Projected LR-Drop Residuals for Source-Only Cosine-to-WSD Loss-Curve Prediction

Jiaju Wu
Peking University
https://github.com/wjjpku/DL-final

June 30, 2026

Abstract

This paper studies a narrow transfer problem in loss-curve prediction: given a strong Multi-Power Law (MPL) baseline and one observed cosine training curve, can we improve predictions for warmup-stable-decay (WSD) schedules without fitting any WSD target loss? The difficulty is an identification problem. The source cosine residual contains both a transferable learning-rate (LR) drop response and non-transferable drift from errors in MPL's LR-dependent parameters. Directly fitting a one-dimensional response amplitude from the raw cosine residual therefore fails. We propose a low-capacity residual estimator,

$$\widehat{L}_s(t) = L_{\mathrm{MPL},s}(t) + a_s \widehat{\kappa}_s \phi_{\lambda_s,s}(t),$$

where the response feature $\phi_{\lambda_s,s}$ is a causal LR-drop convolution, the response rate λ_s is chosen from a schedule-only Herfindahl drop concentration q_s, and the locality factor a_s is a support-projection energy fraction. The only quantity fit from loss residuals is one nonnegative scalar $\widehat{\kappa}_s$, estimated after projecting both the cosine residual and the response feature away from the MPL-LD tangent nuisance subspace. On the current public loss-curve repository, the selected source-only estimator improves same-scale WSD-family MAE by 29.88% on average, improves all 15/15 WSD rows, and remains non-harmful on all controls. Cross-scale transfer improves all 30/30 WSD rows with 24.95% mean MAE reduction. These results support the method as an interpretable residual-identification correction, not as a universal training-loss law.

1 Introduction

Learning-rate schedules shape the transient and tail behavior of neural-network training curves. This matters for schedule search: a useful predictor should estimate how a target schedule will behave before running the target training job. Prior scaling-law work models loss as a function of compute, data, model size, and optimization details [5, 6, 1, 2]. The present work asks a more restricted question focused on WSD-style cooldown behavior [3, 4]. Suppose we already have a strong MPL prediction for each schedule and one observed source curve trained with a cosine decay. Can the source residual tell us how MPL will miss WSD-family schedules?

The answer is not obtained by simply copying the cosine residual. MPL already explains most of the loss curve. The remaining residual is small, structured, and confounded. In particular, a source cosine residual mixes at least two effects:

$$\text{transferable LR-drop response} \quad + \quad \text{non-transferable MPL-LD parameter drift.}$$

The first effect is the signal we want to transfer to WSD schedules. The second is a nuisance term: it reflects local errors in MPL's LR-dependent parameters, and transferring it to a different schedule can make predictions much worse.

This paper frames cosine-to-WSD transfer as a residual-identification problem. We keep the deployable correction deliberately small: the target schedule determines the response shape, the source cosine residual

determines one scalar amplitude, and target WSD losses are used only for evaluation and oracle diagnostics. The selected estimator is

$$\widehat{L}_s(t) = L_{\mathrm{MPL},s}(t) + a_s \widehat{\kappa}_s \phi_{\lambda_s,s}(t). \tag{1}$$

Here $L_{\mathrm{MPL},s}$ is the frozen MPL prediction; $\phi_{\lambda_s,s}$ is a causal response to positive LR drops; λ_s is computed from the effective concentration of the target schedule's LR drops; a_s removes diffuse full-horizon schedule modes; and $\widehat{\kappa}_s$ is the only fitted residual coefficient.

Contributions. The main contribution is an interpretable source-only correction for WSD-family loss-curve prediction. The paper makes four claims:

- **Identification.** Raw cosine residual transfer fails because the residual is contaminated by MPL-LD parameter drift.
- **Method.** MPL-LD tangent projection isolates a one-dimensional LR-drop response amplitude that can be transferred from cosine to WSD targets.
- **Schedule-only response.** The response rate and locality factor can be computed from the LR schedule itself using q_2 half-life and support projection, with no WSD loss calibration.
- **Evidence.** The selected estimator improves all current WSD-family rows in both same-scale and cross-scale audits while abstaining on diffuse or constant controls.

