(assistance under GPT5.6)

3+3+(2) Scaling laws for LLM Pre-training

Fanghui Liu

fanghui.liu@sjtu.edu.cn

Joint work with Yichen Wang and Yudong Chen (UW-Madison)

Institute of Natural Sciences, School of Mathematical Sciences
Shanghai Jiao Tong University (SJTU)

at Peking University, 2026.10



Compute-optimal scaling laws

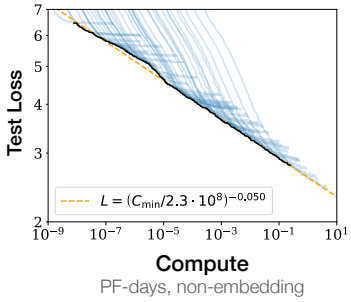


Figure 1: OpenAI / Kaplan et al. (2020)

Scaling laws (e.g., Kaplan, Chinchilla)

$$\mathcal{L}(M, N) \simeq \mathcal{L}_\infty + aM^{-\alpha} + bN^{-\beta}$$

M : model size; N : training token

Method	OpenAI	DeepMind
$M_\star$	$C^{0.73}$	$C^{0.49}$
$N_\star$	$C^{0.27}$	$C^{0.51}$
TPP $N_\star/M_\star$	$C^{-0.46}$	constant
Policy	$\nearrow M \gg \nearrow N$	$\nearrow M \approx \nearrow N$

Is a scaling law universal?

Origin

When does a learning curve follow a power law?

Transfer

If a power law appears, does its exponent survive a change of schedule?

Hestness et al. (2017); Kaplan et al. (2020); Hoffmann et al. (2022)

Is a scaling law universal?

Origin

When does a learning curve follow a power law?

Transfer

If a power law appears, does its exponent survive a change of schedule?

1

Existence

What are the necessary and sufficient conditions for a power-law response?

2

Control

Which macroscopic quantities turn spectrum, target, and schedule into loss?

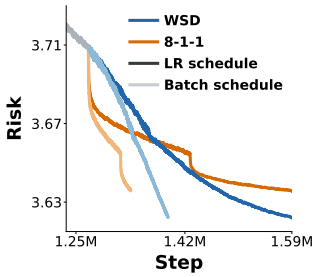
3

Prediction

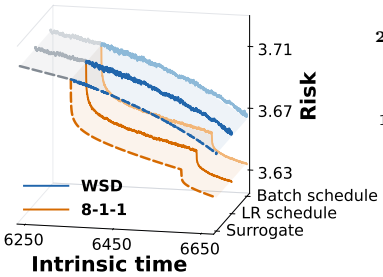
Can a response learned on one schedule predict a held-out LLM schedule?

Hestness et al. (2017); Kaplan et al. (2020); Hoffmann et al. (2022)

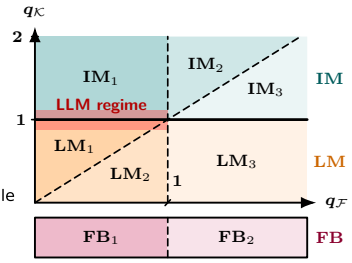
Empirical validation



(a) Loss vs. step



(b) Loss vs. intrinsic time

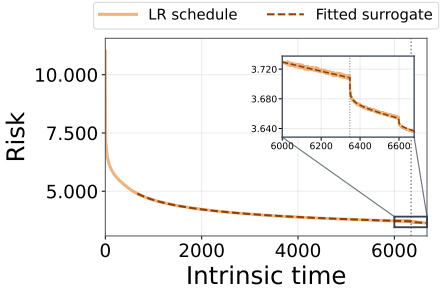


(c) $3 + 3(+2)$ map

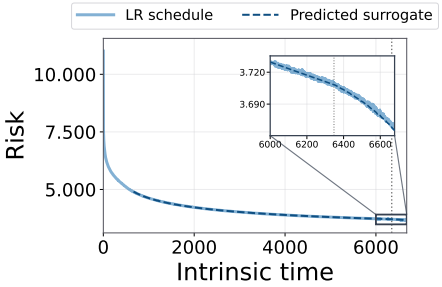
Empirical prediction

(Inspired by Li et al. Functional scaling laws. NeurIPS'25.)

- Consider a 7-parameter model (derived from our theory)
- curve fitting on 8-1-1 schedules
- predict the loss curve under WSD



(a) 8-1-1 fit



(b) WSD prediction

A technical science view

- Why our theory can predict?
- How to derive our theory?
- What is the macro variables of the loss?
- What is the sufficient and necessary condition of power law?



Existence: why can sums of exponentially decaying modes become power laws?

Control: why do $T = \sum \eta$ and $r = B/\eta$ appear?

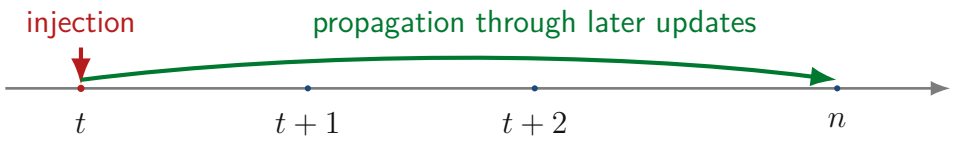
Prediction: when does a schedule preserve, change, or destroy the clean law?

SGD: one noise injection can affect future loss

One stochastic update

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \boldsymbol{\xi}_t, \qquad \mathrm{Cov}_t(\boldsymbol{\xi}_t) = \frac{\boldsymbol{\Sigma}(\boldsymbol{\theta}_t)}{B_t}.$$

For nonlinear models, this response is a small-noise approximation.



What one injection contributes

- **Variance of injection:** $\mathrm{Cov}_t(-\eta_t \boldsymbol{\xi}_t) = (\eta_t^2 / B_t)\boldsymbol{\Sigma}(\boldsymbol{\theta}_t)$.
- **Accumulated history:** Present loss depends on the full history of injections.

Measure progress, then noise per unit progress

1. Why intrinsic time?

(Li et al. Functional scaling laws. NeurIPS'25.)

Step count ignores differences in the learning rate.

$$\Delta T_s = \eta_s, \qquad T_t = \sum_{s<t} \eta_s.$$

Equal values of T represent equal first-order optimization progress.

