FINITE-CORPUS DYNAMICS OF SUPERVISED AND VERIFIER-GUIDED ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-model training commonly separates pretraining and instruction-tuned from supervised learning from subsequent reinforcement learning with verifiable rewards, while the dynamical advantage of stage separation remains theoretically unresolved. We study a solvable finite-corpus attention model that combines mean-squared-error (MSE) training with Group Relative Policy Optimization (GRPO). In single-location regression, MSE jointly learns key and value representations, whereas GRPO uses exact position feedback to refine only the key. In the high-dimensional limit, we derive a dynamical mean-field theory that reduces finite-corpus training to a self-consistent single-channel process. The theory tracks learning across an abrupt MSE-to-GRPO switch and captures verifier-guided refinement after supervised learning. We further compare hard staging with exact-count random interleaving under matched supervision and verifier budgets. Paired experiments reveal schedule crossovers with label noise and sequence length. An exact endpoint decomposition and local force analysis attribute these crossovers to the competition between clean co-adaptation under interleaving and continued exposure to noisy supervised updates.

1 INTRODUCTION

Modern large-model training usually employ multi-stage training stategy. A foundation model is first pretrained or instruction-tuned from supervised data and is then refined with reinforcement-learning objectives. In recent reasoning models, reinforcement learning with verifiable rewards (RLVR), often implemented through Group Relative Policy Optimization (GRPO) or related methods, has become an effective final-stage training paradigm (Shao et al., 2024; Yu et al., 2025), while the dynamical advantage of stage separation remains theoretically unresolved.

GRPO updates a policy from groups of sampled actions and verifiable rewards, using group-relative advantages in place of a separately trained critic. Recent analyses characterize its on- and off-policy objectives (Mroueh et al., 2026). They leave open how verifier-guided updates interact with a representation learned earlier from supervised data when both objectives reuse the same finite corpus.

High-dimensional statistics and statistical physics provide a complementary framework for studying such learning dynamics. In teacher–student models, fresh-sample online SGD can often be reduced to closed dynamics of a few macroscopic overlaps (Goldt et al., 2020). Recent work has extended this perspective to attention models. High-dimensional theories have characterized linear attention and its training dynamics (Zhang et al., 2025; Arnaboldi et al., 2025), while Duranthon et al. (2026) introduced single-location regression as a tractable nonlinear softmax-attention model in which a key learns where to attend and a value learns what to read.

A fixed finite training set introduces an additional difficulty because the same example is revisited and becomes correlated with its own earlier updates. Classical restricted-training-set theory describes this effect through two-time correlations and response functions (Sollich & Barber, 1997; Coolen & Saad, 2000; Heimel & Coolen, 2001; Luo & Wong, 2003). Dynamical mean-field theory (DMFT) provides a modern high-dimensional formulation of such effective stochastic processes (Agoritsas et al., 2018; Gerbelot et al., 2024), and recent work has extended these ideas to multi-pass SGD on finite datasets (Fan & Wang, 2026). Finite-corpus attention with an explicit switch from supervised to policy-dependent training adds a further layer: an earlier perturbation changes both later forces and the probabilities of sampled GRPO action histories.

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

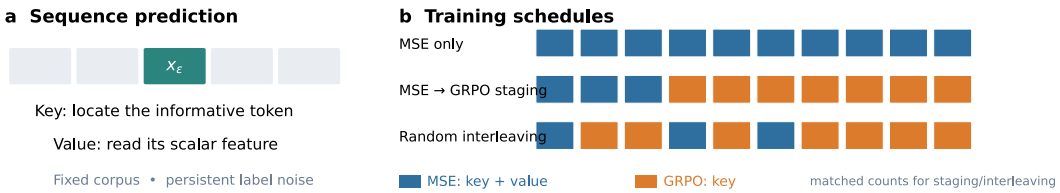


Figure 1: Prediction task and training schedules. The key localizes the informative token and the value reads its scalar feature. We study MSE-only training, hard MSE-to-GRPO staging, and exact-count random interleaving. The staged and interleaved schedules use identical MSE and GRPO update counts and verifier budgets.

We study this problem by extending single-location regression to a fixed finite corpus trained with both MSE and GRPO. MSE updates both the key and value from noisy scalar targets, whereas a position verifier supplies GRPO updates only to the key. This gives a minimal analogue of two-stage large-model training, in which supervised learning first builds the representation and verifier-guided training subsequently refines localization. In the proportional high-dimensional limit, we derive a finite-corpus DMFT that reduces the full training dynamics to a self-consistent single-channel process for one tagged sequence. Its evolution closes through teacher overlaps, two-time covariances, response kernels, and the full action-history response induced by policy-dependent sampling.

We use this framework to study both the dynamics of objective switching and the relative value of stage separation. The theory accurately describes an abrupt MSE-to-GRPO switch in the high-dimensional regime. To compare training schedules fairly, we further contrast hard staging with an exact-count random interleaving that uses the same MSE labels, GRPO feedback, and verifier budget. Paired experiments reveal systematic schedule crossovers as label noise and sequence length vary, while an exact endpoint decomposition and local force analysis identify the competing roles of clean co-adaptation and continued exposure to noisy supervised updates.