2 Problem Setup

Frozen MPL baseline. We use MPL as a fixed baseline predictor [2]. In simplified notation,

$$L_{\mathrm{MPL}}(t) = L_0 + AS(t)^{-\alpha} + B\sum_{k\le t} \Delta\eta_k \, G\big(\eta_k^{-\gamma}(S(t)-S(k))\big), \qquad S(t) = \sum_{i\le t}\eta_i. \tag{2}$$

The LR-dependent MPL parameters are (B, C, β, γ), where G is the saturating kernel used by MPL. We do not refit MPL in this paper. The task is to explain and transfer part of the residual left by this frozen baseline.

Source-only transfer protocol. For each scale, the calibration curve is `cosine_72000.csv`. The target WSD-family curves are `wsd_20000_24000`, `wsdld_20000_24000`, and `wsdcon` endpoints 3e−5, 9e−5, and 18e−5. These targets are evaluated at 25M, 100M, and 400M scales, giving 15 same-scale WSD rows. Cross-scale evaluation applies a source cosine residual from one scale to WSD targets at a different scale, giving 30 WSD rows. Controls include short cosine and constant schedules.

The target WSD losses are not used to choose λ_s, a_s, $\widehat{\kappa}_s$, the calibration window, or any model parameter. They are used only to compute evaluation MAE and oracle diagnostics.

Metric. All tables report percentage MAE change relative to frozen MPL:

$$\Delta_{\mathrm{MAE}} = 100 \cdot \frac{\mathrm{MAE}(\widehat{L}) - \mathrm{MAE}(L_{\mathrm{MPL}})}{\mathrm{MAE}(L_{\mathrm{MPL}})}. \tag{3}$$

Negative values are improvements. We also report row wins, where a win means lower MAE than MPL on that row, and non-harm, where the correction is not worse than MPL.

3 Method

3.1 Residual Decomposition

Let the source cosine residual be

$$r_{\cos}(t) = L_{\cos}(t) - L_{\mathrm{MPL},\cos}(t). \tag{4}$$

The working decomposition is

$$r_{\cos}(t) = \kappa_\star \phi_{\lambda,\cos}(t) + J_{\mathrm{LD}}(t)\Delta\theta_{\mathrm{LD}} + \epsilon(t). \tag{5}$$

The first term is the transferable LR-drop response. The second term is nuisance drift caused by local error in MPL's LR-dependent parameters. The last term is residual noise and unmodeled structure. Equation (5) is not a full dynamical theory; it is the identification assumption that motivates the estimator.

3.2 Causal LR-Drop Response Feature

For a schedule s, define the positive LR drop

$$d_t = [\eta_{t-1} - \eta_t]_+. \tag{6}$$

The response feature is a causal one-pole convolution in cumulative-LR time:

$$z_t = \exp(-\lambda_s \eta_t) z_{t-1} + d_t, \qquad \phi_{\lambda_s,s}(t) = \frac{z_t}{\eta_{\max}}, \qquad z_0 = 0. \tag{7}$$

Equivalently,

$$\phi_{\lambda_s,s}(t) = \frac{1}{\eta_{\max}} \sum_{u \le t} d_u \exp\left(-\lambda_s \sum_{v=u+1}^{t} \eta_v\right). \tag{8}$$

This design has three constraints. It is causal, since only previous drops contribute. It is drop-only, since constant LR intervals inject no new response. It is single-channel, since the same formula is used for WSD, WSD-linear, WSD-con, cosine, and constant controls.

3.3 q_2 Half-Life Response Rate

The response rate must be schedule-dependent but target-loss-independent. We convert the positive LR drops into a probability distribution:

$$D_s = \sum_t d_t, \qquad p_t = \frac{d_t}{D_s}, \qquad q_s = \sum_t p_t^2 = \frac{1}{n_{\mathrm{eff}}}. \tag{9}$$

When the total LR drop occurs in one step, $q_s = 1$. When it is spread uniformly over n steps, $q_s = 1/n$. Thus q_s is the inverse effective number of drop atoms.

The public curves are logged mainly every $\Delta_{\mathrm{obs}} = 128$ training steps. We anchor the response to this observation resolution:

$$\lambda_{\mathrm{obs}} = \frac{\log 2}{\eta_{\max} \Delta_{\mathrm{obs}}}. \tag{10}$$

The selected half-life rule interpolates between a two-observation half-life for diffuse drops and a one-observation half-life for a single sharp drop:

$$H_s = (2 - q_s)\Delta_{\mathrm{obs}}, \qquad \lambda_s = \frac{\log 2}{\eta_{\max} H_s} = \frac{\lambda_{\mathrm{obs}}}{2 - q_s}. \tag{11}$$

This rule is intentionally modest. It is a resolution-based structural prior, not a claim that optimizer dynamics have a universal decay constant.

