LEARNING DYNAMICS OF ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) is emerging as an important paradigm for language-model post-training, combining dense teacher supervision with training on student-generated trajectories. This creates a feedback loop where each update changes the student and thereby the training distribution that generates subsequent updates. Yet, how this feedback shapes learning and generalization remains poorly understood. We conduct a mechanistic analysis of OPD's learning dynamics, distinguishing the two feedback loop stages: conversion of the distillation signal into parameter updates, and broadcast through which these updates change predictions across prefixes and reshape the rollout distribution. This framework helps explain both limited student responsiveness despite substantial teacher-student disagreement, and degeneration when updates that improve local teacher agreement also increase the probability of unsuccessful continuations, exposing the student to unreliable supervision in subsequent rollouts. Finally, we show that offline distillation can prepare a student for OPD without significantly improving immediate held-out accuracy. While OPD alone may degrade generation, it can yield substantial accuracy gains after sufficient offline preparation. Together, our results connect teacher supervision, student geometry, and the evolving rollout distribution in determining the learning dynamics of OPD and its effects on downstream performance.

1 INTRODUCTION

On-policy distillation (OPD) is becoming an important component of language-model post-training (Lu & Thinking Machines Lab, 2025; Song & Zheng, 2026a). By matching a teacher's next-token distributions at student-generated prefixes, it combines dense supervision with training on the states the student actually visits (Gu et al., 2024; Agarwal et al., 2024). Yet, how this local supervision becomes useful learning over an evolving training distribution remains insufficiently understood. It is therefore important to theoretically understand both the conditions under which OPD improves generalization and its role alongside offline distillation.

The difficulty is that each OPD update changes both the student's current behavior and the prefixes on which it will subsequently learn. Reducing teacher-student divergence on currently supervised predictions does not necessarily produce parameters from which further updates are useful (Li et al., 2026b; Armandpour et al., 2026). Modelling this requires a view of how supervision changes behavior beyond the current batch and how that impacts learning at later stages. We conduct a mechanistic analysis, aided by a first-principles account, of OPD learning dynamics to study this process and explain existing empirical challenges and justify their fixes.

We first characterize how the teacher-student log-probability difference can lead to meaningful weight updates, which depends on the overlap between the distillation signal and the logit-space directions the student's current parameters can express. We then describe how learning at one prefix affects predictions at future ones through their dependence on shared parameter directions. We formalize this broadcast relationship through the student's tangent kernel between the two prefixes, building on Ren & Sutherland (2025); Malladi et al. (2023). Finally, we connect each update to its effect on the outcomes of student-generated sequences. The effect of these updates in broader settings also depends on what the training data covers: matching the teacher on narrow tasks may transfer only the capabilities that those tasks exercise. Integrating these effects over training characterizes what OPD can reach from a given starting point, accounting for how much successive updates improve or degrade performance.

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.

This framework helps explain two primary failure modes of OPD: stagnation under stronger teachers and degeneration into incoherent outputs (Li et al., 2026b; Fu et al., 2026; Wang et al., 2026). In the first case, despite substantial teacher–student log-probability differences, we show that the initial distillation signal can induce smaller parameter gradients as training proceeds, contributing to stalled improvement. In the second case, updates that reduce teacher–student disagreement at supervised prefixes also change predictions at other prefixes through shared parameters. These changes can shift the rollout distribution toward continuations the student cannot execute reliably, making those unsuccessful continuations the source of subsequent supervision. A well-prepared student, in contrast, can sustain useful learning as its rollout distribution changes.

The dependence of OPD on the student's own continuation ability and shared-parameter (kernel) geometry suggests that offline preparation need not immediately improve task accuracy. Its deferred value can come from enabling the subsequent OPD process to sustain productive learning as the rollout distribution changes. We test this by training the student on teacher-generated trajectories before applying the same OPD procedure. Across model classes, we find that under sufficient offline training, the subsequent OPD phase produces substantial accuracy improvements, while offline training itself produces limited gains, and OPD with insufficient preparation degrades generation and accuracy. Together, these results motivate three questions that guide the rest of the paper. What determines the *conversion* of teacher supervision into parameter updates? Does the *broadcast* of these updates across prefixes improve subsequent reasoning or steer the student toward unreliable continuations and supervision? Can offline *preparation* strengthen the student's continuation ability and enable productive on-policy learning? Sections 3.2, 3.3 and 3.4 study these three questions respectively.

Related works. We briefly discuss related works here, and leave the expanded version to Appendix Section B. *On-policy distillation.* Training on learner-generated states with expert feedback is a classical approach to addressing distribution shift in imitation learning (Ross et al., 2011), applied to autoregressive distillation by Lin et al. (2020). Gu et al. (2024) develop reverse-KL distillation using student rollouts, while Agarwal et al. (2024) unify on-policy and off-policy training across multiple divergences. Lu & Thinking Machines Lab (2025) demonstrate its effectiveness for reasoning and personalization in language-model post-training. More recently, Yang et al. (2026) interpret OPD as dense KL-constrained reinforcement learning and introduce reward extrapolation. Empirical studies identify teacher-student compatibility and unreliable supervision on student prefixes as important factors in transfer and training failure (Li et al., 2026b; Fu et al., 2026).

Learning dynamics of finetuning. Kernel-based analyses connect parameter updates to changes in language-model predictions (Malladi et al., 2023). Ren & Sutherland (2025) decompose per-step prediction changes using the loss gradient and empirical tangent kernel, analyzing confidence dynamics in SFT and DPO (Rafailov et al., 2024). Tomihari (2026) study how RL updates propagate across examples, while Shen et al. (2026) identify persistent low-dimensional subspaces in OPD updates. Building on these perspectives, we connect the student's evolving tangent kernel to its response to teacher supervision, the quality of its continuations, and the resulting distribution of future training prefixes. This distinguishes limited response to supervision from policy changes that worsen generation, and explains how either can shape subsequent learning.

2 SETUP

On-policy distillation (OPD) trains a student on its own samples. The student generates a continuation from a prompt. At every prefix of that continuation, the training loss pulls the student's next-token distribution toward the next-token distribution of a fixed teacher (Gu et al., 2024; Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025). This section fixes the notation for the objects that recur in the paper, including the policies, the source distribution that a run trains on, the request field that the teacher induces on that source, and the evaluation targets that we score runs against.

Policies and rollouts. A prefix s is a prompt followed by a partial continuation. Prefixes range over a space $\mathcal{S}$, and tokens range over a finite vocabulary $\mathcal{V}$. The student has parameters $\theta \in \mathbb{R}^d$, logits $z_\theta(s) \in \mathbb{R}^{|\mathcal{V}|}$, and next-token distribution $p_\theta(a \mid s) = \mathrm{softmax}(z_\theta(s))_a$. The teacher $q(\cdot \mid s)$ is fixed throughout training. Training proceeds in steps $k = 0, 1, 2, \ldots$, and p_k and z_k denote the student's quantities at step k. Where it simplifies a formula, we use the continuous-time reading $\dot{\theta}_t = u_t$. For

plain gradient steps this gives a first-order local approximation; exact finite-step responses are in Section C.2. Two distributions share the student's weights. The *model policy* p_k is the distribution that the loss acts on. The *rollout policy* $\pi_k = \mathcal{G}[p_k]$ is the distribution that generates the samples. The transform $\mathcal{G}$ collects the temperature, end-of-sequence conventions, etc. We sample prompts $b \sim \beta$ and continue them under π_k. The step-k *rollout population* is $\Pi_k(\xi) = \beta(b) \prod_j \pi_k(a_j \mid s_j)$ for a trajectory $\xi = (b, a_0, a_1, \ldots)$, with $s_0 = b$ and $s_{j+1} = s_j \oplus a_j$ up to termination or the cap. The *training source* $\mu_k = \mathcal{M}(\Pi_k)$ is the distribution over prefixes that the implementation trains on. The map $\mathcal{M}$ collects the rollout counts, the loss masks, the truncation, and the per-token normalization.

Loss, request field, and update. For full-vocabulary reverse Kullback–Leibler divergence, the per-prefix cost is $c_k(s) = D_{\mathrm{KL}}(p_k(\cdot \mid s) \,\|\, q(\cdot \mid s))$. Write $\ell_k(s,a) = \log\big(p_k(a \mid s)/q(a \mid s)\big)$ for the log-ratio of the two policies at a token, and $\bar{\ell}_k(s) = \mathbb{E}_{a\sim p_k}[\ell_k(s,a)]$ for its mean under the student. The negative gradient of the cost with respect to the logits is $B_k(s,a) = p_k(a \mid s)\,\big[\bar{\ell}_k(s) - \ell_k(s,a)\big]$. We call B_k the *request field*[1]. It captures the student-teacher disagreement on a given prefix. The product $\Gamma_k = \mu_k B_k$, visitation times request, is the *request population*. The student's *tangent map* $\Phi_k(s) = \nabla_\theta z_{\theta_k}(s) \in \mathbb{R}^{|\mathcal{V}|\times d}$ pulls a request back into parameter space. This gives the per-prefix gradient $g_k(s) = \Phi_k(s)^\top B_k(s) = -\nabla_\theta c_k(s)$ and the aggregate update direction $u_k = \mathbb{E}_{s\sim\mu_k}[g_k(s)]$. Truncation or detached coefficients can change the numerical request: the response laws use the complete implemented logit force, while the separate conversion-audit field is specified in Section D.3. The full-KL derivation is in Section C.1.

3 The Learning Dynamics of the OPD loop

3.1 The OPD loop and the response law

Recall that Π_k denotes the distribution of student rollouts and $\mu_k = \mathcal{M}(\Pi_k)$ the distribution of prefixes used by the loss, including masks and token weights. At a sampled prefix s, the loss supplies the request $B_k(s) = -\nabla_{z_k(s)} c_k(s)$.

The weighted request field $\Gamma_k = \mu_k B_k$ specifies both *where* supervision is applied and *what change* it requests. One training step follows

$$\underbrace{\Pi_k}_{\text{rollouts}} \xrightarrow{\mathcal{M}} \underbrace{\mu_k}_{\text{source}} \xrightarrow{q} \underbrace{\Gamma_k = \mu_k B_k}_{\text{weighted requests}} \xrightarrow{\Phi} \underbrace{\Delta\theta_k}_{\text{parameter step}} \xrightarrow{\text{response}} \underbrace{\delta_k^p, \delta_k^\pi}_{\text{policy changes}} \longrightarrow \underbrace{\Pi_{k+1}}_{\text{next rollouts}} \tag{1}$$

The response $\delta_k^p(x,a) = \log p_{k+1}(a \mid x) - \log p_k(a \mid x)$ measures the change at a probing prefix x; δ_k^π is its rollout-policy counterpart. Sampling with the changed policy regenerates the prefix (and rollout) distribution, $\mu_{k+1} = \mathcal{M}(\Pi_{k+1})$.

A local request can change behavior at other prefixes. Training differentiates at the sampled prefixes, not through their sampling probabilities. The resulting raw direction is $u_k = \mathbb{E}_{\mu_k}[\Phi_k(s)^\top B_k(s)]$, where $\Phi_k(s) = \nabla_\theta z_k(s)$ is the student's tangent map. For a plain gradient step $\Delta\theta_k = \eta u_k$, the response at any receiver is

$$\delta_k^p(x,a) = \eta \sum_s \underbrace{\mu_k(s)}_{\text{source weight}} \Big[\underbrace{\mathsf{A}_k^p(x)}_{\text{normalization}} \underbrace{K_k(x,s)}_{\text{student coupling}} \underbrace{B_k(s)}_{\text{request}} \Big]_a + O(\eta^2), \tag{2}$$

where $K_k(x,s) = \Phi_k(x)\Phi_k(s)^\top$ and $\mathsf{A}_k^p(x) = I - \mathbf{1}p_k(\cdot \mid x)^\top$. The kernel measures how source and receiver logits move together through shared parameters; A_k^p converts logit motion into log-probability motion. Thus a receiver can change without appearing in the batch (proof and finite-step scope: Section C.2).

Organization of the rest of the section. Conversion (Section 3.2) concerns the student's response to the supervised requests, examined under successful and stalled student-teacher pairings. Broadcast

[1]This is the "student-teacher disagreement" abstraction our theory is primarily concerned with: we adopt the specific reverse KL due to its popularity in OPD, but the theory covers general disagreement-measuring losses.

(Section 3.3) concerns the effects of the update on predictions and subsequent updates: broadcast decomposes into drift, which measures fixed-history prediction changes, and routing, which measures changes in encountered rollout histories. Offline-distillation-based preparation (Section 3.4) can change the starting point of this OPD loop and significantly improve its performance without significant improvements in immediate heldout gains.

3.2 Conversion: Teacher requests and the student's responsiveness

A stronger teacher can transfer less. We open with an intriguing empirical puzzle identified in the OPD community (Li et al., 2026b): a larger teacher can transfer less than an RL-trained sibling, and Li et al. (2026b) suggests cancellation among token-level signals as a possible explanation. We revisit this comparison at 1.5B and 7B (Figure 1), holding the starting DeepSeek-R1-Distill-Qwen student and OPD recipe fixed within each scale. JustRL-1.5B and AceReason-Nemotron-7B yield substantial gains, whereas the larger DeepSeek-7B and DeepSeek-14B teachers transfer little despite higher benchmark scores (Section D.1). We call these runs *healthy* and *stalled*. We find that the teachers' request fields differ substantially in semantic functions, and that the student's response separates in two ways: collective conversion and kernel-based adaptation toward the initial requests.

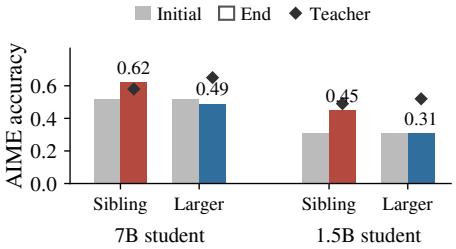


Figure 1: **Teacher strength does not order transfer.** Initial/end AIME accuracy; diamonds mark teachers. Red: sibling-taught; blue: larger-teacher taught.