2. Why the ratio?

Per-step noise variance scales as η_s^2/B_s. Per unit intrinsic time,

$$\frac{\eta_s^2/B_s}{\Delta T_s} = \frac{\eta_s}{B_s} = \frac{1}{r_s}, \qquad r_s := \frac{B_s}{\eta_s}.$$

Hence r_s is the inverse noise intensity per unit progress.

different update counts; same total intrinsic time



Two schedule coordinates: T measures optimization progress, while $r(T)$ controls the accumulation of stochastic error.

Theory of proxy model

A proxy that isolates the mechanism

Frozen random features

$$f_{\boldsymbol{a}}(\boldsymbol{x}) = \langle \boldsymbol{W}^\top \boldsymbol{x}, \boldsymbol{a} \rangle, \qquad y = \langle \boldsymbol{x}, \boldsymbol{\theta}^\star \rangle + \varepsilon.$$

- $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_d)$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d) \succ \boldsymbol{0}$
- Only $\boldsymbol{a}$ is trained. The representation $\boldsymbol{W}$ is fixed.

Noisy online SGD

$$\boldsymbol{a}_{t+1} = \boldsymbol{a}_t - \frac{\eta_t}{B_t} \sum_{i=1}^{B_t} \boldsymbol{W}^\top \boldsymbol{x}_t^i \big(f_{\boldsymbol{a}_t}(\boldsymbol{x}_t^i) - y_t^i \big).$$

Rahimi and Recht (2007); Liu et al. (2021); Mei et al. (2022); Paquette et al. (2024)

Why this proxy?

It separates four effects that evolve together in an LLM:

- spectrum and representation;
- target alignment;
- stochastic injection;
- joint scheduling.

response mechanism
vs.
literal equivalence

From prediction risk to spectral filtering (extension of Paquette et al. (2024))

1. Decompose the risk into spectral modes.

$$\widehat{\boldsymbol{H}} := \boldsymbol{\Lambda}^{1/2}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\Lambda}^{1/2}, \qquad \widehat{\boldsymbol{H}}\widehat{\boldsymbol{u}}_j = \widehat{\lambda}_j\widehat{\boldsymbol{u}}_j,$$

$$\boldsymbol{e}_t := \boldsymbol{\Lambda}^{1/2}(\boldsymbol{W}\boldsymbol{a}_t - \boldsymbol{\theta}^\star), \qquad R_{\sigma,t} := \mathbb{E}[\|\boldsymbol{e}_t\|_2^2|\boldsymbol{W}] = \sum_j \underbrace{\mathbb{E}\big|\langle\widehat{\boldsymbol{u}}_j, \boldsymbol{e}_t\rangle\big|^2}_{\mathcal{E}_{j,t}},$$

2. Propagate each modal energy by one SGD step. For constant (η, B),

$$\mathcal{E}_{j,t+1} = \underbrace{q_\eta(\widehat{\lambda}_j)\mathcal{E}_{j,t}}_{\text{free propagation of old error}} + \underbrace{\frac{\eta^2}{B}\widehat{\lambda}_j^2\big(R_{\sigma,t} + \sigma^2\big)}_{\text{feedback and noise injection}}$$

$$q_\eta(\lambda) = \underbrace{(1-\eta\lambda)^2}_{\text{full-batch contraction}} + \underbrace{\frac{\eta^2\lambda^2}{B}}_{\text{mini-batch fluctuation}}$$

Volterra equation

$$R_{\sigma,t} = \underbrace{F_{\boldsymbol{W}}(t)}_{\text{unresolved target error}} + \sum_{s=0}^{t-1} \underbrace{K_{\boldsymbol{W}}(t-1-s)}_{\text{survival of one injection}} \Big(\underbrace{R_{\sigma,s}}_{\text{feedback}} + \underbrace{\sigma^2}_{\text{label noise}} \Big).$$

Forcing F
Propagates the initial target error without future stochastic feedback.

Memory K
The impulse response of a single SGD-variance injection.

Observed loss R_σ
Accumulates all past injections and feeds earlier risk back into the dynamics.

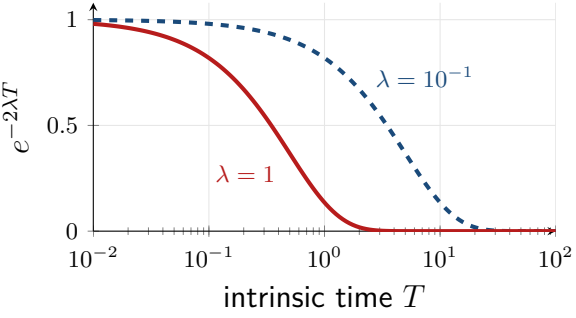
For eigenpairs $(\widehat{\lambda}_j, \widehat{\boldsymbol{u}}_j)$,

$$F_{\boldsymbol{W}}(t) = \sum_j \left| \left\langle \widehat{\boldsymbol{u}}_j, \boldsymbol{\Lambda}^{1/2} \boldsymbol{\theta}^\star \right\rangle \right|^2 q_\eta(\widehat{\lambda}_j)^t, \qquad K_{\boldsymbol{W}}(t) = \frac{\eta^2}{B} \sum_j \widehat{\lambda}_j^2 q_\eta(\widehat{\lambda}_j)^t.$$

The power law of F or K alone is not yet a power law for the total loss.

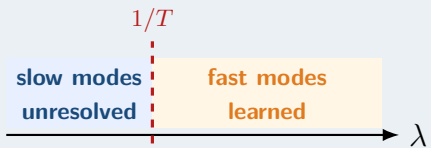
Every mode is exponential, but the active scale moves as $1/T$

$$R(T) \approx \sum_j R_j(0)\, q_\eta(\lambda_j)^t \approx \sum_j R_j(0)\, e^{-2\lambda_j T}, \qquad T = \eta t.$$



A moving spectral cutoff

$$e^{-2T\lambda} \approx \mathbf{1}\{\lambda \lesssim T^{-1}\}.$$



$\lambda T \gg 1$: mostly learned, $\lambda T \lesssim 1$: still unresolved.

As T increases, the cutoff moves toward zero: later loss probes progressively smaller eigenvalues.

Weighted spectra characterize power laws (iff criterion)

Forcing: target-weighted mass

$$\nu^{\mathcal{F}}_{\boldsymbol{W}}((0,x]) = \sum_{0<\widehat{\lambda}_j \le x} \left|\langle \widehat{\boldsymbol{u}}_j, \boldsymbol{\Lambda}^{1/2}\boldsymbol{\theta}^\star\rangle\right|^2.$$

Target energy in directions that take at least order x^{-1} time to learn.