3.4 Support-Projection Locality

A local WSD cooldown response should not be applied to a full-horizon diffuse cosine decay. The locality factor is

$$a_s = \mathbf{1}\{D_s > 0\} \frac{\|(I - P_u) m_s\|_2^2}{\|m_s\|_2^2} = \mathbf{1}\{D_s > 0\} \left[1 - \frac{\ell_s}{T_s - W}\right]_+. \tag{12}$$

Here $T_s - W$ is the post-warmup horizon, ℓ_s is the span of the positive-drop support, m_s is the uniform density on that support, and u is the uniform diffuse mode on the full post-warmup horizon. The equality follows from $\|m_s\|_2^2 = 1/\ell_s$ and $\|P_u m_s\|_2^2 = 1/(T_s - W)$. Therefore a_s is not a learned gate or schedule-family label. It is the remaining local forcing energy after removing the full-horizon diffuse mode.

3.5 MPL-LD Tangent Nuisance Projection

The central step is to remove MPL-LD drift before estimating $\widehat{\kappa}_s$. We form the tangent basis

$$J_{\mathrm{LD}}(t) = \left[\frac{\partial L_{\mathrm{MPL}}}{\partial \log B}, \frac{\partial L_{\mathrm{MPL}}}{\partial \log C}, \frac{\partial L_{\mathrm{MPL}}}{\partial \log \beta}, \frac{\partial L_{\mathrm{MPL}}}{\partial \log \gamma}\right], \tag{13}$$

computed by finite differences of the frozen MPL formula on the source cosine suffix. Let P_{LD} be the orthogonal projection onto the columns of J_{LD}. For a target schedule s, define

$$x_s = (I - P_{\mathrm{LD}})\phi_{\lambda_s,\cos}, \qquad y = (I - P_{\mathrm{LD}})r_{\cos}. \tag{14}$$

The amplitude is the nonnegative ridge projection

$$\widehat{\kappa}_s = \frac{\langle x_s, y\rangle_+}{\|x_s\|_2^2 + 1/N_{\mathrm{cal}}}. \tag{15}$$

Only this scalar is fit from the source residual. The response rate, target feature, locality factor, and nuisance subspace are determined by the LR schedule and the frozen MPL formula. The ridge $1/N_{\mathrm{cal}}$ is a finite-sample identifiability floor on the cosine suffix $t \geq 8000$, which contains $N_{\mathrm{cal}} = 500$ logged points.

3.6 Calibration and Prediction Algorithm

The procedure separates calibration from target evaluation.

Calibration on the source cosine curve.

1. Load the source `cosine_72000.csv` curve and compute the frozen MPL residual $r_{\cos}$.
2. Compute the MPL-LD tangent basis J_{LD} on the calibration suffix and form P_{LD}.
3. For each target schedule, compute q_s, λ_s, and a_s from its LR schedule only.
4. Build the source response feature $\phi_{\lambda_s,\cos}$, residualize it and the source residual using $I - P_{\mathrm{LD}}$, and estimate $\widehat{\kappa}_s$ by Eq. (15).

Prediction on a target schedule.

1. Build the target response feature $\phi_{\lambda_s,s}$ from the target LR schedule.
2. Predict with Eq. (1).
3. Use the target loss curve only after prediction, to compute MAE and diagnostics.

4 Why the Method Should Work

Linear-response intuition. Suppose that after subtracting the MPL baseline there remains a small excess-loss state e_t that reacts to LR decreases with finite memory. A first-order approximation is

$$e_t \approx \rho_t e_{t-1} + b d_t, \qquad \rho_t \approx \exp(-\lambda_s \eta_t). \tag{16}$$

Solving this recursion gives the feature in Eq. (7), up to an unknown scalar mapping from the hidden state to loss. The scalar is precisely the role of $\widehat{\kappa}_s$.

Why raw projection fails. If $r_{\cos}$ were a clean response signal, regressing it directly on $\phi_{\lambda,\cos}$ would be enough. The audits show the opposite. The no-nuisance variant worsens WSD-family MAE by hundreds of percent and wins 0/15 rows. This failure is the main evidence for the identification framing: the source residual is useful only after MPL-LD nuisance directions have been removed.