The teachers disagree at different reasoning positions. On fixed initial-student rollouts, Figure 2 compares where each teacher's strongest 1% of requests occur. We find that the RL-sibling teachers emphasize reasoning-management and initiating positions, with the particular categories varying by scale; larger teachers place relatively more of their strongest requests at digit-bearing positions. This separation connects to findings in the reasoning-interpretability community, that RL tends to improve how well the model "orchestrates" the base reasoning mechanisms (Venhoff et al., 2025a). Full category definitions, numerical comparisons, and the annotation protocol are in Section D.2.

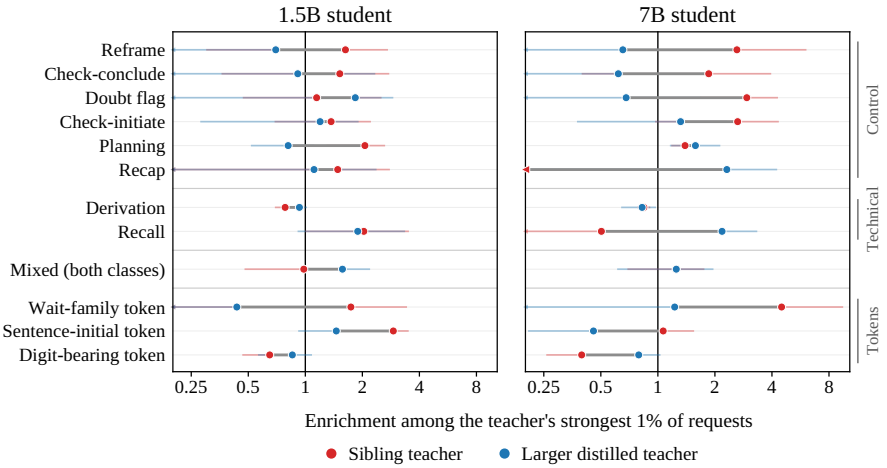


Figure 2: **Teacher requests differ in their semantic and positional distribution.** Each dumbbell pairs the teachers' enrichment among their strongest 1% of initial requests; 1 denotes corpus prevalence. Thin lines show bootstrap intervals; an arrow-shaped point denotes zero enrichment. Category labels describe request locations. Full counts, intervals, and category-wise comparisons are in Section D.2.

From requests to collective motion. Figure 2 shows that the teachers provide systematically different request fields, but this alone does not help us understand which teacher transfers better. The next question is whether the student can convert a given request field into coherent parameter motion, especially the strongest requests (which differ substantially across the healthy and stalled teachers). Let us fix a prefix population ν, and write the gradient $g_t(s) = \Phi_t(s)^\top B_t(s)$, $u_t = \mathbb{E}_\nu g_t(s)$, and $P_t = \mathbb{E}_\nu \|B_t(s)\|^2$. Recall B is the request field ("update force") induced by the teacher-student disagreement. Under gradient flow $\dot{\theta}_t = u_t$, the *coefficient of the logit response along the request field* is

$$\Xi_t[B_t] = \frac{\langle B_t, \dot{z}_t \rangle_\nu}{P_t} = \frac{\|u_t\|^2}{P_t}, \qquad \dot{z}_t = \Xi_t[B_t]B_t + r_t, \quad \langle B_t, r_t \rangle_\nu = 0, \tag{3}$$

where $\langle F, G \rangle_\nu = \mathbb{E}_\nu[F^\top G]$ and $P_t > 0$. With nonzero denominators, this *collective conversion gain* factors as

$$\Xi[B] = \underbrace{\frac{\mathbb{E}_\nu \|\Phi(s)^\top B(s)\|^2}{P}}_{\text{accessibility } A} \underbrace{\frac{\|\mathbb{E}_\nu g(s)\|^2}{\mathbb{E}_\nu \|g(s)\|^2}}_{\text{coherence } C}, \qquad \|u\|^2 = P\Xi = PAC. \tag{4}$$

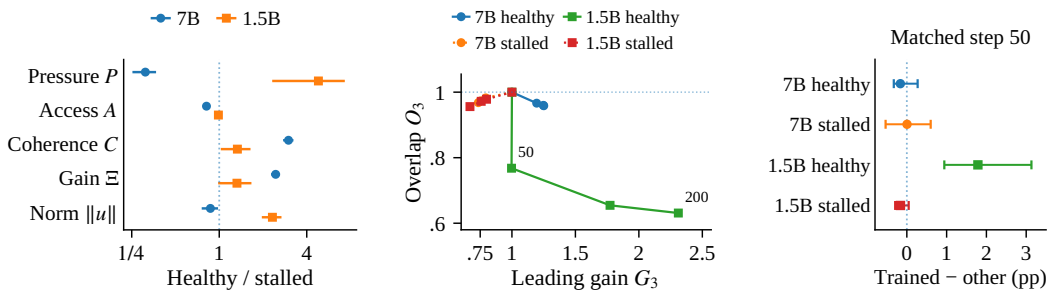
Accessibility averages the self-kernel Rayleigh coefficients $B(s)^\top K(s,s)B(s)/\|B(s)\|^2$, weighted by request energy: it measures how strongly each update force $B(s)$ is *visible* to the student's tangent map $\Phi(s)$. Coherence measures how well the resulting parameter updates $g(s) = \Phi(s)^\top B(s)$ combine: aligned updates reinforce one another, while cancellation lowers C. Thus large requests need not convert efficiently, and similar gradient norms can conceal different gains per unit request. The deeper geometric interpretations and derivations are in Section C.3.

Initial conversion separates the pairings. At 7B, healthy requests have slightly lower accessibility but 2.75–3.24$\times$ higher coherence, yielding 2.27–2.62$\times$ higher Ξ across the two sampled high-strength strata (Figure 3a). Greater stalled pressure largely compensates in gradient norm, masking the weaker conversion of the update forces. At 1.5B, however, *initial* accessibility is similar and coherence differs less; greater healthy pressure yields roughly 2–2.7$\times$ larger raw gradients. This weaker initial contrast in how well the 1.5B model converts the updates motivates examining how responsiveness evolves as the student learns.

Successful runs acquire sensitivity to the same requests. Live accessibility changes with both the request B_t and the kernel K_t. To isolate kernel motion, we freeze each teacher's initial field B_0, prefix bank, and token coordinates, then reapply the same numerical force at every checkpoint. The common initialization anchors both fields to the same student, while preserving their distinct request profiles. With $b_0(s) = B_0(s)/\|B_0(s)\|$, $a_t(s) = b_0(s)^\top K_t(s,s) b_0(s) = \sum_j \lambda_{j,t}(s) |v_{j,t}(s)^\top b_0(s)|^2$. Here $(\lambda_{j,t}, v_{j,t})$ are eigenpairs of the self-kernel, and the request-energy-weighted average of $a_t(s)$ is $A_t[B_0]$. Larger eigenvalues amplify accessible directions; turning strong eigendirections toward B_0 makes more of the same request visible. Both can increase local responsiveness (Section C.4).

At 1.5B, healthy leading-subspace overlap falls to .77 by step 50 while leading strength remains near its initial value (Figure 3b). The eigenframe re-orients towards the request field: the fixed request gains a median 5.5 percentage points of top-three projection, and its paired gain exceeds the other teacher's by 1.79 points (Figure 3c). By step 200, leading strength reaches 2.31$\times$ and median request-direction sensitivity 2.08$\times$ baseline. In contrast, the stalled run contracts to .67$\times$ leading strength and .83$\times$ sensitivity. At 7B, healthy adaptation is primarily amplification of the existing eigenframe: leading strength reaches 1.25$\times$ with overlap .96 against the initial frame, while stalled leading strength falls to .74$\times$.

From responsiveness to the on-policy loop. These results distinguish healthy and stalled pairings in the requests they receive and the responsiveness their tangent geometry develop. Through $\Xi = AC$, accessibility and coherence determine motion along the requested field, exposing distinctions easily hidden by teacher scores or gradient norms alone. We will show that in the next sub-section, JustRL-1.5B, successful with the stronger DeepSeek student, degrades generation in the weaker Qwen2.5-1.5B-Instruct student. We use this as a model organism to trace how policy changes reshape its generative distribution and subsequent supervision.



(a) Initial conversion. (b) Amplification and reorientation. (c) Teacher-relative reorientation.

Figure 3: **Collective conversion and acquired responsiveness distinguish the pairings.** (a) Healthy/stalled ratios; segments show the range across the CONTROL and TECHNICAL strata. (b) Median leading-three eigenvalue gain G_3 and initial-subspace overlap O_3. (c) Step-50 paired trained-minus-other-field projection gain. Definitions and full results are in Sections D.4 and D.5.

3.3 BROADCAST: WHY LOCAL AGREEMENT NEED NOT IMPROVE REASONING

Under JustRL-1.5B, both DeepSeek-R1-Distill-Qwen-1.5B and Qwen2.5-1.5B-Instruct substantially reduce their local training loss. Yet DeepSeek gains accuracy (Section 3.2), while the base Qwen 1.5B model declines to single-digit accuracy on MATH within 100 steps. Why does fitting the same teacher help one learner and harm the other? We trace both the predictions changed by an update and the rollouts (prefixes) on which the student subsequently learns.

Routing and drift: changed predictions vs changed prefix visitation. An OPD update has two coupled effects: it changes *which prefixes* are encountered in future rollouts, and it changes the model's *predictions at a fixed prefix*. We call these effects *routing* and *drift*, respectively.

For a *prefix readout* $Y_t(x)$, such as local loss or entropy, let $J_Y(t) = \mathbb{E}_{\nu_t} Y_t(x)$, where ν_t is the observed prefix population. Then

$$\dot{J}_Y(t) = \underbrace{\sum_x \dot{\nu}_t(x) Y_t(x)}_{R_Y(t)\text{: routing}} + \underbrace{\mathbb{E}_{\nu_t} \dot{Y}_t(x)}_{D_Y(t)\text{: drift}}, \qquad J_Y(b) - J_Y(a) = \int_a^b (R_Y + D_Y)\, dt. \tag{5}$$

Both channels are driven by the K-Γ broadcast. The distinction is *where the same update acts*. Under gradient flow, requests induce a logit response $v_t(x)$ at every receiver prefix x. With a fixed-support, temperature-τ sampler, this gives token log-probability velocity

$$\zeta_t(x, a) = \partial_t \log \pi_t(a \mid x) = \frac{(e_a - \pi_t(\cdot \mid x))^\top v_t(x)}{\tau}, \qquad v_t(x) = \sum_s K_t(x, s) \Gamma_t(s). \tag{6}$$

At the observed prefix itself, $v_t(x)$ changes its prediction, giving drift $D_Y = \mathbb{E}_{\nu_t}[\nabla_z Y_t(x)^\top v_t(x)]$. To reach that prefix, however, a rollout must first generate its ancestor tokens. The earlier token responses add to its log reach-probability velocity, $\lambda_t(x) = \sum_{j<h(x)} \zeta_t(x_j, a_j)$. For fixed prompts and normalized expected prefix counts, this gives routing $R_Y = \mathrm{Cov}_{\nu_t}(Y_t, \lambda_t)$. Routing is positive when prefixes with larger readout values become more common. Thus an update can make a rollout (prefix visitation) more likely without improving the prediction there. The accumulated exposure change also renews the training prefix distribution μ:

$$\mu_b(s) \propto \mu_a(s) e^{\Lambda_{a:b}(s)}, \qquad \Lambda_{a:b}(s) = \int_a^b \lambda_t(s)\, dt \;\; \rightarrow \;\; \Gamma_b = \mu_b B_b. \tag{7}$$

The new rollouts then supply the next requests, closing the OPD feedback loop (see Sections C.5 and C.6). We estimate these effects by *cross-scoring training rollouts and learners*. Write $M_{ij} = \mathbb{E}_{x \sim \nu_i} Y_j(x)$ for the readout of learner j on the frozen bank of rollouts from endpoint i, with $i, j \in$

$\{a, b\}$. Replacing the bank at a fixed learner checkpoint measures routing; replacing the checkpoint on a fixed bank measures drift. Averaging the two replacement orders gives

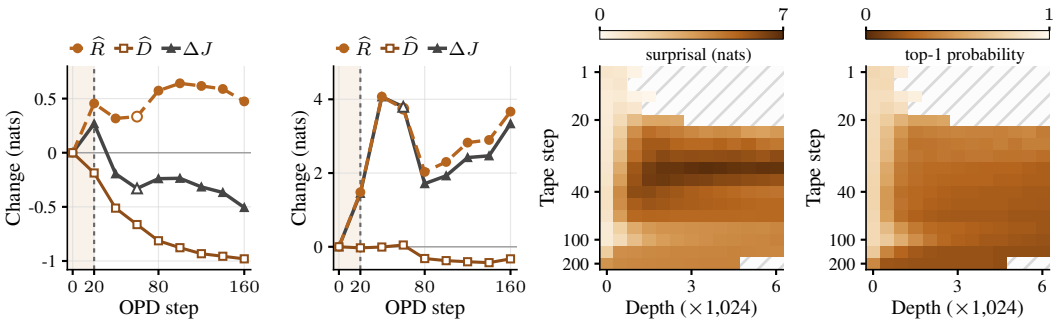
$$\widehat{R}_{a:b} = \tfrac{1}{2}[(M_{ba} - M_{aa}) + (M_{bb} - M_{ab})], \qquad \widehat{D}_{a:b} = \tfrac{1}{2}[(M_{ab} - M_{aa}) + (M_{bb} - M_{ba})]. \quad (8)$$

These contributions sum exactly to $M_{bb} - M_{aa}$. Our curves accumulate them over consecutive windows; its relation to the path integrals and further details are in Sections C.7 and E.1[2].