Our main contributions are:

- We derive a closed finite-corpus DMFT for supervised and verifier-guided softmax attention with policy-dependent action histories.
- We validate the theory across an abrupt MSE-to-GRPO stage switch and characterize verifier-guided refinement after supervised learning.
- We identify noise- and length-dependent crossovers between staging and matched random interleaving, together with a local mechanism analysis.

2 Learning Model

The model separates token localization from scalar prediction and uses two update rules on the same finite corpus. We first define the single-location regression task and the MSE and verifier-guided forces, then specify the three training schedules compared throughout the paper.

2.1 Sequence Regression

Each example contains L tokens in $\mathbb{R}^p$. The informative position ϵ is uniform on $\{1, \dots, L\}$, and independent background vectors satisfy $g_a \sim \mathcal{N}(0, I_p)$, where I_p is the identity matrix. Following the high-dimensional geometry of the independent teacher directions in Duranthon et al. (2026), we impose their limiting normalization exactly at finite p. Thus $k^\star$ and $v^\star$ are orthogonal and each has norm $\sqrt{p}$. With signal strength $\nu \geq 0$, the token and clean target are

$$x_a = g_a + \mathbf{1}\{a = \epsilon\}\sqrt{\nu}\, k^\star/\sqrt{p}, \qquad y_0 = x_\epsilon^\top v^\star/\sqrt{p}. \tag{1}$$

The indicator selects the informative token. The observed label is $y = y_0 + \sigma_y \xi$, with noise amplitude $\sigma_y \geq 0$ and independent $\xi \sim \mathcal{N}(0, 1)$. A corpus consists of n independent examples indexed by μ, with sample density $\alpha = n/p$. Each example retains its tokens and label perturbation throughout training.

The student key $k \in \mathbb{R}^p$ and value $v \in \mathbb{R}^p$ produce token logits $u_a = x_a^\top k/\sqrt{p}$ and readouts $\tilde{y}_a = x_a^\top v/\sqrt{p}$. Their attention policy and scalar prediction are

$$p_a(u) = \frac{e^{u_a}}{\sum_{b=1}^{L} e^{u_b}}, \qquad \widehat{y}(u, \tilde{y}) = \sum_{a=1}^{L} p_a(u)\tilde{y}_a. \tag{2}$$

The clean test risk is $\mathcal{E} = \frac{1}{2}\mathbb{E}_{\text{fresh}}(\widehat{y} - y_0)^2$ on a new example. Teacher orthogonality gives $\mathbb{E}y_0^2 = 1$, so the zero predictor has risk $1/2$. Position accuracy $\mathrm{Acc} = \mathbb{P}_{\text{fresh}}(\arg\max_a u_a = \epsilon)$ and sampled-action success $\mathbb{E}_{\text{fresh}} p_\epsilon$ measure different aspects of localization.

2.2 Supervised and verifier updates

An MSE update selects an example uniformly from the corpus and minimizes $\ell_{\mathrm{M}} = \frac{1}{2}(\widehat{y} - y)^2$. With residual $r = \widehat{y} - y$, prediction derivative $b_a = p_a(\tilde{y}_a - \widehat{y})$, and learning rate $\eta_{\mathrm{M}} > 0$, its negative-gradient token forces are

$$F^{\mathrm{M}}_{a,\mathrm{k}} = -\eta_{\mathrm{M}} r b_a, \qquad F^{\mathrm{M}}_{a,\mathrm{v}} = -\eta_{\mathrm{M}} r p_a. \tag{3}$$

Both forces use the fields at the start of the update.

A verifier-guided update selects a prompt uniformly from a fixed, independently assigned subset $\mathcal{V}$ of fraction $\beta = |\mathcal{V}|/n$. We use a single-step on-policy specialization of GRPO with the reference-policy Kullback–Leibler coefficient set to zero. The policy samples $G_{\mathrm{grp}} \geq 2$ independent positions a_g, with rewards $\rho_g = \mathbf{1}\{a_g = \epsilon\}$. The group mean is $q = G_{\mathrm{grp}}^{-1} \sum_g \rho_g$, its reward standard deviation is $s_\rho = \sqrt{q(1-q)}$, and its advantages are $\widehat{A}_g = (\rho_g - q)/(s_\rho + \varepsilon_{\mathrm{stab}})$, with stabilizer $\varepsilon_{\mathrm{stab}} \geq 0$. Homogeneous groups have zero advantages.

One ascent step on $\widehat{J} = G_{\mathrm{grp}}^{-1} \sum_g \mathrm{stopgrad}(\widehat{A}_g) \log p_{a_g}(u)$ holds the sampled actions and advantages fixed. The centered advantages cancel the common policy-mean term in the log-policy gradient. With histogram counts $c_a = \sum_g \mathbf{1}\{a_g = a\}$ and learning rate $\eta_{\mathrm{R}} > 0$, the resulting force for a mixed-reward group is

$$F^{\mathrm{R}}_{a,\mathrm{k}} = \frac{\eta_{\mathrm{R}} c_a (\mathbf{1}\{a = \epsilon\} - q)}{G_{\mathrm{grp}}(s_\rho + \varepsilon_{\mathrm{stab}})}, \qquad F^{\mathrm{R}}_{a,\mathrm{v}} = 0. \tag{4}$$

The value derivative vanishes because the position policy depends only on k. A pure-GRPO stage therefore holds the readout fixed by construction. Each update draws a fresh group and costs G_{grp} verifier calls, including zero-update groups.