Table 1: Main evaluation of the selected source-only estimator. Values are percentage MAE changes relative to frozen MPL; negative is better.

split	group	mean	median	worst	wins / non-harm
same scale	WSD-family	−29.88	−35.97	−4.67	15/15, 15/15
cross scale	WSD-family	−24.95	−23.21	−3.15	30/30, 30/30
all source-target	WSD-family	−26.59	−24.33	−3.15	45/45, 45/45
same scale	controls	+0.00	+0.00	+0.00	0/9, 9/9
all source-target	controls	+0.00	+0.00	+0.00	0/27, 27/27

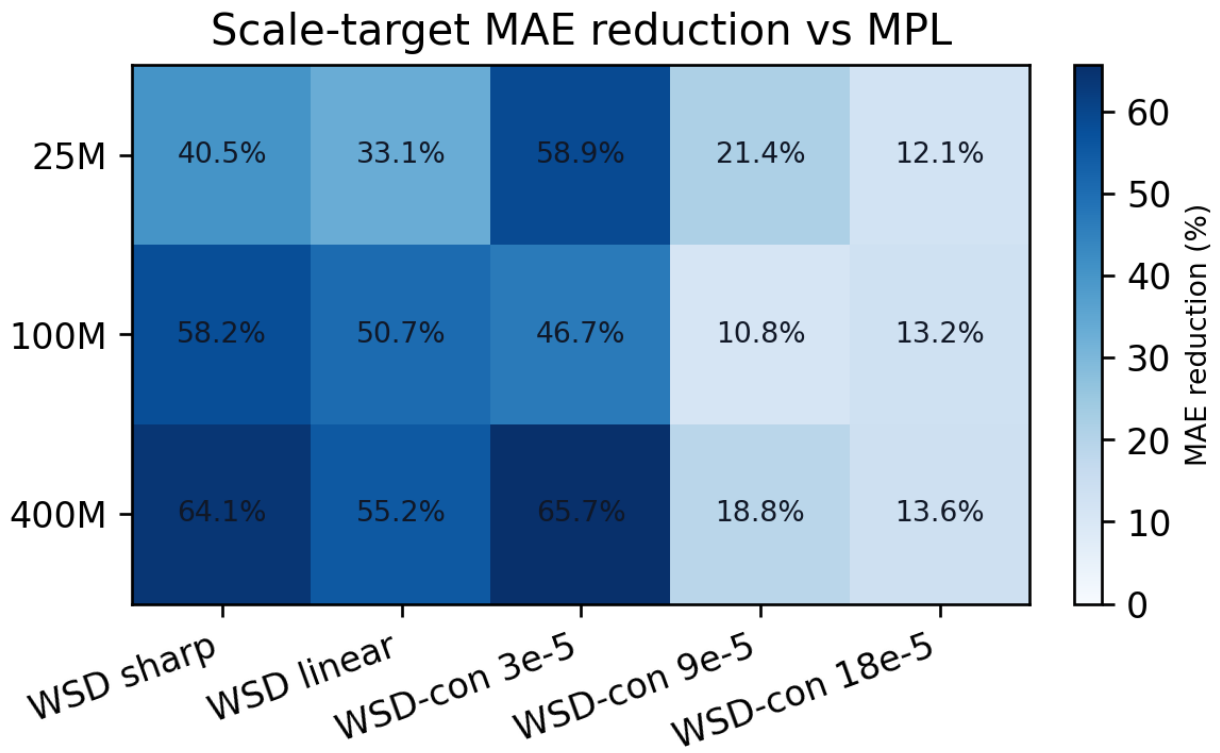

Figure 1: MAE changes for the selected estimator. The correction is calibrated from the source cosine residual only; WSD target losses are used only to evaluate the completed prediction.

Why the selected schedule rules are conservative. The q_2 half-life rule replaces hand-routed schedule classes with an effective drop-count statistic. The support-projection locality replaces an empirical gate with a geometric energy fraction. Both choices preserve the low-capacity nature of the estimator: no target losses and no schedule-family labels enter the deployable prediction.

5 Experiments

5.1 Main Results

Table 1 summarizes the selected q_2 half-life MPL-LD response with support-projection locality. It improves every WSD-family row in both same-scale and cross-scale transfer. The controls remain unchanged because constant schedules have no LR drops and the short-cosine control has zero support-projection locality.

