance of out-of-distribution data. The MM-specific saturating shape is irreplaceable for modeling domain adaptation dynamics.

(ii) Power-law absorption ($r^\mu \tau^\nu$) is essential for domain loss. Replacing it with a linear form (DL-2, $\Psi \cdot r \cdot \tau$) destroys Dom. ρ ($0.954 \to 0.707$, $\Delta = -0.247$) and inflates Dom. Err% from 5.23% to 22.63%, confirming that domain data confers diminishing marginal returns that must be captured by power-law exponents. Replacing r with the coupled dose $I = r \cdot \eta/\eta_0$ in the absorption term (DL-3) causes a smaller but clear degradation (Dom. $\Delta = -0.068$), proving that the absorption side depends on the data mixture ratio r rather than on the LR-coupled intensity—the opposite of the forgetting side.

(iii) The dose coupling $I = r \cdot \eta/\eta_0$ is necessary for general loss. Removing the LR coupling from the forgetting term (GL-2, using r alone) drops Gen. ρ from 0.911 to 0.820 ($\Delta = -0.091$) while leaving Dom. ρ unaffected. Together with DL-3, this establishes a clean asymmetry: general-loss forgetting is driven by the combined intensity of data volume and learning rate (dose), while domain-loss absorption is driven by the data mixture ratio (r) alone.

H Extended Pareto Optimization Analysis: Finance Domain

In Section 4.4, we presented the predictive optimization capabilities of ADIS-LAW on the `BioMed` domain. To verify that our findings generalize to domains with significantly different distribution shifts, we additionally evaluate the Pareto optimization capabilities on the Finance domain.

H.1 Quantitative Comparison: Finance vs. BioMed

Table 8 summarizes the global ranking fidelity, Pareto frontier overlap, and decision quality metrics for both domains. Across both Data-Up (extrapolating from 50B to 100B tokens) and Model-Up (extrapolating from up to 895M to 1.67B parameters), ADIS-LAW achieves exceptional ranking correlations and low Relative Errors. The Mean Relative Regret when applying the predictor's recommended optimal strategy is consistently below 3%, indicating robust utility across different datasets.

Table 8: Comprehensive Comparison of Pareto Strategy Prediction (Finance vs. BioMed). ρ denotes Spearman Rank Correlation. Regret evaluates the performance drop when deploying the proxy-recommended optimal strategy compared to the true Oracle strategy.

Extrapolation	**Domain**	**Spearman $\rho \uparrow$**		**Rel. Err $\varepsilon_{\text{rel}} \downarrow$**		**Chamfer $\downarrow$**	**Mean Regret $\downarrow$**
		Gen.	Dom.	Gen.	Dom.	**Dist.**	
Data-Up	`Finance`	0.989	0.986	0.45%	5.38%	0.0749	0.90%
Data-Up	`BioMed`	0.994	0.977	0.96%	3.36%	0.0588	0.37%
Model-Up	`Finance`	0.794	0.908	1.18%	9.92%	0.0684	1.93%
Model-Up	`BioMed`	0.858	0.914	0.79%	3.91%	0.0381	2.92%

H.2 Finance Target-Scale Pareto Visualization

Figures 7 and 8 visualize the surrogate-predicted strategies mapped against the dense 486-policy target-scale `Finance` oracle evaluation. Much like `BioMed`, the `Finance` recommendations closely hug the true oracle frontier, again validating the geometric consistency of ADIS-LAW.

I Post-Annealing Domain Injection

The main experiments inject domain data *during* the annealing (decay) window, reusing part of the base decay budget. A stricter test of the perturbation formulation is whether it extends to the setting where injection begins only *after* annealing has fully completed. In this protocol, pre-training is run to completion, and domain data is appended from the final checkpoint with a fresh learning rate

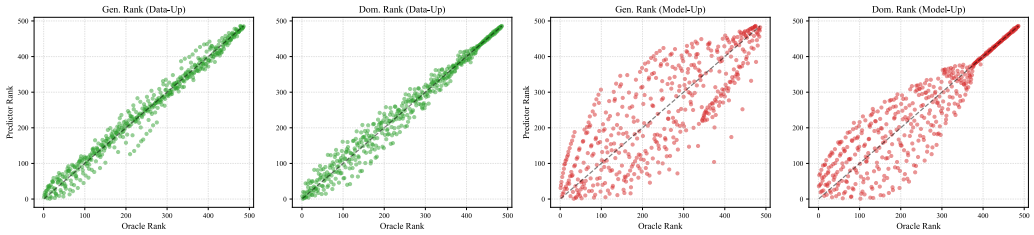

Figure 7: **Predicted vs. oracle rank across 486 synthetic policies on `Finance`.** Global correlations remain tight with minimal noise, mimicking the stability seen on BioMed.