2.3 Training schedules and comparison protocols

The sequence matrix $X^\mu \in \mathbb{R}^{L \times p}$ has token transposes as rows, and the student parameters at training step i are collected as $W_i = [k_i, v_i] \in \mathbb{R}^{p \times 2}$. For either objective, the token-level forces form an $L \times 2$ matrix according to

$$F^o = [F^o_{:,\mathrm{k}}, F^o_{:,\mathrm{v}}], \qquad F^o_{a,:} = (F^o_{a,\mathrm{k}}, F^o_{a,\mathrm{v}}), \qquad o \in \{\mathrm{M}, \mathrm{R}\}.$$

If μ_i denotes the selected training example and F_i the force associated with the update type, both learning rules take the common form

$$W_{i+1} = W_i + \frac{(X^{\mu_i})^\top F_i}{\sqrt{p}}, \qquad F_i \in \{F^{\mathrm{M}}, F^{\mathrm{R}}\}. \tag{5}$$

We study three schedules. **MSE-only training** applies an MSE update at every training step. It is the finite-corpus analogue of the supervised single-location regression training in Duranthon et al. (2026). **MSE-to-GRPO staging** first applies N_{M} MSE updates and then switches to N_{R} GRPO updates. **Random MSE–GRPO interleaving** uses the same N_{M} and N_{R} updates but uniformly randomizes their order. The staged and interleaved schedules use the same learning rates, verifier fraction, group size, initialization, corpus, and total verifier budget. Paired runs also share the verifier subset, objective-specific example streams, and action-sampling uniforms. This matched construction makes interleaving the principal baseline for isolating the effect of training order. Continued MSE is used separately to test the effect of introducing verifier feedback after supervised learning.

3 MEAN-FIELD DYNAMICS

The high-dimensional description couples macroscopic order parameters to the stochastic field history of one tagged training example. The exact finite-dimensional field recursion motivates a single-channel process, and self-consistency then closes the evolution of the overlap, correlation, transverse covariance, and response kernels.

3.1 ORDER PARAMETERS AND EXACT FIELD DYNAMICS

Let $W^\star = [k^\star, v^\star] \in \mathbb{R}^{p\times 2}$ denote the teacher matrix. For training steps i and j, define the teacher overlap, two-time parameter correlation, and transverse covariance by

$$M_i = \frac{W^{\star\top} W_i}{p}, \qquad C_{ij} = \frac{W_i^\top W_j}{p}, \qquad R_{ij} = C_{ij} - M_i^\top M_j. \tag{6}$$

All three quantities are 2×2 matrices over the key and value channels. We retain the full matrices because MSE training couples these two channels.

For each corpus example $\mu = 1, \dots, n$, define its local student field and teacher field by

$$H_i^\mu = \frac{X^\mu W_i}{\sqrt{p}} = [u_i^\mu, \tilde{y}_i^\mu], \qquad T^\mu = \frac{X^\mu W^\star}{\sqrt{p}}, \qquad H_i^\mu, T^\mu \in \mathbb{R}^{L\times 2}. \tag{7}$$

For two examples κ and μ, the token Gram block is $\Gamma^{\kappa\mu} = X^\kappa (X^\mu)^\top / p \in \mathbb{R}^{L\times L}$. Substituting Eq. (5) into Eq. (7) gives the exact recursion

$$H_{i+1}^\kappa = H_i^\kappa + \Gamma^{\kappa\mu_i} F_i. \tag{8}$$

For the selected example X^{μ_i}, the field $H_i^{\mu_i}$ receives an order-one update, whereas H_i^μ for $\mu \neq \mu_i$ receives fluctuations of order $p^{-1/2}$. The accumulation of these weak cross-example effects motivates the high-dimensional process below.

3.2 HIGH-DIMENSIONAL LIMIT AND SINGLE-CHANNEL PROCESS

We take $p, n \to \infty$ with $n/p \to \alpha$ at fixed L, group size, learning rates, and finite horizon in epoch-equivalent time $t = i/n$. At fixed L, the self-Gram matrix of a tagged example satisfies $\Gamma^{\mu\mu} \to I_L$, while the cross-example blocks $\Gamma^{\mu\kappa}$ for $\kappa \neq \mu$ fluctuate at order $p^{-1/2}$. The tagged example therefore experiences order-one changes when it is revisited and many weak changes through updates of the rest of the corpus.

For the numerical construction, macroscopic time is divided into windows of width h. Within a window, the reference dynamics keep the cumulative contribution from other examples fixed and apply repeated visits of the tagged example in sequence. Each visit evaluates its force at the field produced by the preceding visit. The exact visit law and finite-window recursion are given in Appendix C.2. At fixed h this defines the finite-window reference dynamics, while decreasing h tests its approach to step-by-step training.