From prefix diagnostics to completed outcomes. Loss and entropy are read directly from predictions at individual prefixes. Correctness, by contrast, is known only after completing the rollout. To connect local policy motion to a trajectory-determined terminal score Y, set $J_Y = \mathbb{E}_{\Pi_t} Y$ and use the student's continuation values $Q_t^Y(x, a) = \mathbb{E}_t[Y \mid x, a]$ and $V_t^Y(x) = \mathbb{E}_{a\sim\pi_t} Q_t^Y(x, a)$. Applying the same exposure/conditional-change split at an intermediate prefix, then unrolling the continuation:

$$\begin{aligned} J_Y(T) - J_Y(t_\star) &= \int_{t_\star}^{T} \sum_x d_t(x) \mathbb{E}_{a\sim\pi_t} \left[\zeta_t(x, a) \underbrace{(Q_t^Y(x, a) - V_t^Y(x))}_{\text{continuation advantage}} \right] dt \\ &= \frac{1}{\tau} \int_{t_\star}^{T} \sum_{x,s} d_t(x)[\pi_t \odot (Q_t^Y - V_t^Y)](x)^\top K_t(x, s)\Gamma_t(s)\, dt, \end{aligned} \quad (9)$$

where $d_t(x)$ is the unnormalized probability of reaching x. This weights motion by what the *student* can achieve. When the favored alternatives have similarly unsuccessful continuations, fitting their next-token distributions can yield little accuracy progress overall (see Section C.8 for the proof).



(a) Loss change (b) Entropy change (c) Student uncertainty (d) Teacher confidence

Figure 4: **Routing into unreliable prefixes, then fitting them.** Unprepared Qwen: (a,b) cumulative routing, drift, and total changes in local loss and entropy, each on its own scale. The vertical guide at step 20 marks the approximate onset-to-fitting transition; the effects overlap. (c,d) Student surprisal and teacher top-1 probability; darker means greater uncertainty. Columns are 512-token windows. Details are in Section E.1.

Unprepared Qwen: routing first changes where it reasons. In the early phase of training, the teacher pushes the learner toward longer rollouts (such as heavy negative requests on the student's "overly early" answer commitments), but these extensions become high-entropy and incoherent. The teacher supervision there also becomes much less concentrated than in the shallower rollout windows (Figure 4c,d). As seen in Figure 4(a) and (b), from step 1 to 20, loss rises $0.58 \to 0.85$ (routing $+0.455$ outweighs drift -0.187); entropy also rises mainly through routing. Intuitively, the student, which has never been trained to emit long think rollouts, gets pushed into reasoning for long before committing an answer, and executing reasoning states which it is highly unfamiliar with. The teacher, conditioned on these low-quality states, also give much less confident guidance. The drift of the loss is small in this phase, however: the student has not learned to (uniformly) learn the flatter prediction patterns of the teacher across the visited prefixes yet.

[2]While a step-wise estimate for routing and drift is theoretically recommended, it is too expensive for our setting, and we resort to intervals of 10 update steps.

Then drift fits the unreliable population. By update 160, loss has fallen to 0.073(Figure 4a). Entropy remains high within the band around $2 - 4$ nats after the onset phase (beyond step 20). Loss reduction, however, becomes increasingly driven by drift, that is, the student is indeed fitting the teacher's low confidence predictions on the chaotic, low-quality prefixes continually generated during training. To further test this, we find that, on the same batch-41 training prefixes, loss falls $0.79 \rightarrow 0.06$ while entropy falls $4.99 \rightarrow 3.99$: local fitting actually sharpens predictions on this rollout bank toward an uncertain but slightly less entropic teacher (Sections C.9 and E.4.1).

The changed rollout distribution now supply subsequent requests. The OPD update loop therefore spends further updates fitting unreliable requests on the poor continuations it has made common, rather than restoring a reliable path to an answer. Equation (9) explains why this matters for terminal accuracy: policy motion contributes through the value of the *student's* ensuing completions, not local teacher agreement alone. To probe the continuation-value term in Equation (9), we freeze a set of answer-decision prefixes from the initial student's held-out MATH rollouts and intervene on the local "answer-versus-continue" decision token position. The student's probability of the answer branch falls from $0.74 \rightarrow 0.19 \rightarrow 0.07$ at steps 0, 20, 100, below the teacher's 0.30 on the same prefixes. At update 20, while "routing" is strong, counterfactual continuations from the two branches have similar downstream value (correct 0.30 for the "answer" branch and 0.29 for the continue branch), i.e. the student still maintains reasonable initial answer-reaching abilities. By update 100, however, the continue-branch rarely produces the answer (8% of continuations). The teacher's own continuations from the same prefixes is correct 0.60 of the time: the gap is in the student's degraded execution of the branch, not in the branch itself.

For a complete treatment of how we decompose and estimate Equation (9) see Section E.4.

DeepSeek: stable learning acquiring epistemic behaviors. DeepSeek supplies the clean contrast. Loss falls $0.263 \rightarrow 0.015$, mainly through drift (-0.214) rather than routing (-0.033). Motivated by the epistemic-control requests in Section 3.2, we examine WAIT-family initiation at candidate sentence starts. On each prefix, let w_t and q_W denote student and teacher WAIT-family probabilities. We freeze two classes: *over-use* ($w_0 > q_W$) and *under-use* ($w_0 < q_W$).

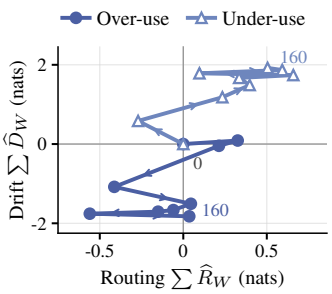


Figure 5: **Conditional checking refinement.** Filled circles: initial over-use; open triangles: initial under-use.

Figure 5 plots cumulative routing and drift of the log-odds $\log[w_t/(1 - w_t)]$. Filled circles move downward: drift suppresses initial over-use. Open triangles move upward: drift promotes initial under-use. The corresponding changes are -1.82 and $+1.92$ nats, with much smaller net routing contributions. Teacher-relative disagreement also decreases on frozen and held-out banks (Section E.2). Thus, unlike fitting on the failed learner's unreliable prefixes, DeepSeek's local improvement refines a reasoning-control behavior and accompanies held-out gains: the student learns *when* to check, not simply to check more. For a more systematic treatment of how well the broadcast works for specific reasoning mechanisms, and an extended study in the safety domain, please refer to Appendix Section F.

Preparation permits a different trajectory. The base Qwen model, prepared with offline distillation training on teacher rollouts, also undergoes substantial μ-motion, but sustains useful learning. Unlike DeepSeek's smoother refinement, its loss and checking behavior are nonmonotone: an excursion and routing-driven overshoot in loss and reasoning (epistemic) behaviors are followed by lower-entropy rollouts and conditional moderation of reasoning behaviors. Its held-out gains suggest that a stronger starting policy with better *continuation-readiness* makes these excursions amenable to correction. Full analysis in Section E.5.

3.4 Preparation: Improving the Starting Point of On-Policy Learning

Section 3.3 suggests a direct intervention. If OPD can fail because early policy changes route the student into reasoning states it cannot yet execute reliably (and resulting in poor supervision and broadcast sources), then improving the student's initial continuation competence may change what the subsequent on-policy loop can reach, even when preparation itself does not necessarily improve

held-out accuracy significantly. Offline distillation supplies teacher-generated reasoning prefixes without requiring the student to reach them first. It may therefore improve continuation skills and reasoning-state management that become useful as OPD changes visitation. This motivates a cold start (Li et al., 2026b) as preparation for subsequent on-policy learning.

We prepare with student/teacher pairings Qwen2.5-1.5B-Instruct/JustRL-1.5B (Qwen et al., 2025; He et al., 2025), Qwen2.5-7B-Instruct/AceReason-Nemotron-7B (Chen et al., 2025), and Llama-3.1-8B-Instruct/Llama-3.1-Nemotron-Nano-8B-v1 (Llama Team, 2024; Bercovich et al., 2025), with students trained on their teachers' solutions to OpenMathInstruct-2 (Toshniwal et al., 2024). We then compare OPD with continued offline distillation on the same DAPO-Math-17k prompt blocks (Yu et al., 2025), matching optimizer steps and sequence presentations. Further details are in Section G.1.

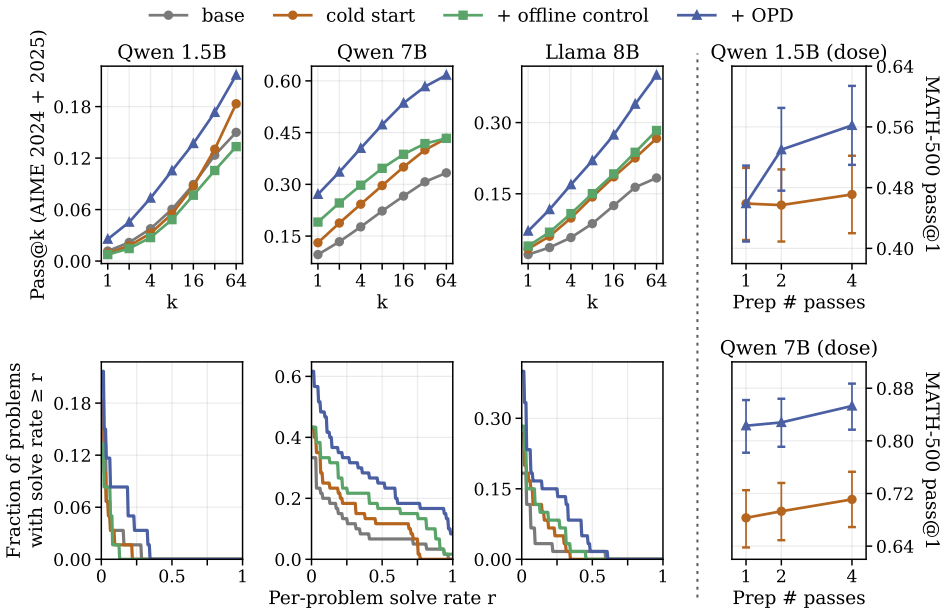


Figure 6: **Offline-prepared OPD improves sampled coverage and reliability, and preparation changes what OPD can achieve.** Left: pooled AIME 2024/2025 pass@k (top) and the fraction of problems with empirical solve rate $\geq r > 0$ (bottom) for the three student–teacher pairs. Right: how preparation influences OPD performance, on MATH. Note how starting accuracy conceals different downstream OPD outcomes. See Section G.1 for further details.

Figure 6 shows higher pass@k accuracy for OPD than offline training through $k = 64$. The gains extend beyond pass@1 to sampled coverage and the distribution of per-problem solve rates. Please refer to Section G.1 for further details. The preparation "dose study" (Figure 6, right) shows offline-preparation's deferred value. At 1.5B, zero to four passes leaves the starting pass rate near 50%, but raises the same 100-step OPD endpoint from 39.3% to 56.2%. A similar trend is observed at 7B.

4 Conclusion

Findings. We studied on-policy distillation as a feedback loop between teacher supervision, student response, and the rollout distribution that supplies future supervision. Our analysis separates two mechanisms in this loop: *conversion*, and *broadcast*. Across our settings, these mechanisms explain both stalled transfer and deterioration despite improving local teacher agreement. Offline preparation changes the starting point of this loop: even with little immediate accuracy improvement, it can enable substantial gains from subsequent OPD.

Limitations and discussion. Our empirical analysis is confined to mathematical reasoning and a select set of LLM models due to resource limitations. We also present a more detailed Q&A in Appendix A to further clarify potential ambiguities in our work.

AI USE STATEMENT

Generative AI tools were also used for language editing, rewriting and presentation. We used generative AI tools to assist with code implementation, debugging, experimental scripting, and parts of experiment execution. The research questions, theoretical development, experimental design, interpretation, and scientific conclusions were determined by the authors. All AI-assisted code and text were reviewed and validated by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Precise experimental details are provided in the appendix, including model specifications, training pipeline and evaluation details. The exact experiment and analysis code, along with the evaluations reported in the paper, are currently being prepared for release.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning, 2023. URL https://arxiv.org/abs/2012.09816.

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc N. M Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking on-policy distillation: Where it helps, where it hurts, and why, 2026. URL https://arxiv.org/abs/2605.10889.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/arora19a.html.

David D. Baek and Max Tegmark. Towards understanding distilled reasoning models: A representational approach, 2025. URL https://arxiv.org/abs/2503.03730.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, and Soumye Singhal. Llama-nemotron: Efficient reasoning models, 2025. URL https://arxiv.org/abs/2505.00949.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter?, 2025. URL https://arxiv.org/abs/2506.19143.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=EgArbnS0BA.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming, 2020. URL https://arxiv.org/abs/1812.07956.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, and Dejian Yang. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes, 2026. URL https://arxiv.org/abs/2603.25562.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, Ning Ding, and Zhiyuan Liu. Justrl: Scaling a 1.5b llm with a simple rl recipe, 2025. URL https://arxiv.org/abs/2512.16649.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Guan Zhe Hong, Nishanth Dikkala, Enming Luo, Cyrus Rashtchian, Xin Wang, and Rina Panigrahy. A implies b: Circuit analysis in llms for propositional logical reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 30749–30803. Curran Associates, Inc., 2025. doi: 10.52202/085713-1033. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2c4477572f06102b2ae3bd2b2ffcebf9-Paper-Conference.pdf.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models, 2026a. URL https://arxiv.org/abs/2603.07079.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=J5i09faOOf.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-policy distillation with curriculum turn-level guidance for multi-turn agents, 2026a. URL https://arxiv.org/abs/2606.15912.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6121–6133. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.494. URL https://aclanthology.org/2020.emnlp-main.494/.

Tao Liu, Taiqiang Wu, Mao Zheng, Xuan Luo, Runming Yang, Xuewei Yang, Junjie Wang, and Yujiu Yang. Simple-OPD: Demystifying warm-up for on-policy distillation, 2026. URL https://arxiv.org/abs/2608.06802.

Meta Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation/.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23610–23641. PMLR, 2023. URL https://proceedings.mlr.press/v202/malladi23a.html.

Sara Vera Marjanovic, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stanczak, and Siva Reddy. Deepseek-r1 thoughtology: Let's think about LLM reasoning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=BZwKsiRnJI.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5142–5151. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/phuong19a.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.10490.

Stephane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning, 2014. URL https://arxiv.org/abs/1406.5979.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, 2014. URL https://arxiv.org/abs/1312.6120.