5.2 Per-Target Behavior

Table 2 shows that the largest gains occur on WSD sharp, WSD linear, and the lowest WSD-con endpoint. Higher WSD-con final-LR rows still improve, but by a smaller margin. This is consistent with the limitation

Table 2: Same-scale WSD-family results by target schedule.

target	mean	worst row	wins
WSD sharp	−46.27	−37.12	3/3
WSD linear	−39.73	−29.81	3/3
WSD-con 3e−5	−41.12	−35.97	3/3
WSD-con 9e−5	−13.40	−9.17	3/3
WSD-con 18e−5	−8.88	−4.67	3/3

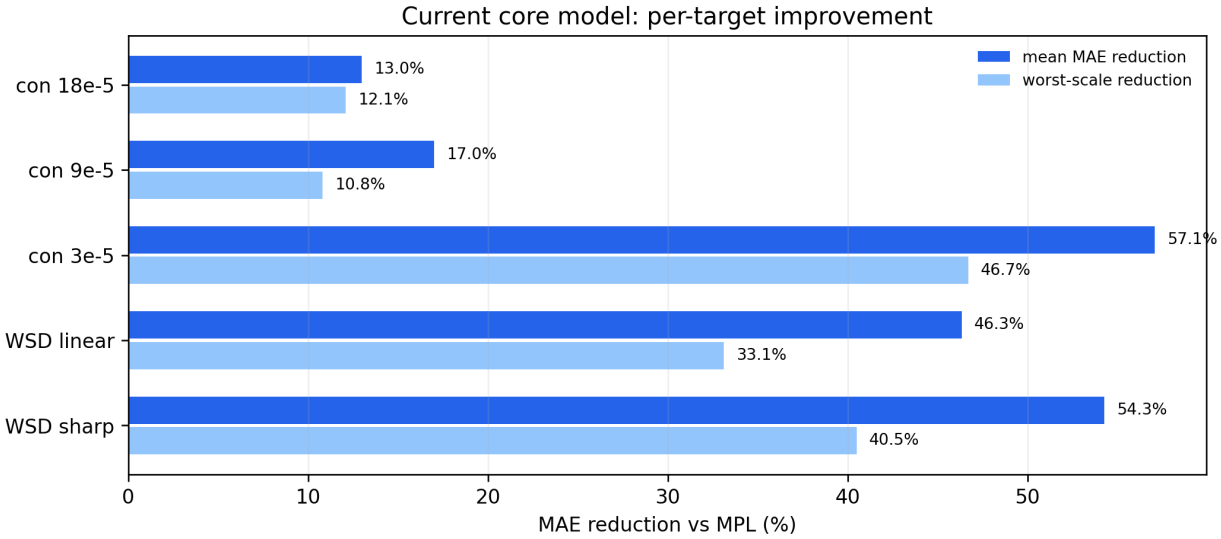

Figure 2: Per-target MAE changes. The selected mechanism-native estimator improves every WSD-family row, with smaller gains on high-end WSD-con targets.

that WSD-con fine ranking is weaker than the coarse distinction between cooldown families.

5.3 Ablations and Negative Controls

Table 3 gives the main ablations. The no-nuisance raw projection is the decisive negative control: it uses the same source residual and response feature but skips MPL-LD tangent projection, and it fails catastrophically. The no-locality variant has a slightly stronger WSD mean (−30.88%) but harms controls, so it is not the selected formula. The older q_∞ rule is numerically almost identical to q_2, but q_2 has the clearer effective-drop-count interpretation. DCT projection remains a useful performance reference, but it is not the main method because its nuisance basis is generic low frequency rather than tied to MPL parameter drift.

5.4 Schedule Diagnostics

The schedule-derived quantities in Table 4 show how the formula behaves. WSD sharp and WSD linear have long but finite drop supports, so $a_s \approx 0.817$. WSD-con curves have a single-step drop, so $q_s = 1$ and $a_s \approx 1$. The short-cosine control has a diffuse full-horizon support and is projected to zero by the locality factor.

5.5 Residual Curves

Figure 5 compares residual curves at 100M. The goal is not to fit every point perfectly. The useful behavior is that the correction reduces WSD-family residuals while leaving safety controls unchanged.