Under the finite-time cavity and response self-averaging assumptions, the field of a tagged example μ follows the single-channel process

$$H_{\tilde{i}}^\mu = T^\mu M_{\tilde{i}} + \zeta_{\tilde{i}}^\mu + \sum_{\tilde{j}<\tilde{i}} F_{\tilde{j}}^\mu G_{\tilde{i}\tilde{j}}. \tag{9}$$

The teacher field T^μ and overlap $M_{\tilde{i}}$ are defined in Eqs. (7) and (6). The accumulated force $F_{\tilde{j}}^\mu \in \mathbb{R}^{L\times 2}$ is built from the token-force matrices in Eqs. (3) and (4). The Gaussian cavity field summarizes the remaining corpus and satisfies $\mathbb{E}[(\zeta_{\tilde{i},a}^\mu)^\top \zeta_{\tilde{j},b}^\mu] = \mathbf{1}\{a=b\} R_{\tilde{i}\tilde{j}}$. The response kernel $G_{\tilde{i}\tilde{j}} \in \mathbb{R}^{2\times 2}$ propagates an earlier parameter perturbation, with $G_{\tilde{j}+1,\tilde{j}} = I_2$ and $G_{\tilde{i}\tilde{j}} = 0$ for $\tilde{i} \leq \tilde{j}$. The final term in Eq. (9) carries the tagged example's previous updates back to its current field.

3.3 CLOSED MACROSCOPIC RECURSION

The single-channel process and the macroscopic kernels form a causal closed system. The core order parameters are the teacher overlap M, transverse covariance R, and response kernel G, while $C = M^\top M + R$ is the full correlation. The auxiliary drive covariance V, feedback kernel Σ, and initial-state propagator Q carry the Gaussian fluctuations and linear response needed to advance these quantities.

Given the feedback kernel Σ, one causal update of M, V, Q, G, R, and C from window $\tilde{i}$ to the next window is

$$\begin{aligned}
&M_{\tilde{i}+1} = M_{\tilde{i}} + \alpha\mathbb{E}[(T^\mu)^\top F^\mu_{\tilde{i}}], && V_{\tilde{i}\tilde{j}} = \alpha\mathbb{E}[(F^\mu_{\tilde{i}})^\top F^\mu_{\tilde{j}}],\\
&Q_{\tilde{i}+1} = Q_{\tilde{i}} + \sum_{\tilde{j}\le\tilde{i}} Q_{\tilde{j}}\Sigma_{\tilde{i}\tilde{j}}, && G_{\tilde{i}+1,\tilde{j}} = G_{\tilde{i}\tilde{j}} + \mathbf{1}\{\tilde{i}=\tilde{j}\}I_2 + \sum_{\tilde{\ell}=\tilde{j}+1}^{\tilde{i}} G_{\tilde{\ell}\tilde{j}}\Sigma_{\tilde{i}\tilde{\ell}},\\
&R_{\tilde{i}+1,\tilde{j}} = Q^\top_{\tilde{i}+1}R_{00}Q_{\tilde{j}} + \sum_{\tilde{\ell}\le\tilde{i},\,\tilde{m}<\tilde{j}} G^\top_{\tilde{i}+1,\tilde{\ell}}V_{\tilde{\ell}\tilde{m}}G_{\tilde{j}\tilde{m}}, && C_{\tilde{i}+1,\tilde{j}} = M^\top_{\tilde{i}+1}M_{\tilde{j}} + R_{\tilde{i}+1,\tilde{j}}.
\end{aligned} \tag{10}$$

Here $0 \le \tilde{j} \le \tilde{i}$ in the new rows of V and G, while the covariance row is reconstructed for $0 \le \tilde{j} \le \tilde{i}+1$ and completed by symmetry. We use $M_0 = 0$, $R_{00} = I_2$, $Q_0 = I_2$, $G_{\tilde{i}\tilde{j}} = 0$ for $\tilde{i} \le \tilde{j}$, and $V_{\tilde{j}\tilde{i}} = V^\top_{\tilde{i}\tilde{j}}$. The propagator Q transports the initial transverse fluctuation.

It remains to determine the feedback kernel Σ. Add a small source $\theta_{\tilde{j}} \in \mathbb{R}^{2L}$ to the vectorized non-retarded field in window $\tilde{j}$ and define the weak force response

$$A_{\tilde{i}\tilde{j}} = \frac{\partial}{\partial\theta_{\tilde{j}}}\mathbb{E}_{\theta_{\tilde{j}}}[\mathrm{vec}\, F^\mu_{\tilde{i}}]\bigg|_{\theta_{\tilde{j}}=0}, \qquad \Sigma_{\tilde{i}\tilde{j}} = \alpha\left[\sum_{a=1}^{L}(A_{\tilde{i}\tilde{j}})_{aa}\right]^\top.$$