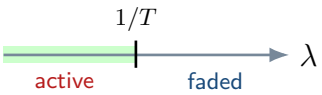
$$\nu^{\mathcal{F}}_{\boldsymbol{W}}((0,x]) \propto x^{q_{\mathcal{F}}} \quad \Longleftrightarrow \quad F_{\boldsymbol{W},>0}(t) \propto T^{-q_{\mathcal{F}}}$$

Memory: squared-spectrum mass

$$\nu^{\mathcal{K}}_{\boldsymbol{W}}((0,x]) = \sum_{0<\widehat{\lambda}_j \le x} \widehat{\lambda}_j^2.$$

Noise-injection weight carried by these directions.

$$\nu^{\mathcal{K}}_{\boldsymbol{W}}((0,x]) \propto x^{q_{\mathcal{K}}} \quad \Longleftrightarrow \quad \frac{B}{\eta^2} K_{\boldsymbol{W}}(t) \propto T^{-q_{\mathcal{K}}}.$$



Intuition: a moving spectral cutoff. Directions with $\lambda \lesssim 1/T$ still contribute:

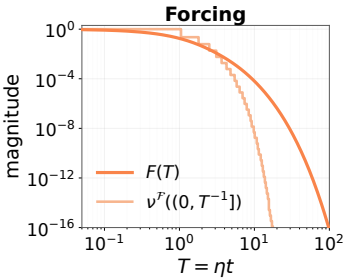
$$\nu((0,x]) \propto x^q, \qquad x \asymp T^{-1} \quad \Longrightarrow \quad T^{-q}.$$

Same cutoff, different weights: forcing and memory can have different decay exponents.

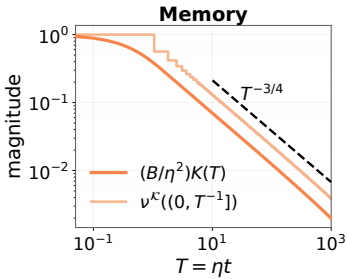
Power-law data (Zipf laws) are neither sufficient nor necessary

The spectrum has slow directions, but the target barely uses them.

$$\lambda_j = j^{-0.8}, \qquad |\theta_j^\star|^2 = e^{-j}.$$

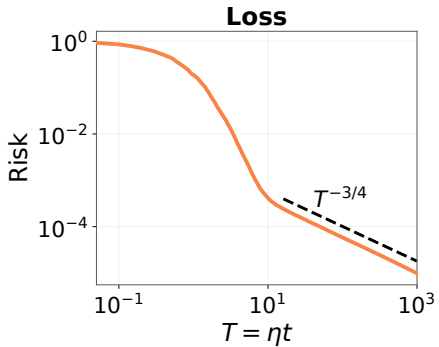


Forcing: faster than every inverse power.



Memory: $T^{-3/4}$.

A total-loss exponent does not identify its mechanism



What sets the tail?

- **Forcing.** Faster than every power law.
- **Memory.** Decays as $T^{-3/4}$.
- **Late SGD loss.** Follows the memory tail.

A fitted loss exponent does not identify what is still being learned.

The loss curve is many-to-one: it cannot reveal the underlying mechanism by itself.

Power-law random features model

The canonical PLRF model makes the cutoff calculation explicit

Assume the microscopic power laws

$$\lambda_j = j^{-2\alpha}, \qquad \theta_j^\star = j^{-\beta}, \qquad 2\alpha + 2\beta > 1.$$

The moving cutoff $\lambda_{j_T} \asymp T^{-1}$ gives $j_T \asymp T^{1/(2\alpha)}$. Summing the unresolved tail yields

$$F(T) \approx \sum_{j \gtrsim j_T} j^{-2(\alpha+\beta)} \propto T^{-q_{\mathcal{F}}}$$

$$\boxed{q_{\mathcal{F}} = \frac{2\alpha + 2\beta - 1}{2\alpha}}$$

Spectrum *and* target alignment

$$K(T) \approx \sum_{j \gtrsim j_T} j^{-4\alpha} \propto T^{-q_{\mathcal{K}}}$$

$$\boxed{q_{\mathcal{K}} = 2 - \frac{1}{2\alpha}}$$

Spectrum alone

The condition $2\alpha + 2\beta > 1$ ensures finite target energy: $\sum_j \lambda_j |\theta_j^\star|^2 < \infty$.

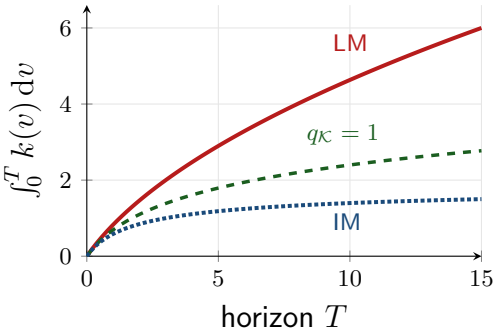
Why use $(q_{\mathcal{F}}, q_{\mathcal{K}})$, rather than only (α, β)?

	(α, β)	$(q_{\mathcal{F}}, q_{\mathcal{K}})$
Meaning	microscopic spectrum and source parameters	decay of target forcing and one-injection memory
Where defined	canonical coordinatewise PLRF assumptions	any dynamics with the corresponding response laws

The honest interpretation

Inside PLRF, this is an algebraic change of coordinates. Outside PLRF, irregular spectra and targets may have no meaningful (α, β), while $(q_{\mathcal{F}}, q_{\mathcal{K}})$ can still describe their macroscopic response.