Table 3: Theory-refinement and safety ablations.

variant	WSD mean	WSD worst	WSD wins	control reading
q_2 half-life + support projection	−29.88	−4.67	15/15	9/9 non-harm
old q_∞ + support projection	−29.87	−4.67	15/15	9/9 non-harm
q_2 density projection	−30.22	−4.67	15/15	worst +8.25, 6/9 non-harm
q_2 no locality	−30.88	−4.67	15/15	worst +56.99, 6/9 non-harm
no-nuisance raw projection	+602.17	+2366.35	0/15	fails on WSD
old MPL-LD fixed ridge	−27.25	−3.00	15/15	9/9 non-harm
DCT performance reference	−32.83	−5.30	15/15	cross-scale worst +26.68

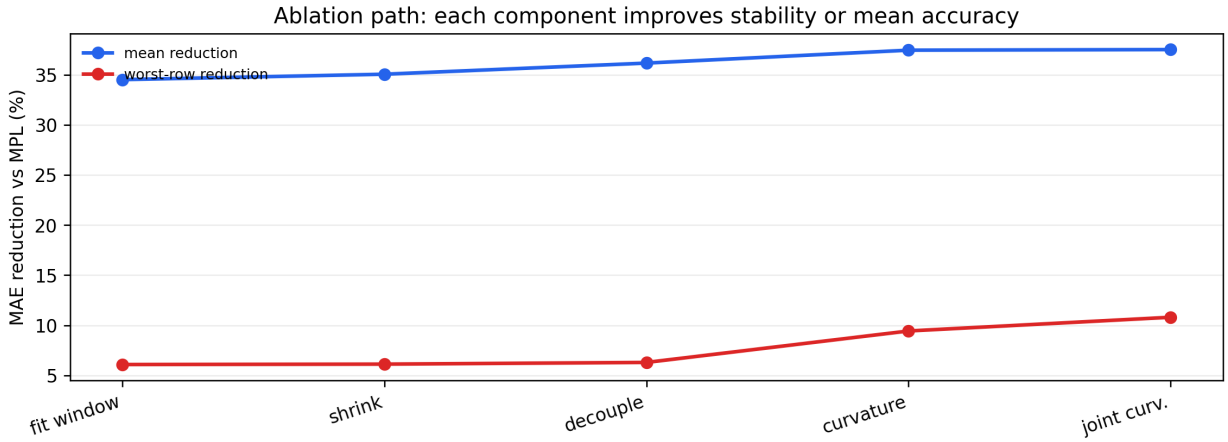

Figure 3: Ablation summary. The selected formula trades a small amount of WSD mean improvement for a cleaner support-projection interpretation and non-harmful controls.

6 Discussion

What is supported. The strongest supported claim is specific: after removing local MPL-LD tangent drift from the source cosine residual, a one-dimensional causal LR-drop response transfers reliably to the current WSD-family targets. The response amplitude is low capacity, the response shape is schedule-derived, and the main protocol does not fit WSD target losses.

Why this is not just residual overfitting. The selected estimator fits one scalar from the source residual. Its response rate, locality factor, nuisance subspace, and ridge scale are determined by schedule statistics, observation resolution, and the frozen MPL formula. Several stronger or more flexible alternatives were rejected because they weakened the interpretation or failed safety checks: direct raw projection, no-locality correction, density projection, and generic DCT projection.

Current research extension. The repository also contains a broader schedule-response amplitude line based on nuisance- projected empirical-Bayes estimation, identifiable-amplitude conversion, and predictive shrinkage. Those audits are useful for future general transfer beyond cosine-to-WSD. They are not the main claim here because the course-facing deliverable is the stricter source-cosine WSD-family setting, and the next-generation shrinkage prior still needs external validation.

Limitations. The q_2 half-life bracket uses the logging interval as a structural prior. It is cleaner than earlier hand-written response endpoints, but it is not a theorem about optimizer dynamics. The calibration suffix $t \geq 8000$ is selected by a source-only identifiability rule; later suffixes remain positive but reduce the gain. The support-projection locality explains why diffuse controls should be suppressed, but it is a boundary condition rather than a complete dynamical derivation. Finally, all main numbers come from the current

Table 4: Schedule-only diagnostics used by the selected formula.

curve	q_2	n_{eff}	support span	a_s
WSD sharp	3.52e−4	2842.3	4000	0.8168
WSD linear	2.50e−4	3999.0	4000	0.8168
WSD-con targets	1.0000	1.0	2	0.9999
Cosine 24k control	5.65e−5	17702.8	21840	0.0000
Constant controls	0.0000	0.0	0	0.0000