The long-range derivative $A_{\tilde{i}\tilde{j}}$ is evaluated causally from local derivatives along the same single-channel path. With $d = 2L$, its recursion is

$$\begin{aligned}
&\Psi_{\tilde{i}\tilde{j}} = \mathbf{1}\{\tilde{i}=\tilde{j}\}I_d + \sum_{\tilde{\ell}=\tilde{j}}^{\tilde{i}-1}(I_L \otimes G^\top_{\tilde{i}\tilde{\ell}})B_{\tilde{\ell}}\Psi_{\tilde{\ell}\tilde{j}},\\
&E_{\tilde{j}\tilde{j}} = 0, && E_{\tilde{i}+1,\tilde{j}} = E_{\tilde{i}\tilde{j}} + \Psi^\top_{\tilde{i}\tilde{j}}\ell_{\tilde{i}},\\
&A_{\tilde{i}\tilde{j}} = \mathbb{E}\left[B_{\tilde{i}}\Psi_{\tilde{i}\tilde{j}} + f_{\tilde{i}}\left(E_{\tilde{i}\tilde{j}} + \Psi^\top_{\tilde{i}\tilde{j}}\ell_{\tilde{i}}\right)^\top\right].
\end{aligned} \tag{11}$$

Here $h_{\tilde{i}} = \mathrm{vec}\, H^\mu_{\tilde{i}}$ and $f_{\tilde{i}} = \mathrm{vec}\, F^\mu_{\tilde{i}}$. The matrix $B_{\tilde{i}} = \partial f_{\tilde{i}}/\partial h_{\tilde{i}}$ is the fixed-history window Jacobian, $\ell_{\tilde{i}} = \nabla_{h_{\tilde{i}}}\log P(\omega_{\tilde{i}} \mid h_{\tilde{i}})$ is the local sampling score, $\Psi_{\tilde{i}\tilde{j}}$ propagates a field perturbation, and $E_{\tilde{i}\tilde{j}}$ accumulates the scores of earlier sampled actions. For pure MSE, $\ell = 0$ and the recursion reduces to the usual Jacobian response. GRPO adds the history-probability term because earlier field perturbations change later action probabilities, including paths whose later updates are MSE. Equations (9), (10), and (11) are therefore non-Markovian but closed once the macroscopic kernels are specified. Appendix E derives the local Jacobians and scores and verifies the historical term on a finite two-window example.

3.4 POPULATION COMPUTATION AND VALIDATION

The closed recursion can be solved forward by sampling a population of single-channel paths. Each path keeps its teacher field, label perturbation, and verifier eligibility fixed while the macroscopic kernels are shared across the population. Force moments and weak responses update Eqs. (10)–(11), and conditional Gaussian sampling extends the cavity histories to the next window. Fresh examples omit the tagged-example return term and provide the clean test risk.

The numerical controls separately test population error, finite-dimensional scaling, window refinement, response propagation, and risk quadrature. Linear spectral and nonlinear one- and two-time

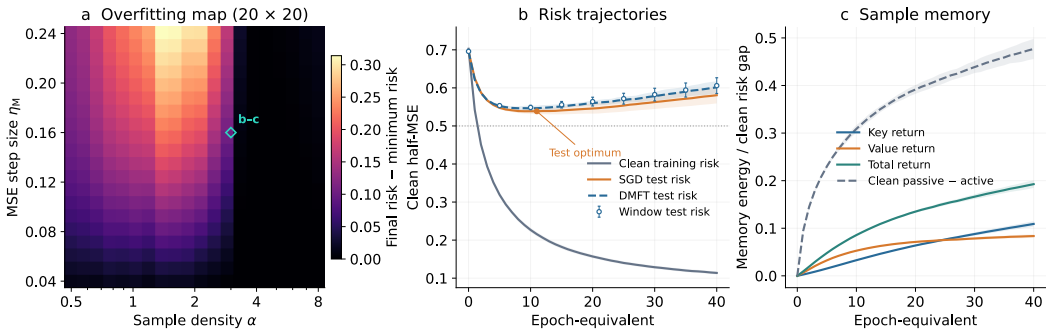


Figure 2: Finite-corpus memory and generalization. (a) Late fresh-example risk rebound across density and MSE learning rate. (b) Clean retained-corpus risk and fresh-example risks from step-by-step SGD, finite-window dynamics, and DMFT. (c) Return-field energies and the clean passive-minus-active gap Δ_{clean}.

comparisons support the closure over the validated horizons. The 64-epoch schedule scans below use step-by-step finite-dimensional training; their crossover locations are measured from those paired simulations.

4 Memory, Localization, and Objective Order

This section separates sample-specific memory from generalization, tests verifier-guided refinement after supervised learning, and compares staging with exact-count interleaving as label noise and sequence length vary.