Three memory regimes

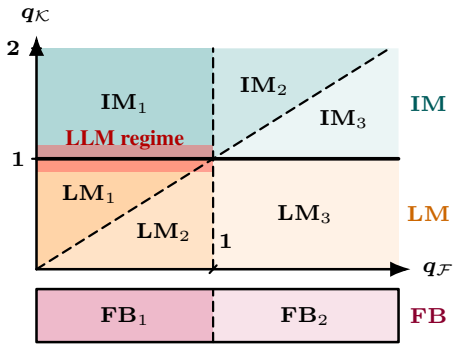


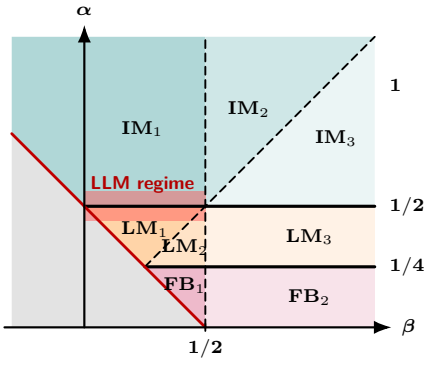
Three memory regimes

- **Long memory (LM):** $\boldsymbol{0 < q_{\mathcal{K}} < 1}$. Accumulated effect grows.
- **Marginal:** $\boldsymbol{q_{\mathcal{K}} = 1}$. $k(v) \sim c/v$: logarithmic growth.
- **Integrable memory (IM):** $\boldsymbol{q_{\mathcal{K}} > 1}$. Total memory is finite.

$\int_0^T k(v)\,\mathrm{d}v$ sums surviving injections: divergence gives LM; convergence gives IM.

The $3 + 3(+2)$ scaling law regimes





- **Long memory (LM):** $0 < q_{\mathcal{K}} < 1$ $(\frac{1}{4} < \alpha < \frac{1}{2})$
  Memory is nonintegrable; the distant past remains important.
- **Integrable memory (IM):** $q_{\mathcal{K}} > 1$ $(\alpha > \frac{1}{4})$ Total memory is finite; recent injections dominate.
- **Finite bulk (FB)** $0 < \alpha < \frac{1}{4}$ Memory stays explicitly coupled to width. $(\sum_{j \le m} \lambda_j^2 \asymp m^{1-4\alpha})$

Joint schedules reduce to two response coordinates

Optimization clock

$$T_t = \sum_{s<t} \eta_s$$

Each step advances the slow-mode filter by η_s.

Noise-control ratio

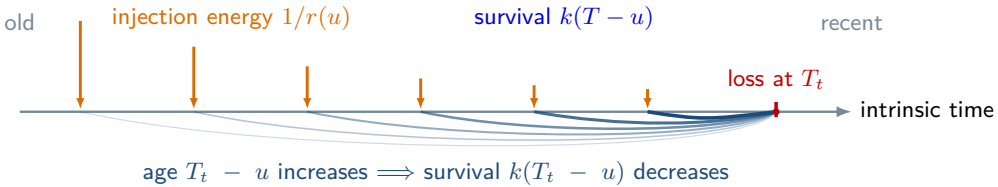
$$r_t = \frac{B_t}{\eta_t}$$

The variance injected per unit intrinsic time is proportional to $1/r_t$.

different (η_t, B_t) factorizations → same T and same path $r(T)$

The leading plain-SGD response predicts approximate factorization invariance in these two coordinates.

A schedule is an input; memory is the system response



Subtracting the clean recursion from the noisy one and passing to intrinsic time gives

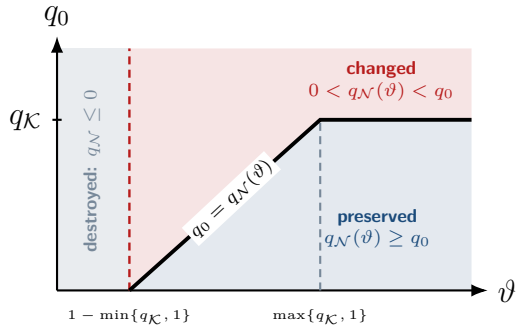
$$\Delta(T) := R_\sigma(T) - R_0(T) \asymp \sigma^2 \int_0^T \underbrace{\frac{1}{r(u)}}_{\text{injected at } u} \underbrace{k(T-u)}_{\text{survives to } T} \, \mathrm{d}u.$$

The total noise response depends on the **entire history** of the ratio path, not only its terminal value.

A schedule can preserve, change, or destroy the clean scaling law

$$k(v) \sim v^{-q_{\mathcal{K}}}, \qquad r(T) \sim T^{\vartheta} \quad \Longrightarrow \quad \Delta(T) := R_\sigma(T) - R_0(T) \asymp T^{-q_{\mathcal{N}}(\vartheta)}.$$

$$q_{\mathcal{N}}(\vartheta) = \begin{cases} \min\{\vartheta + q_{\mathcal{K}} - 1, q_{\mathcal{K}}\}, & \mathrm{LM}(0 < q_{\mathcal{K}} < 1), \\ \min\{\vartheta, q_{\mathcal{K}}\}, & \mathrm{IM}(q_{\mathcal{K}} > 1), \end{cases}$$



Noise response and observed exponent

$$R_0(T) - R_{\mathrm{app}} \sim c_0 T^{-q_0},$$
$$R_\sigma(T) - R_{\mathrm{app}} \asymp T^{-q_{\mathrm{obs}}(\vartheta)},$$
$$q_{\mathrm{obs}}(\vartheta) = \min\{q_0, q_{\mathcal{N}}(\vartheta)\}.$$

At the boundary $q_{\mathcal{K}} = 1$, logarithmic corrections may appear.

The schedule response determines the optimal data law

Since $B_s = r(T_s)\eta_s$, the data consumed by a ratio schedule satisfies

$$D = \sum_{s<t} B_s = \sum_{s<t} r(T_s)\eta_s \approx \int_0^T r(u)\,\mathrm{d}u.$$

Thus stronger noise control also uses more data.

Balance unresolved signal and budget-limited noise

PLRF specialization: $\lambda_j = j^{-2\alpha}$ and $p := 2\alpha + 2\beta - 1 > 0$.

$$\mathcal{R}_\sigma(T; D) \asymp \underbrace{T^{-p/(2\alpha)}}_{\text{Unresolved signal}} + \underbrace{D^{-1}T^{1/(2\alpha)}}_{\text{Budget-limited noise}}.$$

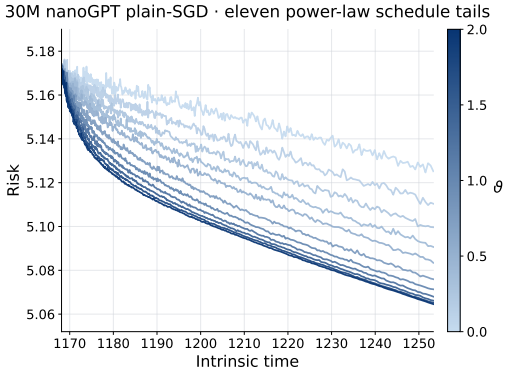
Balancing them gives

$$T^\star \asymp D^{2\alpha/(p+1)} \qquad \Longrightarrow \qquad \mathcal{R}_\sigma^\star(D) \asymp D^{-p/(p+1)}$$