Zhennan Shen, Yanshu Li, Qingyu Yin, Chak Tou Leong, Zhilin Wang, Yanxu Chen, Rongduo Han, Sunbowen Lee, and Yi R. Fung. On the geometry of on-policy distillation, 2026. URL https://arxiv.org/abs/2606.07082.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026a.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models, 2026b. URL https://arxiv.org/abs/2604.00626.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, and Yuankun Chen. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Akiyoshi Tomihari. Learning dynamics in RL post-training for language models. *arXiv preprint arXiv:2601.04670*, 2026.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data, 2024. URL https://arxiv.org/abs/2410.01560.

Dmitrii Troitskii, Koyena Pal, Chris Wendler, and Callum Stuart McDougall. Internal states before wait modulate reasoning patterns. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18640–18649, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1012. URL https://aclanthology.org/2025.findings-emnlp.1012/.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Base models know how to reason, thinking models learn when, 2025a. URL https://arxiv.org/abs/2510.07364.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Understanding reasoning in thinking language models via steering vectors, 2025b. URL https://arxiv.org/abs/2506.18167.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations. *arXiv preprint arXiv:2607.13399*, 2026.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. Reasoning-finetuning repurposes latent representations in base models, 2025. URL https://arxiv.org/abs/2507.12638.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026. URL https://arxiv.org/abs/2602.12125.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement, 2026. URL https://arxiv.org/abs/2608.09836.

Yaocheng Zhang, Jiajun Chai, Yuqian Fu, Songjun Tu, Xiaohan Wang, Wei Lin, Guojun Yin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Are full rollouts necessary for on-policy distillation?, 2026. URL https://arxiv.org/abs/2605.31490.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes, 2026. URL https://arxiv.org/abs/2605.11182.

Zhou Ziheng, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation, 2026. URL https://arxiv.org/abs/2605.27028.

A MECHANISTIC INTERPRETATION AND KEY QUESTIONS

Q1. What do the response law and decompositions in Sections 3.1 to 3.3 explain beyond their familiar mathematical ingredients?

They provide a connected mechanistic account of how dense local supervision can produce either weak transfer or deteriorating generation. OPD updates the policy that generates its next training population. Understanding its dynamics therefore requires following supervision beyond the current loss, through its effects on future behavior and training histories.

The theory separates three links in this process. First, Section 3.2 explains how a teacher's requests become parameter motion: accessibility determines their expression through the student's tangent map, while coherence determines how the resulting gradients combine. Teacher strength or disagreement alone cannot characterize this conversion. Second, Section 3.3 distinguishes changing a prediction from changing the ancestral decisions that make its history likely. Shared-parameter broadcast can increase exposure to a continuation without making the student better at executing it; the resulting histories then change the conditional supervision received from the same fixed teacher. Third, Equation (9) connects the delivered response to completed outcomes through what the *student's remaining policy* can achieve, separating teacher matching from useful reasoning.

Together, these links explain the structure of the empirical puzzles. The stronger-teacher study exposes inefficient conversion and loss of responsiveness; the degenerate run instead exhibits substantial policy motion, routing into unreliable continuations and then fitting them. The latter mechanism motivates preparing the student for the continuations that OPD will elicit, even when preparation does not improve current accuracy. The theoretical contribution is the integrated explanation that makes these mechanisms distinguishable, connects them to outcomes, and motivates an intervention. Which mechanism occurs is established empirically, by the frozen-request and crossed-history measurements.

Q5. Why is every training run reported with a single seed?

Cost. The runs behind this paper were trained and evaluated on GH200 nodes, and the single-seed set already consumed roughly 4,000 GPU-hours, which is at the limit of our computing resources. OPD is known to be computationally demanding, we therefore designed the experiments so that the main comparisons do not rest on seed variance. The mechanism measurements of Sections 3.2 and 3.3 are exact identities evaluated within a run, on fixed receiver banks and fixed checkpoints, and their uncertainty is reported at the level of problems or sites (paired bootstrap intervals). The outcome comparisons of Section 3.4 are paired within a student-teacher pair on the same problems and replicated across three pairs and two student scales, and the effects we interpret are far larger than the problem-level intervals. What single seeds leave unquantified is run-to-run variation of the training trajectories themselves.

Q3. Why does Section 3.3 need both prefix-level routing/drift (Equation (5)) and trajectory-level continuation advantages (Equation (9)) to explain degeneration?

For loss or entropy, $J_Y = \mathbb{E}_{\nu_t} Y_t(x)$ describes predictions on the currently observed histories. Cross-scoring separates changes in those histories from changes in predictions on the *same* histories. In Figure 4, the initial student is already highly uncertain on the deep histories that later become prevalent, while subsequent fitting sharpens them. On identical batch-41 histories, discrepancy falls $0.79 \to 0.06$ and entropy $4.99 \to 3.99$. Together with the stopping/reach probes, these measurements distinguish routing followed by fitting from uniform entropy-increasing drift (Sections E.3 and E.4.1).

But fitting a prefix and improving the completion from it are different effects. For terminal correctness, $J_Y = \mathbb{E}_{\Pi_t} Y$, and the conditional value $V_t^Y(x)$ depends on *every downstream decision*. At an intermediate cut, its derivative separates changes in reaching the cut from changes in the remaining policy. Unrolling the latter gives the continuation-advantage weighting in Equation (9); replacing it with a derivative of the logits at x alone would omit downstream effects (Section C.8). Thus, a teacher-preferred next token need not help if the student cannot complete the reasoning it initiates. The prefix analysis establishes the observed routing-then-fitting pattern; the outcome analysis explains why local fitting need not restore correctness. Entropy enters the argument only as a prefix-level readout.

Q4. Could the responsiveness changes in Section 3.2 (Figure 3) simply reflect changing teacher–student disagreement or generic kernel movement?

A live accessibility measure can change because either the requests or the kernel change. We therefore fix the initial numerical requests B_0, exact histories, and output coordinates across checkpoints. Since only the kernel varies in $A_t[B_0]$, its change measures responsiveness to the *same supervision*, at a fixed discrepancy and on the same training text (Section C.4).

To test whether eigenspace movement reflects teacher-specific adaptation, we evaluate both teachers' frozen fields in the *same checkpoint kernel* and compare their projection gains, separating preferential reorientation from movement that benefits both fields. Healthy 1.5B exhibits preferential reorientation; healthy 7B mainly amplifies existing directions; stalled runs lose sensitivity to the fixed requests (Section D.5). These findings complement the initial accessibility/coherence decomposition, which exposes conversion differences concealed by comparable gradient norms.

Q5. What does preparation in Section 3.4 and Figure 6 contribute beyond offline warm-up, and how does Section 3.3 motivate it?

The contribution is an explanation of why the value of a warm-up stage, already used in prior work (Section B), should be assessed through *subsequent learning*. The failure mechanism identifies a mismatch between making extended reasoning likely and being able to execute it. Offline distillation supplies teacher-generated reasoning histories without requiring the initial student to generate them first. It can therefore prepare continuation skills and reasoning-state management that become useful as OPD changes visitation: *continuation readiness*, a benefit beyond higher starting accuracy – our pass@k results clearly demonstrate this.

B EXPANDED RELATED WORK

B.1 CORE RELATED WORKS

Distillation and the OPD loop. Knowledge distillation transfers teacher outputs (Hinton et al., 2015), including teacher-generated sequences (Kim & Rush, 2016). DAgger, the seminal online imitation algorithm, addresses distribution shift through expert feedback on learner-visited states (Ross et al., 2011); Lin et al. (2020) develop this connection for autoregressive distillation. MiniLLM uses reverse KL and student-generated samples (Gu et al., 2024), GKD studies self-generated sequences with flexible divergences (Agarwal et al., 2024), and Lu & Thinking Machines Lab (2025) demonstrate practical OPD for reasoning and personalization. These methods motivate our question: how does dense supervision on student histories produce successful transfer, stagnation, or degeneration? We develop a systematic first-principles account of the coupled loop through which shared-parameter updates change predictions, renew visited histories, and reshape subsequent supervision.

Kernel response laws support an analysis of the complete on-policy feedback. The neural tangent kernel connects parameter learning to function-space change (Jacot et al., 2018). Malladi et al. (2023) analyze language-model finetuning through kernels, including an Adam extension; Ren & Sutherland (2025) decompose prediction changes through normalization, empirical-kernel, and loss-gradient terms for SFT and preference finetuning, including offline DPO. Tomihari (2026) studies update propagation in RL post-training. OPD needs different theoretical and empirical machinery for two reasons. Its on-policy sampling moves the learning distribution itself, so the analysis must track this motion, which is absent from the offline algorithms (SFT and offline DPO) studied by Ren & Sutherland (2025). Its supervision is also dense and token-level and can be disconnected from the reasoning outcome (such as accuracy), whereas RL gives only sparse, outcome-level learning signals.

Nevertheless, the fundamental response law of the shared-parameter neural learner is shared with Ren & Sutherland (2025). We then follow the induced motion through the rest of the OPD loop: changes at ancestor decisions alter reach probabilities, renew the training source, and thereby change the requests driving future updates. Combining the evolving kernel, routing/drift split, and continuation-value analysis yields a mechanistic account of how optimization changes future supervision and outcomes (Sections 3 and C).

Cumulative weight geometry differs from the geometry that converts supervision. Shen et al. (2026) find low stable rank and early stabilization of dominant subspaces in cumulative weight-matrix changes; constrained-training experiments test those subspaces' functional sufficiency. We instead analyze the current logit Jacobian and the response it delivers to specified teacher requests. The Jacobian can change even when cumulative displacement is low-rank. Frozen histories, coordinates, and requests isolate amplification and reorientation; crossed teacher fields test preferential adaptation in the same kernel (Sections C.4 and D.5). Distinguishing local accessibility from collective coherence further explains why individually expressible requests can cancel. These analyses concern how geometry enables or obstructs transfer.

Connecting local fitting to task performance requires trajectory-level continuation analysis. Policy-gradient theory (Sutton et al., 1999) and cost-to-go-aware imitation learning (Ross & Bagnell, 2014) connect actions to subsequent outcomes. G-OPD interprets distillation through KL-constrained RL and reward extrapolation (Yang et al., 2026). Armandpour et al. (2026) use student continuation-success probabilities to derive an ideal local logit direction and measure its alignment with distillation signals. Our analysis connects this outcome perspective to *delivered* motion: $\sum_s K_t(x,s)\Gamma_t(s)$ includes the effects of all coupled source prefixes in addition to the request at x. Crucially, prefix readouts $\mathbb{E}_{\nu_t} Y_t(x)$ and terminal outcomes $\mathbb{E}_{\Pi_t} Y$ are different objects. At an intermediate cut, terminal change splits into reach and continuation change; unrolling the latter and recombining both terms yields the advantage-weighted contribution of every decision (Equation (9) and Section C.8). This separates reaching a history, fitting there, and completing it successfully—the distinction underlying the degeneration mechanism and continuation readiness.

Degeneration involves source feedback as well as supervision quality. Fu et al. (2026) study unreliable supervision and implementation choices; Jin et al. (2026a) augment high-entropy teacher supervision with forward KL; and Zhu et al. (2026) analyze signals induced by top-K truncation, renormalization, and stop-gradient conventions. Li et al. (2026b) also observe deterioration with prefix depth. Contemporaneous work by Wang et al. (2026) studies teacher mismatch and length

exploitation. Our first-principled analysis reveals a phased failure mechanism: early routing makes already-uncertain continuations prevalent, after which substantial fitting can *sharpen those very histories* without restoring useful generation (Sections E.3 and E.4.1). Our routing-drift machinery, coupled with the learning-value decomposition (for terminal-accuracy readout) distinguish the failure mechanism from a naive, uniform entropy-increasing learning process. The important feedback is that the histories made common by earlier updates become the source of later supervision. The prepared learner's different trajectory further shows why source movement itself is not synonymous with failure, and its continuation-readiness helps in exposing useful subsequent learning.

Offline preparation can improve readiness for learning without improving immediate accuracy. Li et al. (2026b) motivate cold-start SFT by reducing thinking-pattern mismatch and improving overlap. Their §5.1/Figure 8 studies a strong standalone-SFT regime: the SFT initialization already exceeds the direct-OPD endpoint on all three plotted benchmarks, with little further gains from OPD. We study a different regime, where the student-teacher gap is significant, and preparation leaves starting accuracy near the unprepared student's but changes subsequent OPD from deterioration to substantial improvement. The principled motivation is *continuation readiness*: the outcome analysis and observed failure mechanisms in Section 3.3 identify a distinction between increasing a history's reach and executing its continuation successfully. Teacher-generated trajectories provide training on such histories before the student must reach them itself. Our dose study exposes deferred benefits and a sharply nonmonotonic stability pattern at 7B; the prepared-run analysis traces a different subsequent routing/fitting process (Table 14 and Section E.5).

In *contemporaneous work* Simple-OPD (Liu et al., 2026) studies warm-up data, duration, and LoRA parameterization, including benefits from incorrect teacher-compatible traces and differences between SFT and downstream rankings. These findings complement our mechanistic account linking preparation to source feedback and continuation values. The key result is an altered learning trajectory; the preparation dose-response experiments demonstrate this interesting benefit.

B.2 More related works

Reasoning interpretability. Part of the focus of our work is on how LLMs learn to reason (from a teacher model), and part of our analysis methodology is inspired by the discoveries in the reasoning interpretability community. Venhoff et al. (2025a) is especially relevant, as they discovered that RL-trained models tend to acquire the ability to manage "when" to fire a reasoning mechanism (be it an "epistemic control" mechanism such as backtracking or planning, or computation mechanisms like estimating an inequality); Marjanovic et al. (2026) directly studies the chain-of-thought reasoning mechanisms of DeepSeek-R1 models, which our DeepSeek student models are distilled from, also inspiring our analysis on the epistemic control transmission in OPD. More broadly, Bogdan et al. (2025); Baek & Tegmark (2025); Venhoff et al. (2025b); Troitskii et al. (2025); Hong et al. (2025); Ward et al. (2025) identify causal mechanisms behind LLMs' reasoning behaviors in logic, mathematics and other related domains.