4.1 Sample-Specific Memory and Generalization

The single-channel process separates the field of a reused example into a passive component and the return of its own earlier updates,

$$H_{\tilde{i}}^{\mu} = U_{\tilde{i}}^{\mu} + H_{\tilde{i}}^{\mu,\text{ret}}, \qquad U_{\tilde{i}}^{\mu} = T^{\mu} M_{\tilde{i}} + \zeta_{\tilde{i}}^{\mu}, \qquad H_{\tilde{i}}^{\mu,\text{ret}} = \sum_{\tilde{j}<\tilde{i}} F_{\tilde{j}}^{\mu} G_{\tilde{i}\tilde{j}}.$$

We quantify sample-specific memory by its return-field energy and distinguish it from its effect on clean prediction,

$$\begin{aligned} q_{\text{mem}}(\tilde{i}) &= \frac{1}{L}\mathbb{E}\|H_{\tilde{i}}^{\mu,\text{ret}}\|_F^2 = q_{\text{mem}}^{(\text{k})}(\tilde{i}) + q_{\text{mem}}^{(\text{v})}(\tilde{i}), \\ \Delta_{\text{clean}}(\tilde{i}) &= \mathbb{E}\Big[\ell(U_{\tilde{i}}^{\mu}; y_0^{\mu}) - \ell(U_{\tilde{i}}^{\mu} + H_{\tilde{i}}^{\mu,\text{ret}}; y_0^{\mu})\Big]. \end{aligned} \tag{12}$$

A positive Δ_{clean} means that the self-return improves prediction of the retained example's clean target.

Figure 2(a) maps the late test-risk rebound over sample density and MSE learning rate. Panel (b) then compares clean risk on the retained corpus with fresh-example risks from step-by-step SGD, the finite-window reference dynamics, and DMFT. The retained-corpus risk is evaluated against the clean teacher targets; optimization still uses the persistent noisy labels. In the selected regime, fresh-example risk decreases initially and later rises while the retained-corpus fit continues to improve.

Panel (c) shows that key and value return energies continue to grow, while Δ_{clean} remains positive. The self-return therefore remains useful on reused examples even as fresh-example performance deteriorates. Finite-corpus overfitting is therefore identified by the growing gap between sample-specific adaptation and transferable prediction; memory magnitude alone is insufficient.

4.2 Verifier-Guided Refinement After Supervised Learning

We first isolate the effect of introducing verifier feedback after supervised learning. Two runs share the same MSE trajectory up to a prescribed switch. The staged run then uses pure GRPO, whereas the

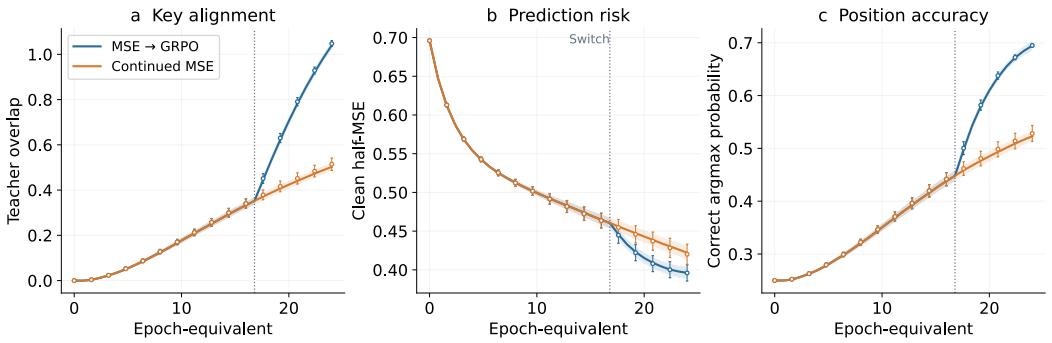


Figure 3: Verifier-guided refinement after supervised learning. Both runs share the same MSE trajectory before the switch; the staged run then uses pure GRPO, while the control continues MSE. The panels show key alignment, clean test half-MSE, and position accuracy. DMFT lines and finite-window results show consistent dynamics across the objective switch.

control continues MSE. Their representations are therefore identical at the switch. This comparison differs from the matched staging-versus-interleaving experiment in Section 2.3, which fixes the counts of both objectives and changes only their order.

Figure 3 first tests the DMFT across this abrupt change of learning rule. The single-channel process follows the MSE force before the switch and the policy-dependent GRPO force afterward, while the same macroscopic closure propagates across both stages. The DMFT tracks the finite-window key alignment, clean prediction risk, and position accuracy through the switch. This experiment extends the validated setting beyond the fixed-update regimes treated by existing multi-pass SGD theory such as Fan & Wang (2026).

The same comparison shows the benefit of verifier-guided refinement in a suitable low-noise regime. MSE first learns both key and value representations; GRPO then uses exact position feedback to refine the key while preserving the learned value vector. The resulting localization gain lowers clean prediction error relative to continued MSE, with the same qualitative ordering in step-by-step finite-dimensional training.

4.3 When does stage separation help?

Large-model training commonly separates supervised representation learning from a later reinforcement-learning stage. We study a minimal analogue in which MSE first trains the key and value, followed by pure GRPO on the key. To hold information access fixed while changing temporal organization, the baseline is exact-count random interleaving: staging and interleaving receive the same supervised labels, verifier subset, numbers of MSE and GRPO updates, and verifier-call budget; only the update order differs.

Figure 4 shows two systematic crossovers. At low label noise, interleaving benefits from continued co-adaptation of localization and readout. Increasing persistent label noise reverses the ordering because the final pure-GRPO block freezes the value vector while continuing to refine the key. At fixed noise, shorter sequences favor interleaving and longer sequences favor staging. The crossover locations are obtained from paired finite-dimensional experiments.