Experiments

Intrinsic time is insufficient

- The same intrinsic time
- Different optimizer steps and tokens

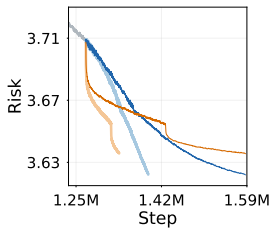


Matched paths decided by intrinsic time and ratio

300M / plain SGD / OWT / 6.5B tokens; Risk = validation CE

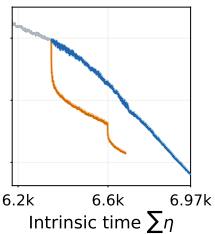
Optimizer step

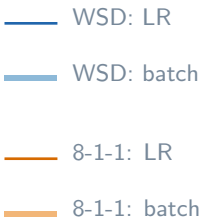
Each step advances T by η.



Intrinsic time

Same $r(T)$; each pair overlaps.





Finding. Matched $r(T)$ makes LR/batch factorizations overlap in intrinsic time.

The theory becomes a predictive forcing-memory surrogate

$$\widehat{L}(T) = \underbrace{L_\infty}_{\text{floor}} + \underbrace{A_{\mathcal{F}}(1+T)^{-q_{\mathcal{F}}}}_{\text{unlearned signal}} + \underbrace{\int_0^T \frac{A_0 + A_1(1+u)^{-q_{\mathcal{F}}}}{r(u)} \big(1 + c_{\mathcal{K}}(T-u)\big)^{-q_{\mathcal{K}}}\, \mathrm{d}u}_{\text{schedule-controlled injection accumulated through memory}}.$$

- $A_{\mathcal{F}}(1+T)^{-q_{\mathcal{F}}}$: the signal part that is unfinished to learn
- $1/r(u)$: schedule-dependent noise injection at time u
- $\big(1 + c_{\mathcal{K}}(T-u)\big)^{-q_{\mathcal{K}}}$: how much the noise can survival from u to T

Fit response

Estimate seven finite-window parameters from one 8-1-1 trajectory.

Change input

Replace only the observed ratio path $r(u)$ by WSD.

Predict

Keep every fitted response parameter frozen.

Joint-schedule adaptation of the fit-then-transfer protocol in Li et al. (2025).

Independent fits stay near $q_{\mathcal{K}} \approx 1$

Independent seven-parameter fits

Each dataset/model is trained and fitted independently.

Dataset / model	Tokens	$q_{\mathcal{K}}$	$q_{\mathcal{F}}$
OpenWebText / 124M	2.5B	1.017	0.374
FineWeb / 124M	2.5B	0.952	0.405
peS2o V2 / 124M	2.5B	0.995	0.365
OpenWebText / 300M	6.5B	0.989	0.291

Finding. Across datasets and model scales, the fitted exponents satisfy $q_{\mathcal{K}} \approx 1$ and $q_{\mathcal{F}} < 1$, placing the tested LLMs near the $\mathrm{LM}_1/\mathrm{IM}_1$ boundary of the phase map.

A controlled test in 300M nanoGPT

Held fixed

- model and initialization;
- mature checkpoint and future data tape;
- ratio path $r(T)$ and terminal intrinsic time.

Changed

- the factorization of the same path into B_t and η_t.

Prescribed ratio path $r(T)$

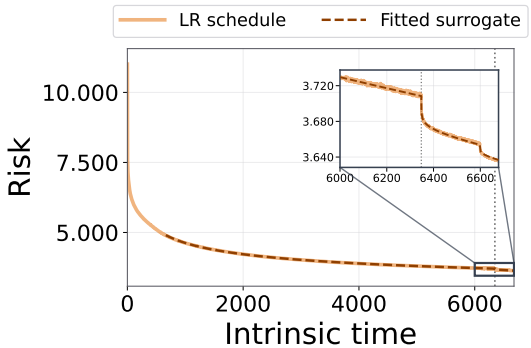
Fixed batch	Fixed learning rate
$B_t = B_0$	$\eta_t = \eta_0$
$\eta_t = B_0 / r(T_t)$	$B_t = \eta_0 r(T_t)$

Prediction: the two runs can differ in optimizer step, but should align in the correct intrinsic clock.

300M nanoGPT, plain SGD, OpenWebText, 6.5B tokens; WSD and 8-1-1 ratio paths.

Learn the response from 8-1-1 only

300M / plain SGD / OpenWebText / 6.5B tokens / $q_{\mathcal{K}} = 1$ fixed

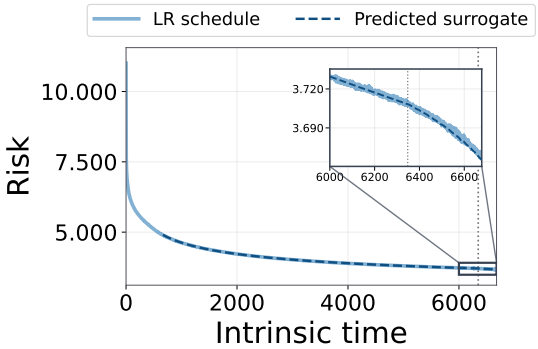


Fit protocol

- **Training curve.** Fit six parameters to 8-1-1 validation loss only.
- **Solid.** Validation loss.
- **Dashed.** Fitted surrogate.
- **Held out.** No WSD loss enters the fit.

The surrogate predicts WSD

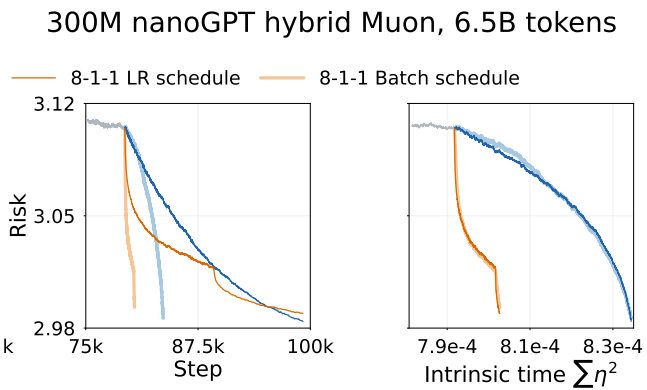
300M / plain SGD / OpenWebText / 6.5B tokens / $q_{\mathcal{K}} = 1$ fixed



Transfer protocol

- **No refitting.** Replace only $r(u)$; all response parameters stay frozen.
- **Solid.** Validation loss.
- **Dashed.** Fitted surrogate.
- **Training data.** Parameters come only from 8-1-1 in the same 300M experiment.