OPD and its variants. On-policy distillation is becoming increasingly popular in industry post-training pipelines (DeepSeek-AI et al., 2026; Team et al., 2026). Many variants are being proposed, including selecting which tokens to perform supervision on (Jin et al., 2026b; Yu et al., 2026), and curriculum-styled variants (Ziheng et al., 2026; Zhang et al., 2026; Li et al., 2026a). The survey on OPD Song & Zheng (2026b) is also a good reference for understanding the research landscape in this area.

Learning dynamics. Many classic works lie in this space and are related to the lineage of LLM-era learning dynamics work which inspired the seminal Ren & Sutherland (2025). Saxe et al. (2014); Lee et al. (2019); Phuong & Lampert (2019); Chizat et al. (2020); Allen-Zhu & Li (2023) are all representative works in this space.

C Derivations and proofs for the main text

This appendix follows Sections 2 and 3. It gives the assumptions and proofs for the main-text identities, then the specializations used to interpret the experiments. The measurement protocols are in Sections D, E and G. The matched-pair results used only by "The broadcast is coarse" remain with that extension in Section F.1.

Main-text result	Complete derivation
Section 2: policies and request field	Section C.1: finite-tree conventions and exact-loss request
Section 3.1, Equation (2)	Section C.2: exact response and explicit finite-step remainder
Section 3.2, Equations (3) and (4)	Sections C.3 and C.4: conversion and frozen-field geometry
Section 3.3, Equations (5) and (6)	Section C.5: routing and conditional drift
Section 3.3, Equation (7)	Section C.6: source normalization and support
Section 3.3, Equation (8)	Section C.7: endpoint accounting and path-integral error
Section 3.3, Equation (9)	Section C.8: continuation values and terminal outcomes
Section 3.3: entropy versus fitting	Section C.9: entropy and checking sensitivities

C.1 Setup: policies, requests, and standing assumptions

Connection to Section 2. We distinguish the model distribution, the sampling distribution, and the numerical field pulled back into parameter space. This distinction is needed for both the proofs and the experimental diagnostics.

Notation. A prompt is denoted by b, a source prefix by s, and a receiver prefix by x. A complete trajectory is ξ; $\tau > 0$ is reserved for sampling temperature. The integer k indexes optimizer updates, j indexes generated-token depth, and t indexes continuous training time. All gradients of scalar functions are column vectors. The matrix $\Phi_\theta(x) \in \mathbb{R}^{|\mathcal{V}|\times d}$ has row a equal to $\nabla_\theta z_\theta(x,a)^\top$, so $K_\theta(x,s) = \Phi_\theta(x)\Phi_\theta(s)^\top$. We write $\mathsf{A}^p = I - \mathbf{1}p^\top$ for the log-softmax normalization matrix, A for scalar accessibility, and $\mathcal{A}^Y$ for continuation advantage. A logit-space readout gradient is denoted by $\chi \in \mathbb{R}^{|\mathcal{V}|}$; its parameter-space pullback is $\psi = \Phi^\top\chi \in \mathbb{R}^d$. Norms without a subscript are Euclidean; $\|\cdot\|_{\mathrm{op}}$ denotes the induced operator norm.

Assumption C.1 (Finite generation trees and regular parameter paths)**.** The prompt set $\mathcal{B}$ and vocabulary $\mathcal{V}$ are finite, and generation has a finite cap $H_{\max}$. Each prefix contains its prompt and its full generated ancestry. Appending a token is deterministic. A node is terminal after a stopping token or at the cap, and has no further decisions. The terminal rule and chat template are fixed. The prompt law β is fixed; prompts of zero probability are omitted.

The parameter path $\theta : I \to \mathbb{R}^d$, on a compact interval I, is continuous and piecewise C^1 with finitely many pieces. For every prefix under consideration, $z_\theta(x)$ is C^2 on an open neighborhood of the path and of any parameter segment used for a finite-step calculation. A third-order expansion will explicitly assume C^3. The model distribution is the full-vocabulary policy $p_\theta = \mathrm{softmax}(z_\theta)$. The teacher q is fixed and is positive on every coordinate used in an exact-KL calculation.

For the sampling identities, each nonterminal x has a fixed nonempty allowed set $\mathcal{V}_\pi(x) \subseteq \mathcal{V}$ and

$$\pi_\theta(a \mid x) = \frac{\exp(z_\theta(x,a)/\tau)}{\sum_{c\in\mathcal{V}_\pi(x)}\exp(z_\theta(x,c)/\tau)} \quad (a \in \mathcal{V}_\pi(x)), \qquad \pi_\theta(a \mid x) = 0 \quad (a \notin \mathcal{V}_\pi(x)).$$

The temperature $\tau > 0$, allowed sets, and any observation or source masks are fixed. The full-vocabulary sampler is the case $\mathcal{V}_\pi(x) = \mathcal{V}$.

Let $\mathcal{S}$ and $\mathcal{T}$ be the resulting finite sets of nonterminal and terminal prefixes. A trajectory $\xi = (b, a_0, \dots, a_{N-1})$ ends at its first terminal node, with $N \le H_{\max}$, and has probability

$$\Pi_t(\xi) = \beta(b)\prod_{j=0}^{N-1}\pi_t(a_j \mid x_j).$$

For any node x, including a terminal node, $d_t(x)$ is its reach probability. It is the same product truncated at depth $h(x)$. On terminal nodes $d_t(x) = \Pi_t(x)$; on nonterminal nodes the sum $\sum_x d_t(x)$

is generally not one. Every allowed trajectory has positive probability. Finiteness and the stated regularity justify every interchange of sums and derivatives below. Identities for derivatives hold on each smooth piece, hence almost everywhere, and integrate across the entire interval. The results take the path as given and are independent of any existence or uniqueness theorem for an autonomous training flow.

The raw direction is

$$g_t(s) = \Phi_t(s)^\top B_t(s), \qquad u_t = \sum_s \mu_t(s) g_t(s), \qquad \Gamma_t(s) = \mu_t(s) B_t(s).$$

Here B_t is a declared numerical logit-space field. The specialization $\dot\theta_t = u_t$ is called the raw flow. An actual optimizer path uses its realized $\dot\theta_t$. Restricted diagnostic fields are extended by zero outside their recorded token coordinates, and their Jacobians are restricted consistently; these diagnostic supports leave the sampler's allowed set unchanged.

Lemma C.2 (Log-softmax differentiation and centering)**.** *Write $\pi = \pi_\theta(\cdot \mid x)$ as a vector in the full vocabulary, with zeros outside its fixed allowed set, and let $\mathsf{A}^\pi = I - \mathbf{1}\pi^\top$. For each allowed token a,*

$$\nabla_\theta \log \pi_\theta(a \mid x) = \tau^{-1} \Phi_\theta(x)^\top (e_a - \pi), \qquad \sum_a \pi(a \mid x) \nabla_\theta \log \pi(a \mid x) = 0.$$

Consequently $\zeta_t(x,a) := \partial_t \log \pi_t(a \mid x)$ satisfies $\sum_a \pi_t(a \mid x)\zeta_t(x,a) = 0$. If $\mathbf{1}^\top w = 0$, then $(\mathsf{A}^\pi)^\top w = w$. Sums involving $\log \pi$ or ζ range only over allowed tokens.

Proof. For allowed a, the logarithm of the displayed sampler in Assumption C.1 is

$$\log \pi(a \mid x) = \tau^{-1} z(x,a) - \log \sum_{c \in \mathcal{V}_\pi(x)} \exp(z(x,c)/\tau).$$

Its partial derivative with respect to any full-vocabulary logit $z(x,c)$ is $\tau^{-1}(\mathbf{1}\{a = c\} - \pi(c \mid x))$, including the value zero when c is disallowed. The chain rule proves the gradient formula. Averaging over allowed a yields

$$\tau^{-1} \Phi(x)^\top \left(\sum_a \pi(a \mid x) e_a - \pi \sum_a \pi(a \mid x) \right) = 0.$$

Pairing this equality with $\dot\theta_t$ gives the zero-mean velocity. Finally, $(\mathsf{A}^\pi)^\top w = (I - \pi\mathbf{1}^\top)w = w - \pi(\mathbf{1}^\top w) = w$. □

Proposition C.3 (Exact reverse-KL request)**.** *At a fixed prefix, let $c = \sum_a p_a \log(p_a/q_a)$, $\ell_a = \log(p_a/q_a)$, and $\bar\ell = \sum_a p_a \ell_a$, with $p = \mathrm{softmax}(z)$ and $q_a > 0$. Then*

$$B_a = -\partial_{z_a} c = p_a(\bar\ell - \ell_a), \qquad \mathbf{1}^\top B = 0, \qquad g = \Phi^\top B = -\nabla_\theta c. \tag{10}$$

For fixed source weights μ, $u = \sum_s \mu(s) g(s)$ is the negative gradient of $\sum_s \mu(s) c(s)$.

Proof. Positivity makes the logarithms differentiable. Since $\partial_{z_a} p_b = p_b(\mathbf{1}\{a = b\} - p_a)$ and $\partial_{p_b}[p_b \log(p_b/q_b)] = \ell_b + 1$, we have

$$\begin{aligned} \partial_{z_a} c &= \sum_b p_b(\mathbf{1}\{a = b\} - p_a)(\ell_b + 1) \\ &= p_a(\ell_a + 1) - p_a \sum_b p_b(\ell_b + 1) = p_a(\ell_a - \bar\ell). \end{aligned}$$

Thus $\sum_a B_a = \bar\ell \sum_a p_a - \sum_a p_a \ell_a = 0$. The chain rule gives $\nabla_\theta c = \Phi^\top \nabla_z c$; linearity of the finite sum proves the last assertion when μ is held fixed. □

This is a full-distribution identity. For an implemented objective, such as a truncated-support loss or detached-coefficient surrogate, B must be that objective's own complete numerical negative logit derivative, including the softmax coupling. For the conversion diagnostic, B denotes the explicitly specified audit field in Section D.3. The response and conversion results hold for this field's pullback, whether or not that pullback coincides with exact-KL descent. Holding μ fixed applies to computing one update; subsequent source populations can change.

C.2 Response law: exact finite steps and controlled remainders

Connection to Section 3.1, Equation (2). The integral is exact for an arbitrary realized step. The main-text kernel expression is its raw-gradient, first-order specialization.

Proposition C.4 (Exact response and Taylor remainder)**.** *Let $\Delta\theta = \theta_{k+1} - \theta_k$, and let $\theta_r = \theta_k + r\Delta\theta$ for $0 \le r \le 1$. For any fixed receiver x and token a,*

$$\delta_k^p(x,a) = \int_0^1 \nabla_\theta \log p_{\theta_r}(a \mid x)^\top \Delta\theta \, dr. \tag{11}$$

If $M_k(x,a) = \sup_{0\le r\le 1} \|\nabla_\theta^2 \log p_{\theta_r}(a \mid x)\|_{\mathrm{op}}$, then

$$\delta_k^p(x,a) = \nabla_\theta \log p_k(a \mid x)^\top \Delta\theta + R_k(x,a), \qquad |R_k(x,a)| \le \tfrac{1}{2} M_k(x,a)\|\Delta\theta\|^2.$$

The corresponding exact logit displacement is $z_{k+1}(x) - z_k(x) = \int_0^1 \Phi_{\theta_r}(x)\Delta\theta \, dr$.

Proof. Define $f(r) = \log p_{\theta_r}(a \mid x)$. Its first two derivatives are

$$f'(r) = \nabla_\theta \log p_{\theta_r}(a \mid x)^\top \Delta\theta, \qquad f''(r) = \Delta\theta^\top \nabla_\theta^2 \log p_{\theta_r}(a \mid x)\Delta\theta.$$

The fundamental theorem of calculus gives Equation (11). Applying it also to f' and changing the order of integration over $0 \le v \le r \le 1$ gives

$$f(1) - f(0) - f'(0) = \int_0^1 \int_0^r f''(v)\, dv\, dr = \int_0^1 (1-v) f''(v)\, dv.$$

The Hessian is continuous on a compact segment, so $M_k(x,a) < \infty$. The bound follows from $|f''(v)| \le M_k(x,a)\|\Delta\theta\|^2$ and $\int_0^1 (1-v)\, dv = 1/2$. Applying the same first-derivative calculation to each logit proves the vector identity. □

Corollary C.5 (Small-step response law)**.** *For $\Delta\theta_k = \eta u_k$ with $u_k = \sum_s \mu_k(s)\Phi_k(s)^\top B_k(s)$,*

$$\delta_k^p(x,a) = \eta \sum_s \mu_k(s)[\mathsf{A}_k^p(x) K_k(x,s) B_k(s)]_a + R_k(x,a), \qquad |R_k(x,a)| \le \tfrac{1}{2}\eta^2 M_k(x,a)\|u_k\|^2.$$

For fixed θ_k, u_k and sufficiently small η, the remainder is $O(\eta^2)$, as in Equation (2)*.*

Proof. At temperature one, Lemma C.2 gives $\nabla_\theta \log p_k(a \mid x)^\top = e_a^\top \mathsf{A}_k^p(x)\Phi_k(x)$. Substituting u_k in Proposition C.4 therefore yields

$$\nabla_\theta \log p_k(a \mid x)^\top u_k = \sum_s \mu_k(s) e_a^\top \mathsf{A}_k^p(x)\Phi_k(x)\Phi_k(s)^\top B_k(s),$$

which is the asserted kernel term. The remainder bound follows by substituting $\Delta\theta_k = \eta u_k$. To justify the asymptotic notation, choose $\eta_0 > 0$ so that all segments with $|\eta| \le \eta_0$ lie in a compact neighborhood where the Hessian is continuous. Its supremum on this neighborhood supplies a single bound independent of η. A family of varying directions requires a uniform bound on their norms as well. □

Velocity and the optimizer. At every differentiable time on an actual parameter path,

$$v_t(x) := \dot z_t(x) = \Phi_t(x)\dot\theta_t, \qquad \zeta_t(x,a) = \tau^{-1}(e_a - \pi_t(\cdot \mid x))^\top v_t(x).$$

These follow directly by the chain rule and Lemma C.2. Under gradient flow, substituting $\dot\theta_t = u_t$ gives $v_t(x) = \sum_s K_t(x,s)\Gamma_t(s)$ and hence Equation (6). With momentum, preconditioning, clipping, or weight decay, the actual $\dot\theta_t$ in general differs from the raw u_t. However, with the rich tradition of theoretical analysis on the learning dynamics of neural networks including the seminal Ren & Sutherland (2025); Jacot et al. (2018); Arora et al. (2019), gradient-flow analysis is still genuinely informative in obtaining rich theoretical insights. We leave extensions to the more convoluted optimizer settings to future work.