Appendix G.4 gives an exact endpoint decomposition of these trends. With endpoint clean risks $\mathcal{E}_S$ and $\mathcal{E}_X$ for staging and interleaving,

$$\Delta(L,\sigma) := \mathcal{E}_S(L,\sigma) - \mathcal{E}_X(L,\sigma) = b(L) - N(L,\sigma), \tag{13}$$

where $b(L)$ is the clean-label benefit of interleaving and $N(L,\sigma)$ is its differential noise penalty. Staging is favored when the latter exceeds the clean co-adaptation benefit. The local analysis also shows that noisy MSE contributes value-force variance proportional to $\eta_{\mathrm{M}}^2\sigma_y^2\|p\|_2^2$, while at a uniform-policy reference $\|p\|_2^2 = 1/L$. GRPO is length-sensitive through the reward-active group probability $1 - L^{-G_{\mathrm{grp}}} - (1-1/L)^{G_{\mathrm{grp}}} = O(G_{\mathrm{grp}}/L)$. These local scalings explain why noise and sequence length alter the balance, while the endpoint crossover integrates the full training trajectory.

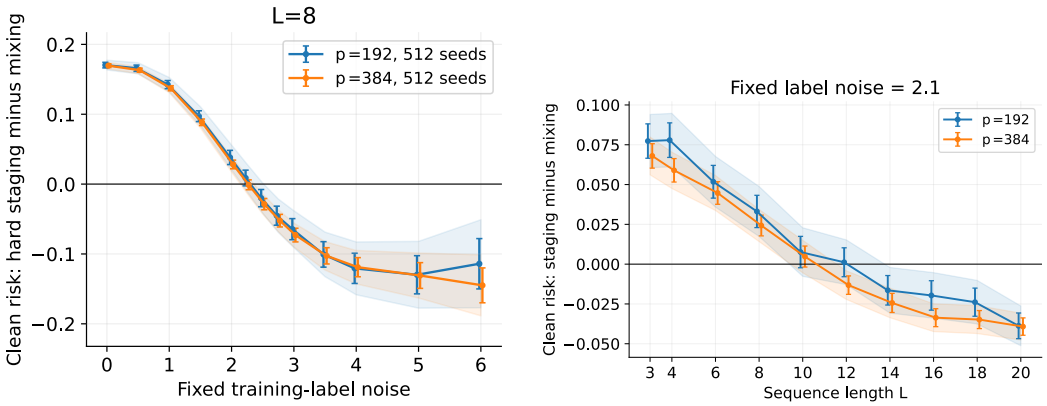


Figure 4: Matched staging versus exact-count random interleaving. Left: the schedule gap changes sign as persistent label noise increases at $L = 8$. Right: at fixed noise, the same gap changes sign with sequence length. Positive staging-minus-interleaving gaps favor interleaving.

5 Related work

Our work connects finite-data learning dynamics, high-dimensional theories of attention, and verifier-guided training. We separate the literature on fresh-sample online learning from finite-corpus dynamics, since repeated reuse of the same examples is the source of the sample-specific memory studied here.

Online and finite-corpus learning dynamics. Statistical-mechanics analyses of online learning often consider a one-pass setting in which each SGD step receives a fresh independent example. In this regime, high-dimensional teacher–student dynamics can close on a finite set of overlaps and correlations, leading to deterministic ODE descriptions of the learning trajectory (Goldt et al., 2020). Recent work has extended such analyses to sequence models and attention-like architectures; in particular, Arnaboldi et al. (2025) derive high-dimensional one-pass SGD dynamics for sequence single-index models and simplified single-layer attention networks. Fresh-sample dynamics omit revisits and the resulting sample-specific retarded self-interaction that is central here.

Finite training sets require a different dynamical description. Restricted-training-set analyses showed that repeated presentation of $O(p)$ examples produces temporally correlated effective processes with nontrivial response and memory kernels (Sollich & Barber, 1997; Coolen & Saad, 2000; Heimel & Coolen, 2001; Luo & Wong, 2003). Dynamical cavity methods provide a complementary route to such self-consistent effective processes (Agoritsas et al., 2018), while Gerbelot et al. (2024) establish rigorous high-dimensional DMFT equations for a broad class of first-order methods on Gaussian empirical-risk problems. Most recently, Fan & Wang (2026) derive Poisson-driven DMFT equations for multi-pass SGD in high-dimensional multi-index models and recover the one-pass limit when the effective sample supply becomes infinite. Our setting differs by coupling two attention fields, repeated sample reuse, and policy-dependent verifier sampling within the same finite corpus.

High-dimensional theories of attention. A substantial recent literature analyzes tractable attention models in high dimension. For linear attention, Lu et al. (2025) derive sharp asymptotics for in-context linear regression, including sample-complexity, task-diversity, and memorization-to-generalization transitions. Related high-dimensional sequence models yield sharp statistical and algorithmic thresholds for attention architectures (Troiani et al., 2025). These works primarily characterize asymptotic generalization or inference; our focus is the temporal dynamics created by finite-corpus reuse.