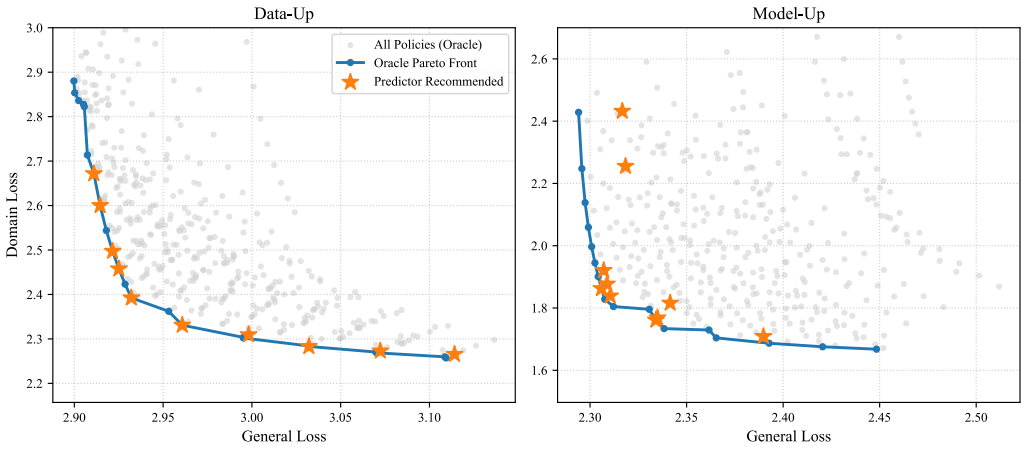

Figure 8: **Finance Pareto frontier alignment.** The proxy-recommended strategies (stars) closely trace the true target-scale frontier (blue line).

schedule. This increases the total token budget from D_{base} to $D_{\text{base}} + D_{\text{cpt}}$, so injection occurs at a late global progress fraction $\tau_d = D_{\text{base}}/(D_{\text{base}} + D_{\text{cpt}})$ rather than within the decay window. If ADIS-LAW remains predictive under this regime shift, it suggests that the perturbation geometry captures a structural property of domain injection that is independent of when injection occurs relative to the pre-training schedule. We evaluate this on `BioMed`.

Two leave-one-out protocols are tested. **D-LOO** holds out one data budget (25B, 50B, or 100B) at a fixed model scale ($N = 115$M); **M-LOO** holds out one model scale (115M, 366M, 895M, or 1.67B) at a fixed data budget ($D_{\text{base}} = 50$B). In each fold, ADIS-LAW is fit on the training scales and evaluated on the held-out scale.

Table 9: Post-annealing domain injection on `BioMed`. ADIS-LAW is evaluated under leave-one-out protocols across data budgets and model scales, reporting both pre-training (PT) and CPT-phase metrics.

Protocol	**Gen.** $\rho \uparrow$	**Gen. Err%** $\downarrow$	**Dom.** $\rho \uparrow$	**Dom. Err%** $\downarrow$
PT phase				
D-LOO	0.987	0.69%	0.937	1.22%
M-LOO	0.987	1.30%	0.934	1.33%
CPT phase				
D-LOO	0.908	1.00%	0.977	3.84%
M-LOO	0.904	1.34%	0.982	2.85%

Table 9 reports the results. The PT phase is predicted with high fidelity across both protocols (Gen. $\rho = 0.987$, Gen. Err $< 1.4\%$), confirming that the base trajectory formulation transfers without modification to the post-annealing setting. For the CPT phase, general-loss prediction is stable (Gen.

$\rho \approx 0.91$, Gen. Err $< 1.4\%$), and domain-loss prediction is strong in both settings: Dom. $\rho = 0.977$ for D-LOO and 0.982 for M-LOO, with Dom. Err below 4%. Notably, the cross-model-scale domain-loss ranking is the highest observed across all experimental settings in this paper (Dom. $\rho = 0.982$), indicating that the perturbation formulation transfers cleanly across model scales when injection occurs after the base run.

These results demonstrate that ADIS-LAW is not merely a curve fit to annealing-window injection. The perturbation geometry generalizes to post-annealing injection without modification, suggesting that the separation of base trajectory and domain adaptation dynamics in Eqs. (5)–(6) captures a structural property of domain injection that is largely independent of when injection occurs relative to the pre-training schedule.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state our contributions (a phenomenological scaling law, ablation, and cross-scale extrapolation) and the empirical numbers reported there match the results in Section 4; limitations are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 5, we discuss the phenomenological nature of the law, the limited model/data scale, the high-intensity failure mode and underestimation of transient loss peaks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The modelling assumptions are stated in Sections 3.2–3.4, and the corresponding derivations are provided in Appendices B and C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 and Appendix D describe model sizes, data budgets, schedules, the injection-strategy grid, fitting procedures, and evaluation protocols; trajectory smoothing is specified in Figure 2. The $S_1/S_2/\phi$ definitions are given in Sections 3.1–3.2.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The pre-training and CPT runs use a mixture of public and internal corpora (including an internal finance corpus), and the trained checkpoints are not released; we will release the fitting and evaluation code for ADIS-LAW and the baselines upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies the model and data scales, the injection-strategy grid, the WSD schedule shape, the smoothing applied to trajectories, and the LOO-CV protocol; the fitted parameter set θ is described in Section 3.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report point estimates of Spearman correlation ρ and relative error $RelErr\%$ averaged across LOO-CV splits and domains. In Appendix E.1, we report means, standard deviations, and confidence intervals for ρ and $RelErr\%$ under D-Up and M-Up.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All training runs use the WSD schedule on standard transformer-decoder architectures (115M to 1.67B parameters) with 25–100B tokens. Detailed hardware configuration and time-of-execution details (64 NVIDIA B200 GPUs, approx. 1 month) are provided in Appendix D.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly available pre-training corpora and an internal finance corpus; no human subjects or sensitive data are involved, in line with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: ADIS-Law models LLM training dynamics to enable more principled and efficient selection of continual pre-training strategies. Thus this work has no societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release new high-risk models or scraped datasets in this paper; the fitting code released upon acceptance does not pose misuse risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: Dataset sources are described in Appendix D.3; the architecture source and prior scaling-law work are cited in the paper. No third-party assets are redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper does not release any new datasets or pre-trained model checkpoints.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are the object of study; they are not used as a non-standard component of the methodology beyond the standard pre-training pipeline.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.