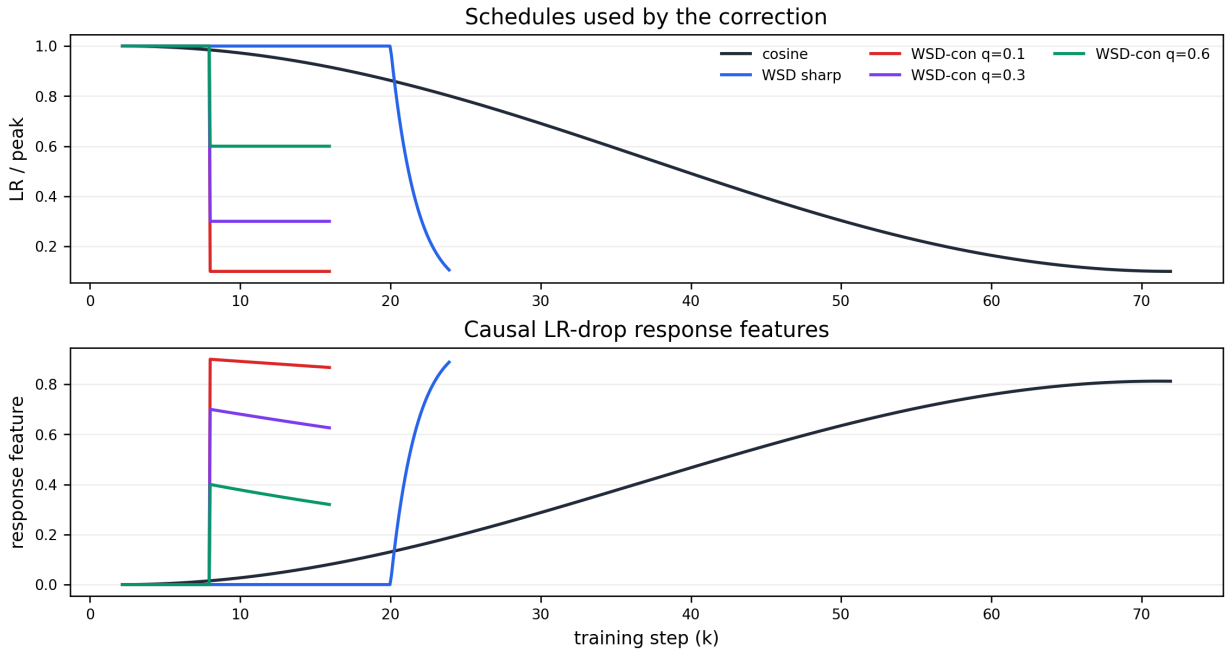

Figure 4: Schedule features used by the formula. The construction separates concentrated LR-drop responses from diffuse full-horizon decay without using target loss.

public loss-curve repository. New held-out WSD schedules or new training runs are required before claiming broad generalization.

7 Conclusion

This paper rewrites the project around a single residual-identification story. MPL is the frozen baseline. Source cosine residuals contain both transferable LR-drop response and non-transferable MPL-LD drift. A causal response feature becomes useful only after projecting away the MPL-LD tangent nuisance. The final deployable correction is

$$\widehat{L}_s(t) = L_{\text{MPL},s}(t) + a_s \widehat{\kappa}_s \phi_{\lambda_s,s}(t), \qquad \lambda_s = \frac{\lambda_{\text{obs}}}{2 - q_s}, \qquad q_s = \sum_t \left(\frac{d_t}{\sum_u d_u} \right)^2 .$$

With support-projection locality, the method improves all current WSD-family rows (−29.88% same-scale mean and −24.95% cross-scale mean) while leaving controls non-harmful. Its main value is interpretability under a strict source-only calibration protocol: the correction is small, schedule-derived, and explicit about the remaining validation gap.