Several recent papers instead study how attention models are learned. Zhang et al. (2025) analyze gradient-descent dynamics of multi-head linear self-attention and obtain explicit reduced dynamics exhibiting abrupt or saddle-to-saddle acquisition of in-context learning, depending on the key–query parametrization. Arnaboldi et al. (2025) study high-dimensional SGD dynamics for a nonlinear

sequence single-index formulation containing simplified softmax attention, but in a one-pass fresh-sample regime. For nonlinear attention more broadly, mean-field approaches have also been used to study optimization landscapes and network-state dynamics (Kim & Suzuki, 2024; Poc-López & Aguilera, 2024); these analyses concern parameter-distribution mean-field optimization and internal self-attention dynamics, respectively, whereas our effective process tracks memory generated by finite-training-set reuse. Closest to our model, Duranthon et al. (2026) analyze single-location regression and show a high-dimensional statistical advantage of softmax over linear attention. We adopt this interpretable task but study repeated finite data, coupled key–value training dynamics, and objective switching.

Verifier objectives and training order. GRPO and related verifier-guided methods motivate the sampled-group update studied here (Shao et al., 2024; Yu et al., 2025). Phasic training and sequential-objective analyses show that separating objectives can change the representation learned by a shared model (Cobbe et al., 2021; Dherin, 2023). Our comparison isolates this effect under finite-data reuse by matching the numbers of MSE and GRPO updates, verifier calls, data, and initialization, and changing only their temporal ordering. The resulting theory therefore connects objective scheduling to the sample-memory and response structure of a finite-corpus attention model.

6 Conclusion

We developed a finite-corpus dynamical theory for attention trained by supervised regression and verifier-guided reinforcement learning, using single-location regression as a controlled analogue of multi-stage model training. In the high-dimensional limit, repeated sample reuse reduces to a self-consistent single-channel process whose macroscopic evolution is governed by overlaps, two-time covariances, response kernels, and the full response of policy-dependent action histories. The resulting DMFT accurately describes an abrupt switch from MSE to GRPO and shows how exact position feedback can refine localization while preserving a previously learned value representation. A matched comparison with exact-count random interleaving shows that the relative advantage of stage separation changes with label noise and sequence length. The endpoint decomposition and local force analysis identify the underlying tradeoff between continued clean co-adaptation and continued exposure to noisy supervised updates. These results provide a tractable account of how finite-data reuse, objective switching, and supervision quality jointly shape multi-stage learning, and establish a basis for extending finite-corpus DMFT to richer attention models and reinforcement-learning objectives.

Reproducibility statement

Appendix H show the detail experiments settings.

Ethics statement

The study uses synthetic Gaussian sequences and an exact-position verifier, with no human-subject or personal data. It examines learning dynamics and objective order in a controlled model. The conclusions concern the stated synthetic model and training protocols.

AI use statement

Generative AI tools assisted with theoretical analysis, code development, and checked by the author.

References

Elisabeth Agoritsas, Giulio Biroli, Pierfrancesco Urbani, and Francesco Zamponi. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002, 2018.

Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborová. Asymptotics of SGD in sequence-single index models and single-layer attention networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Karl W. Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 2020–2027. PMLR, 2021.

A. C. C. Coolen and David Saad. Dynamics of learning with restricted training sets. *Physical Review E*, 62:5444–5487, 2000.

Benoit Dherin. Implicit biases in multitask and continual learning from a backward error analysis perspective, 2023.

Olivier Duranthon, Pierre Marion, Claire Boyer, Bruno Loureiro, and Lenka Zdeborová. Statistical advantage of softmax attention: Insights from single-location regression. In *International Conference on Learning Representations*, 2026.

Zhou Fan and Leda Wang. High-dimensional learning dynamics of multi-pass stochastic gradient descent in multi-index models, 2026.

Cédric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborová. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024.

Sebastian Goldt, Madhu S. Advani, Andrew M. Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124010, 2020.

J. A. F. Heimel and A. C. C. Coolen. Supervised learning with restricted training sets: a generating functional analysis. *Journal of Physics A: Mathematical and General*, 34(42):9009–9026, 2001.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 24527–24561. PMLR, 2024.

Yue M. Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025.

Peixun Luo and K. Y. Michael Wong. Dynamical and stationary properties of on-line learning from finite training sets. *Physical Review E*, 67:011906, 2003.

Youssef Mroueh, Nicolas Dupuis, Brian Belgodere, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, Jiri Navratil, Jerret Ross, and Jesus Rios. Revisiting group relative policy optimization: Insights into on-policy and off-policy training. In *International Conference on Learning Representations*, 2026.

Ángel Poc-López and Miguel Aguilera. Dynamical mean-field theory of self-attention neural networks, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024.

Peter Sollich and David Barber. On-line learning from finite training sets. *Europhysics Letters*, 38(6): 477–482, 1997.

Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of learning in sequence multi-index models and deep attention networks: high-dimensional asymptotics and sharp thresholds. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 60147–60182. PMLR, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, 2025.

Yedi Zhang, Aaditya K. Singh, Peter E. Latham, and Andrew M. Saxe. Training dynamics of in-context learning in linear attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 76047–76087. PMLR, 2025.