C.3 COLLECTIVE RESPONSE AND ITS GEOMETRY

Connection to Section 3.2, Equations (3) and (4). We fix the checkpoint, output coordinates, and normalized bank ν in this subsection. Zero-weight sites are omitted. The space $\mathcal{H}_\nu$ of fields on the remaining finite bank has the inner product

$$\langle F, G\rangle_\nu = \sum_s \nu(s) F(s)^\top G(s), \qquad \|F\|_\nu^2 = \langle F, F\rangle_\nu.$$

Define $\mathcal{J} : \mathbb{R}^d \to \mathcal{H}_\nu$ by $(\mathcal{J}v)(s) = \Phi(s)v$. This is a finite-dimensional inner-product space, so orthogonal projections and adjoints are well-defined.

Proposition C.6 (Collective response and conversion factorization)**.** *Let $g(s) = \Phi(s)^\top B(s)$, $u = \mathbb{E}_\nu g$, $P = \|B\|_\nu^2$, and $D_g = \mathbb{E}_\nu\|g\|^2$. The adjoint is $\mathcal{J}^* B = u$, and $\mathcal{K} = \mathcal{J}\mathcal{J}^*$ is self-adjoint and positive semidefinite. For the same-bank raw direction and $P > 0$,*

$$\dot z = \mathcal{K}B = \Xi B + r, \qquad \Xi = \frac{\langle B, \mathcal{K}B\rangle_\nu}{P} = \frac{\|u\|^2}{P} \geq 0, \qquad \langle B, r\rangle_\nu = 0.$$

If also $D_g > 0$, then

$$A = \frac{D_g}{P}, \qquad C = \frac{\|u\|^2}{D_g}, \qquad \Xi = AC, \qquad \|u\|^2 = PAC, \qquad 0 \leq C \leq 1.$$

Here $C = 1$ exactly when $g(s) = u$ at every positive-weight site, and $C = 0$ exactly when $u = 0$. If $P > 0$ but $D_g = 0$, then $u = 0$, $A = \Xi = 0$, and C is undefined. If $P = 0$, then $B = g = u = 0$ on the bank and A, C, Ξ are not defined by these ratios.

Proof. For any field F and parameter vector v,

$$\langle F, \mathcal{J}v\rangle_\nu = \sum_s \nu(s) F(s)^\top \Phi(s) v = \left(\sum_s \nu(s)\Phi(s)^\top F(s)\right)^\top v.$$

This proves the adjoint formula. It also gives

$$\langle F, \mathcal{K}G\rangle_\nu = (\mathcal{J}^*F)^\top(\mathcal{J}^*G) = \langle \mathcal{K}F, G\rangle_\nu, \qquad \langle F, \mathcal{K}F\rangle_\nu = \|\mathcal{J}^*F\|^2 \geq 0.$$

By the chain rule, the logit velocity generated by u is $\mathcal{J}u = \mathcal{K}B$. Thus $\langle B, \dot z\rangle_\nu = \|u\|^2$. Since $P > 0$, the strictly convex quadratic $\alpha \mapsto \|\dot z - \alpha B\|_\nu^2$ has its unique minimum at $\alpha = \langle B, \dot z\rangle_\nu / P$. This is Ξ, and the residual has the stated orthogonality.

For $D_g > 0$, multiplication of the defined ratios gives $\Xi = AC$ and $\|u\|^2 = PAC$. Expanding a centered squared norm gives

$$\mathbb{E}_\nu\|g - u\|^2 = \mathbb{E}_\nu\|g\|^2 - 2u^\top \mathbb{E}_\nu g + \|u\|^2 = D_g - \|u\|^2 \geq 0.$$

This proves $C \leq 1$ and its equality condition; nonnegativity of C and its zero condition follow from its numerator. If $D_g = 0$, every nonnegative summand $\nu(s)\|g(s)\|^2$ vanishes, so $g = u = 0$. If $P = 0$, the same argument first gives $B = 0$, then $g = u = 0$. These cases cannot be handled by cancelling zero denominators. □

Proposition C.7 (Joint singular modes)**.** *Stack the blocks $\sqrt{\nu(s)}\Phi(s)$ into $\mathsf{J} \in \mathbb{R}^{N_\nu \times d}$ and $\sqrt{\nu(s)}B(s)$ into $\mathbf{B} \in \mathbb{R}^{N_\nu}$. Let the compact singular-value decomposition be $\mathsf{J} = U\Sigma V^\top$, of rank r, with orthonormal columns U_j, V_j and positive singular values σ_j. If $P > 0$,*

$$u = \sum_{j=1}^r \sigma_j (U_j^\top \mathbf{B}) V_j, \qquad \Xi = \sum_{j=1}^r \sigma_j^2 \frac{|U_j^\top \mathbf{B}|^2}{\|\mathbf{B}\|^2}. \tag{12}$$

The component of $\mathbf{B}$ orthogonal to $\mathrm{range}(\mathsf{J})$ produces zero parameter response.

Proof. The stacking gives $\mathsf{J}^\top\mathbf{B} = \sum_s \nu(s)\Phi(s)^\top B(s) = u$ and $\|\mathbf{B}\|^2 = P$. Substitution of the singular-value decomposition proves the first identity. Since $V_i^\top V_j = \mathbf{1}\{i = j\}$, its squared norm is $\sum_j \sigma_j^2 |U_j^\top \mathbf{B}|^2$; division by P proves the second. Finally $\mathsf{J}^\top \mathbf{B}_\perp = V\Sigma U^\top \mathbf{B}_\perp = 0$ when $U^\top \mathbf{B}_\perp = 0$. If $r = 0$, these are empty sums and $u = \Xi = 0$. □

Joint modes span multiple prefixes and tokens. Their parameter-space components are orthogonal, whereas individual per-prefix gradients need not be. Strong cancellation between local gradients is therefore compatible with large self-kernel responses but a small joint response; independently sorted self-kernel spectra do not determine Equation (12).

Proposition C.8 (Coherence and interference)**.** *For independent $s, s' \sim \nu$,*

$$\|u\|^2 = \mathbb{E}\langle g(s), g(s')\rangle = \mathbb{E}_\nu\|g(s)\|^2 - \tfrac{1}{2}\mathbb{E}\|g(s) - g(s')\|^2.$$

If $B = \sum_{m=1}^{M} B^{(m)}$ is any finite field decomposition and $u^{(m)} = \mathbb{E}_\nu[\Phi^\top B^{(m)}]$, then $\|u\|^2 = \sum_m \|u^{(m)}\|^2 + 2\sum_{m<n}\langle u^{(m)}, u^{(n)}\rangle$.

Proof. Independence gives the explicit finite-sum identity

$$\mathbb{E}\langle g(s), g(s')\rangle = \sum_{s,s'} \nu(s)\nu(s')g(s)^\top g(s') = \left(\sum_s \nu(s)g(s)\right)^\top \left(\sum_{s'} \nu(s')g(s')\right).$$

This is $\|u\|^2$. Expanding the squared difference and using the identical marginals yields $\mathbb{E}\|g(s) - g(s')\|^2 = 2D_g - 2\|u\|^2$. The field decomposition gives $u = \sum_m u^{(m)}$ by linearity, and the last formula follows by expanding the square and collecting both copies of each off-diagonal inner product. □

Proposition C.9 (Finite-sample correction of the mean-gradient energy)**.** *Let $g_1, \dots, g_n$, $n > 1$, be independent identically distributed parameter vectors with finite second moment and mean u. Define $\bar{g} = n^{-1}\sum_i g_i$ and $\widehat{D}_g = n^{-1}\sum_i \|g_i\|^2$. Then*

$$\widehat{U} = \frac{n\|\bar{g}\|^2 - \widehat{D}_g}{n-1} = \frac{1}{n(n-1)}\sum_{i\neq j} g_i^\top g_j$$

has expectation $\|u\|^2$.

Proof. Expanding $n^2\|\bar{g}\|^2$ gives $\sum_i \|g_i\|^2 + \sum_{i\neq j} g_i^\top g_j$. Subtracting $n\widehat{D}_g = \sum_i \|g_i\|^2$ and dividing by $n(n-1)$ proves the second expression. Independence gives $\mathbb{E}[g_i^\top g_j] = (\mathbb{E}g_i)^\top(\mathbb{E}g_j) = \|u\|^2$ for each of its $n(n-1)$ terms. Thus its expectation is $\|u\|^2$. The finite second moment makes every term integrable. Positive clipping and division by the random $\widehat{D}_g$ are nonlinear, so the bias of the resulting coherence ratio falls outside this linear expectation calculation. Other sampling designs, including sampling without replacement from a fixed finite population, require their own design-specific bias calculation. □

Policy-level observation of a logit response. For a fixed vector $w(x)$ with $\mathbf{1}^\top w(x) = 0$,

$$w(x)^\top[z_{\theta'}(x) - z_\theta(x)] = w(x)^\top[\log p_{\theta'}(\cdot \mid x) - \log p_\theta(\cdot \mid x)].$$

Indeed, $\log p_\theta = z_\theta - \log\sum_a e^{z_\theta(a)}\mathbf{1}$, and both scalar-normalizer terms disappear under $w^\top$. The identity holds for finite displacements, for zero-extended contrasts on a recorded support, and for centered and uncentered training fields alike. The experimental contrast is specified in Section D.3.

C.4 Why the frozen spectral probe measures accessibility

Connection to Section 3.2 and the fixed-bank diagnostic in Section 3.4. All time dependence below is in the tangent map; the bank, coordinates, and numerical field are fixed.

Proposition C.10 (Frozen-field Rayleigh expansion)**.** *Let $P_0 = \mathbb{E}_{\nu_0}\|B_0\|^2 > 0$, and define on nonzero-request sites*

$$\omega_0(s) = \frac{\nu_0(s)\|B_0(s)\|^2}{P_0}, \qquad b_0(s) = \frac{B_0(s)}{\|B_0(s)\|}.$$

Zero-request sites have weight zero and are omitted from sums involving b_0. For any nonzero field on a specified bank,

$$A[B] = \frac{\mathbb{E}_\nu[B(s)^\top K(s,s)B(s)]}{\mathbb{E}_\nu\|B(s)\|^2}. \tag{13}$$

Let $(\lambda_{j,t}(s), v_{j,t}(s))$ be a complete orthonormal eigendecomposition of $K_t(s,s)$. Then

$$a_t(s) = b_0(s)^\top K_t(s,s) b_0(s) = \sum_j \lambda_{j,t}(s) |v_{j,t}(s)^\top b_0(s)|^2,$$
$$A_t[B_0] = \sum_s \omega_0(s) a_t(s). \tag{14}$$

The weights sum to one. In particular, $A_t[B_0]$ is constant whenever the self-kernels are constant on the bank.

Proof. For each site, $\|\Phi(s)^\top B(s)\|^2 = B(s)^\top \Phi(s)\Phi(s)^\top B(s) = B(s)^\top K(s,s) B(s)$. Substituting into the definition $A = D_g/P$ proves Equation (13). In the frozen case, write $B_0(s) = \|B_0(s)\| b_0(s)$ at each nonzero site and divide the weighted sum by P_0; this gives $A_t[B_0] = \sum_s \omega_0(s) b_0(s)^\top K_t(s,s) b_0(s)$. The weights sum to $P_0/P_0 = 1$.

Each $K_t(s,s) = \Phi_t(s)\Phi_t(s)^\top$ is symmetric and positive semidefinite: for every vector v, $v^\top K_t(s,s) v = \|\Phi_t(s)^\top v\|^2 \geq 0$. The finite-dimensional spectral theorem therefore gives $K_t = \sum_j \lambda_{j,t} v_{j,t} v_{j,t}^\top$ with $\lambda_{j,t} \geq 0$. Inserting this sum into the quadratic form proves Equation (14). All coefficients except K_t are fixed, which proves the constancy assertion. Repeated eigenvalues do not affect the full sum: the sum of squared projections within an eigenspace is basis-invariant. □

The live quantity $A_t[B_t]$ can also change through its field and energy weights. The frozen collective gain is $\Xi_t[B_0] = A_t[B_0] C_t[B_0]$ when the denominators are nonzero; both factors use B_0, whereas $A_t[B_0] C_t[B_t]$ mixes two fields and is a different quantity.

C.5 Routing and conditional drift on moving prefix populations

Connection to Section 3.3, Equations (5) and (6). Let $m^{\mathrm{obs}}(x) \geq 0$ be a fixed weight on the nonterminal prefixes of the observation tree. Assume at least one eligible prefix has positive reach probability, and define

$$\nu_t(x) = \frac{m^{\mathrm{obs}}(x) d_t(x)}{Z_t^{\mathrm{obs}}}, \qquad Z_t^{\mathrm{obs}} = \sum_x m^{\mathrm{obs}}(x) d_t(x) > 0. \tag{15}$$

This is normalization of expected weighted counts, which allows fixed depth masks, context classes defined by a fixed reference learner, and empirical minibatches of unequal length.

Proposition C.11 (Routing and conditional drift)**.** *Let $Y_t(x)$ be a scalar readout that is continuous and piecewise C^1 along the parameter path, and put $J_Y(t) = \sum_x \nu_t(x) Y_t(x)$. For a positive-reach prefix with ancestry $(x_j, a_j)_{j<h(x)}$, define*

$$\lambda_t(x) = \partial_t \log d_t(x) = \sum_{j<h(x)} \zeta_t(x_j, a_j).$$

Then, almost everywhere,

$$\dot\nu_t(x) = \nu_t(x)\big(\lambda_t(x) - \mathbb{E}_{\nu_t}\lambda_t\big),$$
$$\dot J_Y(t) = R_Y(t) + D_Y(t), \qquad R_Y(t) = \mathrm{Cov}_{\nu_t}(Y_t, \lambda_t), \quad D_Y(t) = \mathbb{E}_{\nu_t}\dot Y_t.$$

If $Y_t(x) = Y(z_t(x), x)$ for a C^1 local function Y with no explicit time dependence, then $\dot Y_t(x) = \nabla_z Y(z_t(x), x)^\top v_t(x)$. Under raw flow, $v_t = K_t \Gamma_t$. For every subinterval $[a,b]$, $J_Y(b) - J_Y(a) = \int_a^b (R_Y + D_Y)\, dt$.