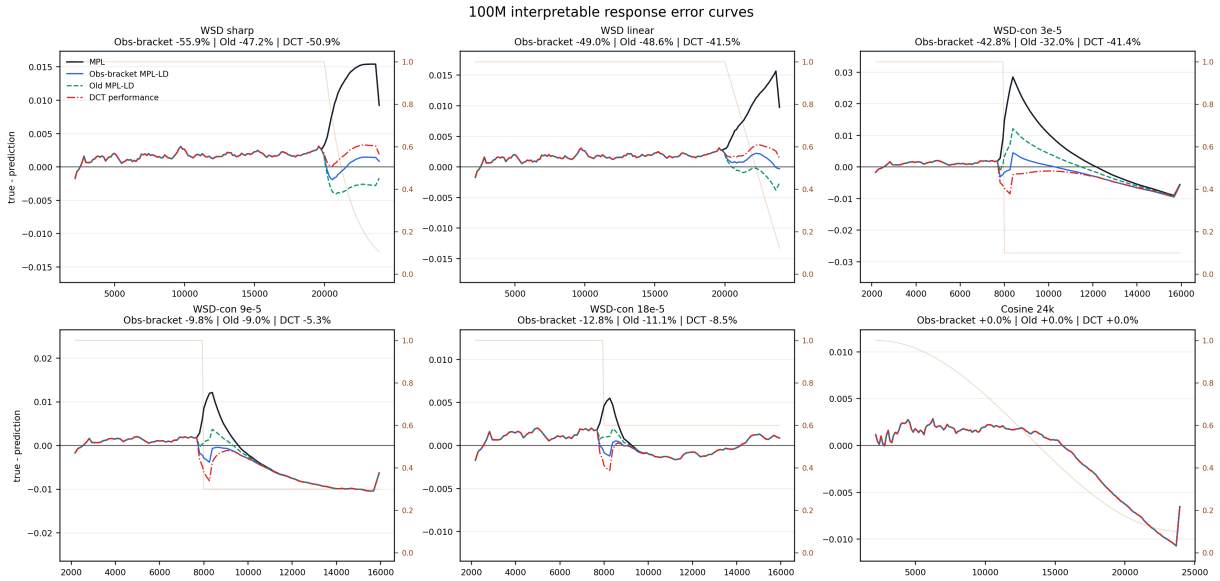

Figure 5: Representative error curves at 100M. The correction is fitted from the source cosine residual only and then evaluated on WSD-family targets and controls.

A Implementation Details

Calibration window. The source residual is fit on the suffix $t \geq 8000$. In the source-only audit, this is the earliest candidate start at which the projected response-feature retention falls below the finite-sample floor for the observation-bracket endpoint checks. The selection does not enumerate WSD target losses.

Finite-difference tangent basis. The MPL-LD tangent basis is computed by perturbing

$$\log B, \qquad \log C, \qquad \log \beta, \qquad \log \gamma$$

around the frozen MPL parameters, then evaluating the resulting loss-curve differences on the source cosine suffix. The projection is an ordinary least-squares orthogonal projection onto these four columns.

Target-loss usage. The leakage audit records the boundary explicitly: calibration uses the source cosine curve, the source LR schedule, and the frozen MPL prediction; target feature construction uses only the target LR schedule; prediction uses $L_{\mathrm{MPL},s}$, $\phi_{\lambda_s,s}$, a_s, and $\widehat{\kappa}_s$. Target WSD losses are used only for MAE evaluation, oracle $\kappa_\star$, and oracle lambda-grid diagnostics. Oracle quantities are not deployable model inputs.

Reproduction artifacts. The main artifacts used in this paper are:

- `results/interpretable_theory_refinement/REPORT.md`
- `results/schedule_response_robustness/REPORT.md`
- `results/schedule_response_robustness/LEAKAGE_AUDIT.md`
- `paper/figs/`

References

[1] H. Tissue, V. Wang, and L. Wang. Scaling law with learning rate annealing. *arXiv:2408.11029*, 2024.

[2] K. Luo, H. Wen, S. Hu, et al. A multi-power law for loss curve prediction across learning rate schedules. *arXiv:2503.12811*, 2025.

[3] K. Wen, Z. Li, T. Ma, et al. Understanding warmup-stable-decay learning rates: a river valley loss landscape perspective. *arXiv:2410.05192*, 2024.

[4] A. Dremov, A. Hagele, A. Kosson, and M. Jaggi. Training dynamics of the cooldown stage in warmup-stable-decay learning rate scheduler. *arXiv:2508.01483*, 2025.

[5] J. Kaplan, S. McCandlish, T. Henighan, et al. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

[6] J. Hoffmann, S. Borgeaud, A. Mensch, et al. Training compute-optimal large language models. *NeurIPS*, 2022.