Go beyond SGD: Muon suggests optimizer-specific coordinates



A separate short calibration selects

$$\widetilde{T}_t = \sum_{s<t} \eta_s^2, \qquad \widetilde{r}_t = \frac{B_t}{\eta_t^2}.$$

Matched factorizations nearly collapse in this empirical clock.

Interpretation: the correct schedule coordinates may depend on the optimizer's effective homogeneity.

Takeaway messages

Existence	F and K follow power laws exactly when their cumulative weighted low-spectrum masses have the corresponding scaling.
Control	The response exponents $(q_{\mathcal{F}}, q_{\mathcal{K}})$, intrinsic clock $T = \sum \eta$, and ratio path $r = B/\eta$ determine how clean learning competes with accumulated stochastic memory.
Prediction	These coordinates organize matched nanoGPT runs, and one fitted forcing-memory response predicts unseen schedules without refitting.

- Yichen Wang, Fanghui Liu, Yudong Chen. From Spectra to Joint Schedules in LLM Pre-training: 3+3+(2) Scaling-Law Regimes. https://arxiv.org/abs/2609.40148
- Code and configurations: https://github.com/yichenblue/spectra-to-schedules-in-pretraining

Go beyond: Deep Learning (Theory) as Technical Science

Core task: Problem Formulation and Modeling

Start from common problems and build effective descriptions.

Methods: Analysis, Computation, Experiment

Inspire, test, and refine by each other.

Knowledge: General Laws and Their Scope

Explain mechanisms, characterize limits and response laws.

Goal: Advance Understanding, Guide Design

Develop testable predictions and design judgments.

> The concern is not mathematics itself, but the disconnection between theory from its motivating problems and empirical validation.

Q & A

fanghui.liu@sjtu.edu.cn

Selected references

1. Kaplan et al. (2020). *Scaling Laws for Neural Language Models*.
2. Hoffmann et al. (2022). *Training Compute-Optimal Large Language Models*.
3. Paquette et al. (2024). *4+3 Phases of Compute-Optimal Neural Scaling Laws*.
4. Li et al. (2025). *Functional Scaling Laws in Kernel Regression: Loss Dynamics and Learning Rate Schedules*.
5. Karpathy (2022). *nanoGPT*.
6. Jordan et al. (2024). *Muon: An Optimizer for Hidden Layers in Neural Networks*.

Backup

What does random matrix theory do?

1. **Encode the spectral dynamics by the resolvent.**
   Both forcing and memory are spectral functions of $\widehat{\boldsymbol{H}}$. For an analytic g,

$$\boldsymbol{G}_m(z) := (\widehat{\boldsymbol{H}} - z\boldsymbol{I})^{-1}, \qquad g(\widehat{\boldsymbol{H}}) = -\frac{1}{2\pi i}\oint_{\mathcal{C}} g(z)\boldsymbol{G}_m(z)\,\mathrm{d}z.$$

2. **Replace the random resolvent by a deterministic equivalent.**
   For the required bilinear and trace observables, RMT gives

$$\boldsymbol{G}_m(z) \;\rightsquigarrow\; \mathscr{R}_m\big(z;\mathfrak{m}(z)\big),$$

   where one scalar fixed point $\mathfrak{m}(z)$ captures the effect of the random projection.

3. **Read the learning laws from low frequency.**
   Expanding the deterministic equivalent near $z = 0$ gives

$$z \longrightarrow 0 \quad\Longrightarrow\quad \big(R_{\mathrm{app}}, q_{\mathcal{F}}, q_{\mathcal{K}}\big) \quad\Longrightarrow\quad 3 + 3(+2) \text{ regimes.}$$

The low spectrum generates the power law. RMT computes how random features modify it.

Backup: the one-step filter and exact components

For a constant schedule (η, B), Gaussian covariates give the one-step second-moment factor

$$q_\eta(\lambda) = 1 - 2\eta\lambda + \left(1 + \frac{1}{B}\right)\eta^2\lambda^2.$$

For eigenpairs $(\widehat{\lambda}_j, \widehat{\boldsymbol{u}}_j)$,

$$F_{\boldsymbol{W}}(t) = \sum_j \left|\left\langle \widehat{\boldsymbol{u}}_j, \boldsymbol{\Lambda}^{1/2}\boldsymbol{\theta}^\star \right\rangle\right|^2 q_\eta(\widehat{\lambda}_j)^t,$$

$$K_{\boldsymbol{W}}(t) = \frac{\eta^2}{B}\sum_j \widehat{\lambda}_j^2 q_\eta(\widehat{\lambda}_j)^t.$$

On the long-time scale $\lambda = O(T^{-1})$,

$$q_\eta(\lambda)^t \approx e^{-2T\lambda}, \qquad T = \eta t.$$

Backup: what the full spectral theorem adds

The informal criterion on the main slide suppresses four technical ingredients:

- joint width-time limits with $m, t \to \infty$;
- uniform control over the moving spectral window $\lambda \asymp T^{-1}$;
- stability and no escape of mass outside that window;
- slowly varying factors and exact Gamma constants.

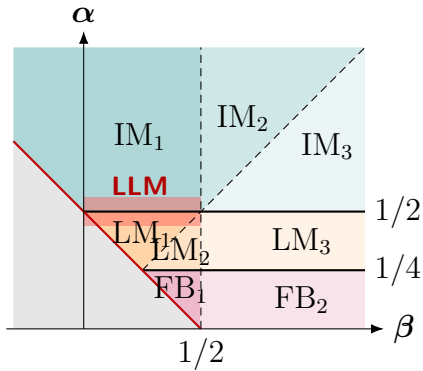
In regular-variation form, schematically,

$$A_{\mathcal{F}}(x) \sim x^{q_{\mathcal{F}}} L_{\mathcal{F}}(1/x) \quad \Longleftrightarrow \quad F_{>0}(T) \sim \Gamma(q_{\mathcal{F}}+1)(2T)^{-q_{\mathcal{F}}} L_{\mathcal{F}}(T),$$

with the analogous statement for memory.