Proof. For fixed x, the prompt b_x and all tokens in its ancestry are fixed while training time changes. Thus

$$d_t(x) = \beta(b_x) \prod_{j<h(x)} \pi_t(a_j \mid x_j), \qquad \partial_t \log d_t(x) = \sum_{j<h(x)} \partial_t \log \pi_t(a_j \mid x_j).$$

This proves the formula for λ_t; at a root the product is empty and $\lambda_t = 0$. Also $\dot d_t(x) = d_t(x)\lambda_t(x)$. Differentiating the finite normalizer gives

$$\frac{\dot Z_t^{\mathrm{obs}}}{Z_t^{\mathrm{obs}}} = \frac{\sum_x m^{\mathrm{obs}}(x) d_t(x) \lambda_t(x)}{Z_t^{\mathrm{obs}}} = \mathbb{E}_{\nu_t}\lambda_t.$$

The quotient rule now gives the asserted formula for $\dot{\nu}_t$ at each eligible site. For a site with zero fixed mask, both ν_t and $\dot{\nu}_t$ are zero, so the same expression holds with that site omitted.

The product rule and the definition of covariance yield

$$\begin{aligned}\dot{J}_Y &= \sum_x \dot{\nu}_t(x) Y_t(x) + \sum_x \nu_t(x) \dot{Y}_t(x) \\ &= \mathbb{E}_{\nu_t}(Y_t \lambda_t) - (\mathbb{E}_{\nu_t} Y_t)(\mathbb{E}_{\nu_t} \lambda_t) + \mathbb{E}_{\nu_t} \dot{Y}_t.\end{aligned}$$

For a local readout, the chain rule gives $\dot{Y}_t = \nabla_z Y^\top \dot{z}_t$; the raw specialization follows from Section C.2. All functions are continuous and piecewise C^1 on a finite partition of the compact interval. Integrating the derivative on each piece and telescoping the boundary values proves the final equality. □

The two roles of the same parameter change. Define $\mathcal{F}(\theta_g, \theta_e) = \sum_x \nu_{\theta_g}(x) Y_{\theta_e}(x)$. Applying the multivariable chain rule to $\mathcal{F}(\theta_t, \theta_t)$ gives its generator derivative $\sum_x \dot{\nu}_t Y_t = R_Y$ and evaluator derivative $\sum_x \nu_t \dot{Y}_t = D_Y$. Both are contributions of the same parameter change to one derivative. A readout depending on the full parameter vector, rather than only the logits at x, uses $\dot{Y}_t(x) = \nabla_\theta Y_\theta(x)^\top \dot{\theta}_t$ instead. This distinction will matter for continuation values.

If prompt weights or observation masks vary explicitly, the formula for λ_t alone is insufficient. On positive coordinates the quotient-rule calculation instead centers $\partial_t \log[m_t^{\mathrm{obs}}(x) d_t(x)]$; $\partial_t \log d_t$ includes $\partial_t \log \beta_t(b_x)$ for a changing prompt law. Zero weights or changes of support require separate boundary accounting. The fixed-weight assumptions are therefore substantive.

C.6 Finite-window exposure and regenerated source weights

Connection to Section 3.3, Equation (7). Use a superscript tr when training and observation employ different fixed prompt laws or samplers. In particular, τ_{tr} can differ from the observation temperature. For a fixed training mask $m^{\mathrm{tr}}(s) \geq 0$, let

$$\mu_t(s) = \frac{m^{\mathrm{tr}}(s) d_t^{\mathrm{tr}}(s)}{Z_t^{\mathrm{tr}}}, \qquad Z_t^{\mathrm{tr}} = \sum_s m^{\mathrm{tr}}(s) d_t^{\mathrm{tr}}(s) > 0.$$

The following proposition assumes one fixed eligible support throughout the interval, as guaranteed by Assumption C.1. Merely sharing some positive-support states at the endpoints is not sufficient for the unrestricted normalization formula.

Proposition C.12 (Source reweighting across a training window)**.** *For $a \leq b$, let $\Lambda_{a:b}(s) = \int_a^b \lambda_t^{\mathrm{tr}}(s)\, dt$, where λ_t^{tr} is the ancestral log reach-probability velocity under the training sampler. On every allowed prefix,*

$$\frac{d_b^{\mathrm{tr}}(s)}{d_a^{\mathrm{tr}}(s)} = e^{\Lambda_{a:b}(s)}, \qquad \Lambda_{a:c}(s) = \Lambda_{a:b}(s) + \Lambda_{b:c}(s) \quad (a \leq b \leq c). \tag{16}$$

For each eligible source prefix,

$$\mu_b(s) = \frac{\mu_a(s) e^{\Lambda_{a:b}(s)}}{\mathbb{E}_{v \sim \mu_a} e^{\Lambda_{a:b}(v)}}, \qquad \Gamma_b(s) = \mu_b(s) B_b(s). \tag{17}$$

A zero fixed mask stays zero.

Proof. By Proposition C.11 with the training sampler, $\partial_t \log d_t^{\mathrm{tr}}(s) = \lambda_t^{\mathrm{tr}}(s)$ on each smooth piece. Integrating and telescoping gives $\log d_b^{\mathrm{tr}}(s) - \log d_a^{\mathrm{tr}}(s) = \Lambda_{a:b}(s)$. Exponentiating proves the ratio, and additivity of the integral proves the second identity.

Because the eligible set is the same at both endpoints, substitution into the entire source normalizer is valid:

$$\begin{aligned}Z_b^{\mathrm{tr}} &= \sum_s m^{\mathrm{tr}}(s) d_a^{\mathrm{tr}}(s) e^{\Lambda_{a:b}(s)} \\ &= Z_a^{\mathrm{tr}} \sum_s \mu_a(s) e^{\Lambda_{a:b}(s)}.\end{aligned}$$

The expectation is strictly positive and finite, since it is a finite sum of positive terms on the nonempty eligible set. Dividing $m^{\mathrm{tr}}(s)d_b^{\mathrm{tr}}(s)$ by this expression for Z_b^{tr} proves Equation (17). The expression for Γ_b defines the regenerated request population with the current numerical field B_b. If $m^{\mathrm{tr}}(s) = 0$, both endpoint weights are zero directly from their definition. □

An eligible prefix can be absent from a finite recorded batch while having positive population weight at every checkpoint. The reference in Equation (17) is the population μ_a, which includes such prefixes. At a root, $\Lambda = 0$, although normalization by the total expected token count can change its source weight.

C.6.1 Finite-batch extraction and native checkpoints

Proposition C.13 (Transport of a normalized batch extraction)**.** *Let $\mathfrak{B}$ be a finite batch space and $P_a^{\mathrm{tr}}, P_b^{\mathrm{tr}}$ two batch laws with identical positive support. For each batch $\mathcal{B}$, let $\widehat{\mu}_{\mathcal{B}}(s) \geq 0$ be an extraction rule with $\sum_s \widehat{\mu}_{\mathcal{B}}(s) = 1$, determined only by the batch text. Set $\mu_t(s) = \mathbb{E}_{P_t^{\mathrm{tr}}}\widehat{\mu}_{\mathcal{B}}(s)$ and $\Lambda_{a:b}(\mathcal{B}) = \log[P_b^{\mathrm{tr}}(\mathcal{B})/P_a^{\mathrm{tr}}(\mathcal{B})]$. Then*

$$\mu_b(s) = \mathbb{E}_{P_a^{\mathrm{tr}}}\big[\widehat{\mu}_{\mathcal{B}}(s)e^{\Lambda_{a:b}(\mathcal{B})}\big], \qquad \mathbb{E}_{P_a^{\mathrm{tr}}}e^{\Lambda_{a:b}(\mathcal{B})} = 1. \tag{18}$$

Proof. Expand the first expectation and cancel its positive probability factor:

$$\sum_{\mathcal{B}} P_a^{\mathrm{tr}}(\mathcal{B})\widehat{\mu}_{\mathcal{B}}(s)\frac{P_b^{\mathrm{tr}}(\mathcal{B})}{P_a^{\mathrm{tr}}(\mathcal{B})} = \sum_{\mathcal{B}} P_b^{\mathrm{tr}}(\mathcal{B})\widehat{\mu}_{\mathcal{B}}(s) = \mu_b(s).$$

Taking $\widehat{\mu}_{\mathcal{B}}(s)$ out of this calculation and replacing it by one gives $\sum_{\mathcal{B}} P_b^{\mathrm{tr}}(\mathcal{B}) = 1$. If the batch responses are independent, its probability is their product, so its log likelihood ratio is the sum of their complete-trajectory log ratios. Conditional independence given a fixed shared prompt schedule gives the same factorization conditional on that schedule. □

For example, if $N_s(\mathcal{B})$ is the eligible count at s, batch normalization gives $\widehat{\mu}_{\mathcal{B}}(s) = N_s(\mathcal{B})/\sum_v N_v(\mathcal{B})$ when the denominator is positive. Its expectation is generally not $\mathbb{E}N_s/\mathbb{E}\sum_v N_v$. Thus the random normalization must stay inside Equation (18); no additional ratio of expectations is introduced. A rule for batches with no eligible positions must be specified before defining a normalized extraction. If extraction also depends on the learner after fixing the text, transport uses the endpoint rule $\widehat{\mu}_{b,\mathcal{B}}$, not an unchanged rule from time a.

Native checkpoint identity. For discrete checkpoints $k_1 < k_2$, assume that every token in the chosen prefix has positive sampler probability at every intermediate checkpoint. Then

$$\begin{aligned}\Lambda_{k_1:k_2}(s) &:= \sum_{k=k_1}^{k_2-1}\sum_{j<h(s)} \log\frac{\pi_{k+1}^{\mathrm{tr}}(a_j \mid s_j)}{\pi_k^{\mathrm{tr}}(a_j \mid s_j)} \\ &= \sum_{j<h(s)} \log\frac{\pi_{k_2}^{\mathrm{tr}}(a_j \mid s_j)}{\pi_{k_1}^{\mathrm{tr}}(a_j \mid s_j)} = \log\frac{d_{k_2}^{\mathrm{tr}}(s)}{d_{k_1}^{\mathrm{tr}}(s)}.\end{aligned}$$

The first equality telescopes the finite sum over checkpoints for each token; the second multiplies the ancestral probability factors and cancels the fixed prompt law. The normalizer calculation in Proposition C.12 then applies unchanged on a fixed eligible support. The identity holds for any optimizer and any learning rate. If a token has zero probability at an intermediate checkpoint, the per-step logarithms above need not exist even when both endpoint probabilities are positive; the endpoint ratio can still be used directly where it is defined.

C.6.2 Changing support and nonsmooth readouts

Proposition C.14 (Explicit support accounting)**.** *Let P_a, P_b be probability laws on a common finite universe, let $\mathcal{C} = \operatorname{supp} P_a \cap \operatorname{supp} P_b$, and let F be a bounded real function. Define $e^{\Lambda(v)} =$*

$P_b(v)/P_a(v)$ *only for* $v \in \mathcal{C}$. *Then*

$$\mathbb{E}_{P_b} F - \mathbb{E}_{P_a} F = \sum_{v \in \mathcal{C}} P_a(v) F(v)(e^{\Lambda(v)} - 1) + \sum_{v \in \operatorname{supp} P_b \setminus \mathcal{C}} P_b(v) F(v) - \sum_{v \in \operatorname{supp} P_a \setminus \mathcal{C}} P_a(v) F(v).$$

The absolute value of the last two terms together is at most $\|F\|_\infty [P_b(\mathcal{C}^c) + P_a(\mathcal{C}^c)]$.

Proof. Split each expectation into its contribution on $\mathcal{C}$ and on its complement. On $\mathcal{C}$, replace $P_b(v)$ by $P_a(v)e^{\Lambda(v)}$ and subtract. This gives the identity. Applying the triangle inequality and $|F(v)| \le \|F\|_\infty$ to the remaining sums gives the bound. □

In particular, $\sum_{v \in \mathcal{C}} P_a(v) e^{\Lambda(v)} = P_b(\mathcal{C})$, which need not be one. Common-support ratios do not recover newly created mass or, by themselves, the full normalization of a changed source. The proposition applies to complete-trajectory laws, batch laws, or normalized source laws. Changed prompt sets and masks require the same care.

The local top-16 loss in Section E.1 can change its expression when the selected support changes. If a readout path has finitely many such changes and is piecewise C^1 with one-sided limits, integration over each smooth piece gives

$$J(b) - J(a) = \int_a^b \dot{J}(t)\, dt + \sum_{t_i \in (a,b)} [J(t_i^+) - J(t_i^-)],$$

with endpoints taken on their respective one-sided branches. The formula follows by telescoping the integrals and inserting the jumps; a continuous branch switch has zero jump. Smooth-flow integrals are asserted only under this regularity; the finite four-cell identity below remains valid whenever its four scores are finite. For an exogenously fixed source, the response and outcome laws still hold, while the source's own generating rule replaces the on-policy source-tilt formula.

C.7 Exact symmetric endpoint accounting and its path limit

Connection to Section 3.3, Equation (8). The four-cell estimator is exactly additive at finite endpoints; the difference between its individual components and the continuous routing and drift integrals is computed and bounded in Proposition C.16.