The slowly varying factor prevents a constant-factor statement from hiding large oscillations, but it is not needed for the main conceptual message.

Backup: PLRF phase map in microscopic coordinates



$$q_{\mathcal{F}} = \frac{2\alpha + 2\beta - 1}{2\alpha}, \qquad q_{\mathcal{K}} = 2 - \frac{1}{2\alpha}.$$

- $\alpha = 1/4$: finite bulk to decaying memory.
- $\alpha = 1/2$: long to integrable memory.
- red line: finite target-energy boundary.

Backup: finite-bulk memory remains width dependent

For $0 < \alpha < 1/4$, the unresolved squared spectral mass grows with width:

$$\sum_{j \le m} \lambda_j^2 \asymp m^{1-4\alpha}.$$

The noisy-clean gap therefore has the finite-bulk form

$$\boxed{R_{\sigma,t} - R_{0,t} \asymp \sigma^2 m^{1-4\alpha} \sum_{s<t} \frac{\eta_s^2}{B_s} = \sigma^2 m^{1-4\alpha} \int_0^{T_t} \frac{du}{r(u)}.}$$

This width coupling is why FB is kept outside the width-independent LM/IM response map.

Backup: the optimal ratio allocates data where noise matters

For a terminal horizon T and data budget D, the LM/IM noise term is minimized under $\int_0^T r(u)\,du = D$ by

$$\boxed{r^\star(u) \propto \sqrt{k(T-u)\,[F(u)+\sigma^2]}.}$$

The schedule spends more samples when an injection is

- large at the time it is created, and
- likely to survive until the terminal horizon.

A representative balance is

$$\mathcal{R}_\sigma(T) \asymp T^{-p/(2\alpha)} + D^{-1}T^{1/(2\alpha)}, \qquad p = 2\alpha + 2\beta - 1,$$

which gives

$$T^\star \asymp D^{2\alpha/(1+p)}, \qquad \mathcal{R}_\sigma^\star(D) \asymp D^{-p/(1+p)}.$$

Backup: exact 300M matched-factorization protocol

Each macro cell advances intrinsic time by $\Delta T = 0.005$. For integer g, the two implementations are

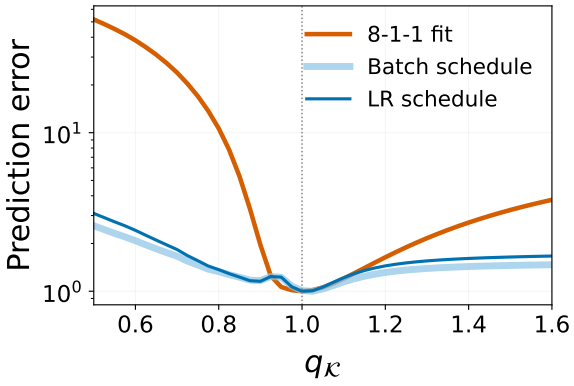
$$\begin{aligned} &\text{fixed batch:} && g \text{ updates}, && B = 16, && \eta = 0.005/g, \\ &\text{fixed learning rate:} && 1 \text{ update}, && B = 16g, && \eta = 0.005. \end{aligned}$$

Both consume exactly $16g$ contexts from the same contiguous tape segment, advance T by 0.005, and realize

$$\frac{B}{\eta} = 3200g.$$

Thus the match is exact at macro boundaries even though the optimizer-update counts differ.

Transfer error identifies an effective memory coordinate



For each fixed $q_{\mathcal{K}}$:

1. fit the other six parameters on 8-1-1;
2. freeze every parameter;
3. measure WSD prediction error.

The held-out error is minimized near

$$\boxed{q_{\mathcal{K}} \approx 1}.$$

Scope. This is a finite-window transfer-sensitivity coordinate, not an independent measurement of an asymptotic LLM spectrum.

The resolvent links risk to random matrix theory

So far, the loss dynamics are exact conditional on $\boldsymbol{W}$. To predict their large-width behavior, write

$$\widehat{\boldsymbol{H}} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\Lambda}^{1/2}, \qquad \boldsymbol{v} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\theta}^\star.$$

Both dynamical components are spectral functions of $\widehat{\boldsymbol{H}}$:

$$F_{\boldsymbol{W}}(t) = \boldsymbol{v}^\top q_\eta(\widehat{\boldsymbol{H}})^t \boldsymbol{v}, \qquad \frac{B}{\eta^2} K_{\boldsymbol{W}}(t) = \operatorname{tr}\left[\widehat{\boldsymbol{H}}^2 q_\eta(\widehat{\boldsymbol{H}})^t\right].$$

The resolvent

$$\boldsymbol{G}_m(z) := (\widehat{\boldsymbol{H}} - z\boldsymbol{I})^{-1}$$

turns any analytic spectral filter f into

$$\boldsymbol{v}^\top f(\widehat{\boldsymbol{H}})\boldsymbol{v} = -\frac{1}{2\pi\mathrm{i}} \oint_{\mathscr{C}} f(z)\, \boldsymbol{v}^\top \boldsymbol{G}_m(z)\boldsymbol{v}\, \mathrm{d}z,$$

$$\operatorname{tr} f(\widehat{\boldsymbol{H}}) = -\frac{1}{2\pi\mathrm{i}} \oint_{\mathscr{C}} f(z) \operatorname{tr} \boldsymbol{G}_m(z)\, \mathrm{d}z.$$

One resolvent controls the target-weighted forcing and the trace-weighted memory.

A scalar fixed point replaces the random projection

For the bilinear and trace observables above, random matrix theory gives

$$\boldsymbol{v}^\top \boldsymbol{G}_m(z)\boldsymbol{v} \simeq \boldsymbol{v}^\top \mathscr{R}_m(z)\boldsymbol{v}, \qquad \operatorname{tr} \boldsymbol{G}_m(z) \simeq \operatorname{tr} \mathscr{R}_m(z),$$

where

$$\mathscr{R}_m(z) = \operatorname{diag}\left(\frac{1}{\lambda_j \mathfrak{m}(z) - z}\right)_{j=1}^{d}, \qquad \mathfrak{m}(z) = \left(1 + \frac{1}{m}\sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j \mathfrak{m}(z) - z}\right)^{-1}.$$

The analytic branch is selected by $\mathfrak{m}(z) \to 1$ as $|z| \to \infty$.

Its low-frequency structure separates

$$F(t,m) = -\frac{1}{2\pi i}\oint_{\mathcal{C}} q_\eta(z)^t \boldsymbol{v}^\top \mathscr{R}_m(z)\boldsymbol{v}\,\mathrm{d}z = \underbrace{F_0(m)}_{\text{null-space floor}} + \underbrace{F_{pp}(t,m)}_{\text{unresolved target modes}} + \underbrace{F_{ac}(t,m)}_{\text{finite-width distortion}}.$$

The cutoff gives the exponents; RMT adds the floor and finite-width corrections needed for the full phase map.

The pure-point spectrum gives the temporal power law

The pure-point component propagates target energy aligned with the population spectral directions:

$$F_{pp}(T,m) \asymp \sum_j \lambda_j (\theta_j^\star)^2 e^{-2T\lambda_j}.$$

For the canonical PLRF model,

$$\lambda_j = j^{-2\alpha}, \qquad \theta_j^\star = j^{-\beta},$$

and hence

$$F_{pp}(T,m) \asymp \sum_j j^{-2(\alpha+\beta)} e^{-2Tj^{-2\alpha}}.$$

Using $x = j^{-2\alpha}$, this becomes

$$F_{pp}(T,m) \asymp \frac{1}{2\alpha} \int_0^1 x^{q_{\mathcal{F}}-1} e^{-2Tx}\,\mathrm{d}x \asymp T^{-q_{\mathcal{F}}}, \qquad q_{\mathcal{F}} := \frac{2\alpha+2\beta-1}{2\alpha}.$$

Equivalently, the active modes satisfy $j \gtrsim j_T \asymp T^{1/(2\alpha)}$, and summing their remaining target energy gives the same exponent.

The temporal exponent comes from unresolved target energy at the moving spectral cutoff.

The zero atom gives the finite-width floor

The null-space contribution is the residue of the DE resolvent at zero:

$$F_0(m) = -\operatorname*{Res}_{z=0} \boldsymbol{v}^\top \mathscr{R}_m(z)\boldsymbol{v}.$$

Near zero, the scalar fixed point satisfies

$$\mathfrak{m}(z) \sim -\gamma_m z, \qquad \frac{1}{m}\sum_{j=1}^{d} \frac{\gamma_m \lambda_j}{1+\gamma_m\lambda_j} = 1.$$

For $\lambda_j = j^{-2\alpha}$, this gives

$$\gamma_m \asymp m^{2\alpha}, \qquad F_0(m) \asymp \sum_j \frac{\lambda_j (\theta_j^\star)^2}{1+\gamma_m\lambda_j}.$$

Separating directions above and below the width cutoff,

$$F_0(m) \asymp m^{-2\alpha} \sum_{j \lesssim m} j^{-2\beta} + \sum_{j \gtrsim m} j^{-2(\alpha+\beta)} \asymp \begin{cases} m^{-p} = m^{-2\alpha q_{\mathcal{F}}}, & \beta < \frac{1}{2}, \\ m^{-2\alpha}, & \beta > \frac{1}{2}, \end{cases}$$

where $p := 2\alpha + 2\beta - 1$.

The continuous spectrum gives finite-width distortion

The branch cut of $\mathfrak{m}(z)$ generates an absolutely continuous spectral component. By Stieltjes inversion,

$$\mu^{\mathcal{F}}_{m,ac}(\mathrm{d}x) = \frac{1}{\pi}\,\mathrm{Im}\big[\boldsymbol{v}^\top \mathscr{R}_m(x+\mathrm{i}0)\boldsymbol{v}\big]\,\mathrm{d}x.$$

For a smooth target, $\beta > \frac{1}{2}$, the corresponding forcing has order

$$F_{ac}(T,m) \asymp \frac{1}{m}\int_{m^{-2\alpha}}^{1} x^{-1/(2\alpha)} e^{-2Tx}\,\mathrm{d}x.$$

Its behavior changes at $\alpha = \frac{1}{2}$:

$$F_{ac}(T,m) \asymp \begin{cases} O(m^{-2\alpha}), & \alpha < \frac{1}{2}, \\ m^{-1}T^{-1+1/(2\alpha)} = m^{-1}T^{-(q_{\mathcal{K}}-1)}, & \alpha > \frac{1}{2}, \end{cases}$$

where

$$q_{\mathcal{K}} := 2 - \frac{1}{2\alpha}.$$

For $\alpha < 1/2$, this contribution is absorbed into F_0. For $\alpha > 1/2$, it becomes a distinct time-dependent correction.

The three spectral pieces yield the phasewise forcing

Combining the null-space, pure-point, and continuous components gives

$$F(T,m) = F_0(m) + F_{pp}(T,m) + F_{ac}(T,m).$$

Away from the critical lines, and for $1 \ll T \le C_0 m^{2\alpha}$,

$$F(T,m) \asymp \begin{cases} m^{-2\alpha q_{\mathcal{F}}} + T^{-q_{\mathcal{F}}}, & \mathrm{FB}_1, \mathrm{LM}_1, \mathrm{LM}_2, \mathrm{IM}_1, \\ m^{-2\alpha} + T^{-q_{\mathcal{F}}}, & \mathrm{FB}_2, \mathrm{LM}_3, \\ m^{-2\alpha} + T^{-q_{\mathcal{F}}} + m^{-1}T^{-(q_{\mathcal{K}}-1)}, & \mathrm{IM}_2, \mathrm{IM}_3. \end{cases}$$

$$\begin{aligned} \beta < \tfrac{1}{2} \iff q_{\mathcal{F}} < 1 : &\quad \text{target-dependent width floor,} \\ \beta > \tfrac{1}{2},\ \alpha < \tfrac{1}{2} : &\quad \text{representation floor only,} \\ \beta > \tfrac{1}{2},\ \alpha > \tfrac{1}{2} : &\quad \text{additional feature-distortion transient.} \end{aligned}$$

The boundary $q_{\mathcal{F}} = q_{\mathcal{K}}$ does not change the forcing formula itself; it becomes relevant when forcing is compared with memory.

This is the forcing input to the Volterra system, not yet the observed loss.