Proposition C.15 (Symmetric endpoint accounting)**.** *Let* ν_a, ν_b *be any two probability vectors on the union of two finite banks, and let* Y_a, Y_b *be finite readouts defined on that union. With* $M_{ij} = \sum_x \nu_i(x) Y_j(x)$, *define*

$$\widehat{R}_{a:b} = \tfrac{1}{2}[(M_{ba} - M_{aa}) + (M_{bb} - M_{ab})], \qquad \widehat{D}_{a:b} = \tfrac{1}{2}[(M_{ab} - M_{aa}) + (M_{bb} - M_{ba})]. \quad (19)$$

Using $\langle \nu, Y \rangle := \sum_x \nu(x) Y(x)$ *for the distribution–readout pairing,*

$$\widehat{R}_{a:b} = \left\langle \nu_b - \nu_a, \frac{Y_a + Y_b}{2} \right\rangle, \qquad \widehat{D}_{a:b} = \left\langle \frac{\nu_a + \nu_b}{2}, Y_b - Y_a \right\rangle,$$

and $\widehat{R}_{a:b} + \widehat{D}_{a:b} = M_{bb} - M_{aa}$. *This holds without common support or a parameter-path assumption.*

Proof. Set $\Delta\nu = \nu_b - \nu_a$ and $\Delta Y = Y_b - Y_a$. Bilinearity gives

$$M_{bb} - M_{aa} = \langle \Delta\nu, Y_a \rangle + \langle \nu_a, \Delta Y \rangle + \langle \Delta\nu, \Delta Y \rangle.$$

Expanding the four terms in Equation (19) yields

$$\widehat{R}_{a:b} = \langle \Delta\nu, Y_a \rangle + \tfrac{1}{2}\langle \Delta\nu, \Delta Y \rangle, \qquad \widehat{D}_{a:b} = \langle \nu_a, \Delta Y \rangle + \tfrac{1}{2}\langle \Delta\nu, \Delta Y \rangle.$$

These are the stated pairings and their sum is the first display. Equivalently, replacing the bank first assigns the interaction to drift, while replacing the evaluator first assigns it to routing; their average splits the interaction equally. No division by any source probability appears. □

Proposition C.16 (Difference from the path-integrated channels). *Suppose ν_t, Y_t are C^1 on $[a, b]$, and write $R_{a:b} = \int_a^b \langle \dot\nu_t, Y_t\rangle\, dt$ and $D_{a:b} = \int_a^b \langle \nu_t, \dot Y_t\rangle\, dt$. Then*

$$\widehat R_{a:b} - R_{a:b} = \int_a^b \left\langle \dot\nu_t, \frac{Y_a + Y_b}{2} - Y_t \right\rangle dt = -(\widehat D_{a:b} - D_{a:b}).$$

If $L_\nu = \sup_t \|\dot\nu_t\|_1$ and $L_Y = \sup_t \|\dot Y_t\|_\infty$, each channel error is at most $L_\nu L_Y (b-a)^2/2$. If both functions are C^2, define $M_\nu = \sup_t \|\ddot\nu_t\|_1$ and $M_Y = \sup_t \|\ddot Y_t\|_\infty$; the sharper bound is

$$|\widehat R_{a:b} - R_{a:b}| = |\widehat D_{a:b} - D_{a:b}| \le \frac{(b-a)^3}{12}(L_\nu M_Y + M_\nu L_Y).$$

Proof. Substitute $\nu_b - \nu_a = \int_a^b \dot\nu_t\, dt$ in the pairing for $\widehat R$ and subtract $R_{a:b}$. Since the sums of the two channels both equal $J_Y(b) - J_Y(a)$, their errors have opposite signs. The fundamental theorem of calculus gives

$$\left\| \frac{Y_a + Y_b}{2} - Y_t \right\|_\infty \le \tfrac12 L_Y[(t-a) + (b-t)] = \tfrac12 L_Y(b-a).$$

Applying $|\langle v, y\rangle| \le \|v\|_1 \|y\|_\infty$ and integrating proves the C^1 bound.

For the sharper result, insert $\frac12(Y_a+Y_b) - Y_t = \frac12(\int_t^b \dot Y_r\, dr - \int_a^t \dot Y_r\, dr)$ into the exact error. Reordering these finite-dimensional integrals gives

$$\widehat R - R = \tfrac12 \int_{a\le r<t\le b} \big[\langle \dot\nu_r, \dot Y_t\rangle - \langle \dot\nu_t, \dot Y_r\rangle\big]\, dr\, dt.$$

Add and subtract $\langle \dot\nu_r, \dot Y_r\rangle$ in the integrand. Its absolute value is bounded by $L_\nu \|\dot Y_t - \dot Y_r\|_\infty + \|\dot\nu_t - \dot\nu_r\|_1 L_Y \le (L_\nu M_Y + M_\nu L_Y)(t-r)$. Finally $\int_a^b \int_a^t (t-r)\, dr\, dt = (b-a)^3/6$, which proves the coefficient $1/12$. □

For a partition of a fixed interval with largest width h, the finite totals telescope exactly. Summing the C^1 bound gives a channel error at most $L_\nu L_Y (b-a)h/2$, tending to zero as $h \to 0$; the C^2 bound gives order h^2 instead.

Scope. The path limit is a population statement; finite-sample differences between empirical banks, checkpoint substitutions, prompt-composition changes, and support-switch jumps lie outside it, and Sections E.1 and E.6 give the corresponding experimental controls. At a fixed finite partition the allocation is a symmetric endpoint convention; a causal reading of the two channels would require a separate mediation analysis.

C.8 Continuation values and completed outcomes

Connection to Section 3.3, Equation (9). A terminal score is a fixed function of a completed response, whereas its expectation from a prefix changes with the remaining policy. We make that distinction explicit, including responses that terminate before the chosen intermediate depth.

Fix a bounded, parameter-independent function $Y : \mathcal{T} \to \mathbb{R}$. Write $J_Y(t) = \mathbb{E}_{\Pi_t} Y$ for the terminal expectation; in this subsection, J_Y denotes this quantity rather than the prefix-averaged diagnostic of Section C.5. For a nonterminal x, define

$$V_t^Y(x) = \mathbb{E}_t[Y \mid x \text{ is reached}], \qquad Q_t^Y(x,a) = V_t^Y(x \oplus a), \qquad \mathcal{A}_t^Y(x,a) = Q_t^Y(x,a) - V_t^Y(x).$$

For terminal x, set $V_t^Y(x) = Y(x)$. These are finite conditional expectations under the student's current sampler with its remaining generation budget. Positivity on the fixed allowed tree makes them well-defined; a receiver outside the training source is still allowed. Conditioning on the next action gives $V_t^Y(x) = \sum_a \pi_t(a \mid x) Q_t^Y(x,a)$ and $\sum_a \pi_t(a \mid x)\mathcal{A}_t^Y(x,a) = 0$.

Proposition C.17 (Recombining exposure and continuation change)**.** *For an integer $0 \le h \le H_{\max}$, let $\mathcal{C}_h$ contain every allowed node at depth h and every terminal node of depth less than h. Then $\sum_{x\in\mathcal{C}_h} d_t(x) = 1$, and*

$$\begin{aligned} J_Y(t) &= \sum_{x\in\mathcal{C}_h} d_t(x) V_t^Y(x), \\ \dot{J}_Y(t) &= \sum_{x\in\mathcal{C}_h} \dot{d}_t(x) V_t^Y(x) + \sum_{x\in\mathcal{C}_h} d_t(x) \dot{V}_t^Y(x). \end{aligned} \tag{20}$$

The first derivative term is the sum of the outcome contributions from decisions of depth less than h; the second is the sum from decisions of depth at least h. In particular, along any path satisfying the regularity conditions,

$$\dot{J}_Y(t) = \sum_{x\in\mathcal{S}} d_t(x) \mathbb{E}_{a\sim\pi_t(\cdot|x)}[\zeta_t(x,a)\mathcal{A}_t^Y(x,a)]. \tag{21}$$

Under raw flow this equals

$$\frac{1}{\tau}\sum_{x\in\mathcal{S}}\sum_{s} d_t(x)[\pi_t \odot \mathcal{A}_t^Y](x)^\top K_t(x,s)\Gamma_t(s).$$

Integrating from $t_\star$ to T yields Equation (9)*. If training and evaluation samplers differ, d_t, π_t, τ in this expression are the evaluation quantities, while Γ_t uses the training source and field.*

Proof. Finite conditional values and the cut. Every trajectory ends by depth $H_{\max}$. It either reaches depth h or terminates earlier, and therefore meets exactly one element of $\mathcal{C}_h$. The corresponding reach events are disjoint and exhaustive. Their probabilities sum to one, and conditioning the terminal score on this partition gives the first equality in Equation (20). At a nonterminal node, $V_t^Y(x)$ is a finite sum of terminal labels multiplied by products of suffix transition probabilities; at a terminal node it is the fixed label. Thus these values are continuous and piecewise C^1. The product rule proves the second equality.

Differentiating a continuation. For each nonterminal x, set

$$\mathfrak{h}_t(x) = \sum_a \pi_t(a \mid x)\zeta_t(x,a)\mathcal{A}_t^Y(x,a).$$

Differentiate the finite recursion $V_t^Y(x) = \sum_a \pi_t(a \mid x) V_t^Y(x \oplus a)$. By Lemma C.2, $\dot{\pi}_t = \pi_t\zeta_t$ and $\sum_a \pi_t\zeta_t = 0$, so

$$\dot{V}_t^Y(x) = \mathfrak{h}_t(x) + \sum_a \pi_t(a \mid x)\dot{V}_t^Y(x \oplus a). \tag{*}$$

At a terminal child the last derivative is zero. For a descendant y of x, let $d_t(y \mid x)$ be the product of transition probabilities from x to y, with $d_t(x \mid x) = 1$. Induction on the remaining depth proves

$$\dot{V}_t^Y(x) = \sum_{y\in\mathcal{S}\,:\,y\succeq x} d_t(y \mid x)\mathfrak{h}_t(y). \tag{**}$$

Indeed, if all children are terminal, (*) is precisely (**). Otherwise insert (**) for each nonterminal child into (*). Its descendant subtrees are disjoint, and multiplying a child's descendant reach probability by $\pi_t(a \mid x)$ gives $d_t(y \mid x)$. This proves the induction step and counts every remaining decision exactly once.

Decisions after the cut. For a nonterminal y of depth at least h, the ancestry has a unique cut ancestor $x \in \mathcal{C}_h$. Also $d_t(x)d_t(y \mid x) = d_t(y)$. Multiplying (**) by $d_t(x)$ and summing over the cut therefore gives

$$\sum_{x\in\mathcal{C}_h} d_t(x)\dot{V}_t^Y(x) = \sum_{y\in\mathcal{S}\,:\,h(y)\ge h} d_t(y)\mathfrak{h}_t(y).$$

Terminal cut nodes have no descendants with decisions and contribute zero.

Decisions before the cut. The reach derivative at a cut node is $\dot{d}_t(c) = d_t(c)\sum_{j<h(c)}\zeta_t(c_j, a_j)$. Reorder the finite sum $\sum_{c\in\mathcal{C}_h}\dot{d}_t(c)V_t^Y(c)$ by each ancestral decision (x, a) with $h(x) < h$. The coefficient of $\zeta_t(x, a)$ is

$$\sum_{c\in\mathcal{C}_h\,:\,c\succeq x\oplus a} d_t(c)V_t^Y(c) = d_t(x)\pi_t(a \mid x)Q_t^Y(x, a).$$

To verify this equality, condition on reaching x and selecting a. The subsequent cut nodes form a disjoint, exhaustive partition of that conditional event, and the conditional expectation of their values is $Q_t^Y(x, a)$. This includes a terminal child $x \oplus a$, which is itself in the cut. Summing over a and subtracting $V_t^Y(x)\sum_a \pi_t(a \mid x)\zeta_t(x, a) = 0$ gives

$$\sum_{c\in\mathcal{C}_h}\dot{d}_t(c)V_t^Y(c) = \sum_{x\in\mathcal{S}\,:\,h(x)<h} d_t(x)\mathfrak{h}_t(x).$$

Combining the two depth ranges proves Equation (21), including the boundary cases $h = 0$ and $h = H_{\max}$.

Kernel form and integration. Let $a_t^Y(x) = [\pi_t \odot \mathcal{A}_t^Y](x)$, extended by zero outside the allowed tokens. Its coordinates sum to zero. Substituting $\zeta_t = \tau^{-1}(e_a - \pi_t)^\top v_t$ gives

$$\mathfrak{h}_t(x) = \tau^{-1}\left[a_t^Y(x)^\top v_t(x) - \Big(\sum_a \pi_t(a \mid x)\mathcal{A}_t^Y(x, a)\Big)\pi_t^\top v_t(x)\right] = \tau^{-1}a_t^Y(x)^\top v_t(x).$$

Under raw flow, $v_t(x) = \sum_s K_t(x, s)\Gamma_t(s)$, proving the bilinear formula. For an actual optimizer path, the same calculation uses $v_t(x) = \Phi_t(x)\dot{\theta}_t$. Integrating the continuous, piecewise differentiable J_Y proves the accumulated outcome formula. □

The suffix derivative in (**) already includes changes to every downstream prediction. Adding a second derivative of Q_t^Y to the full decision sum in Equation (21) would double-count that motion. Conversely, at a fixed cut prefix, $\dot{V}_t^Y(x)$ cannot generally be replaced by a derivative with respect to the logits at x alone. Local entropy and terminal correctness are different observables even though their changes follow the same underlying parameter path.

Reordering the finite source-receiver sum gives the source-resolved form

$$\dot{J}_Y(t) = \sum_s \mu_t(s)\varphi_{Y,t}(s), \qquad \varphi_{Y,t}(s) = \frac{1}{\tau}\sum_x d_t(x)[\pi_t \odot \mathcal{A}_t^Y](x)^\top K_t(x, s)B_t(s).$$

A source contributes through every coupled receiver, not just through its own continuation value. The experiments in Section E therefore combine prefix dynamics with held-out outcomes.

C.9 Entropy sensitivity and ancestral exposure

Connection to Section 3.3, the entropy-versus-fitting argument. In this subsection $H_t(x) = -\sum_a p_t(a \mid x)\log p_t(a \mid x)$ is full-vocabulary model entropy. Its policy is $p_t = \mathrm{softmax}(z_t)$ at temperature one; the sampling policy in the routing term remains π_t at temperature τ. Fixed-bank evidence is in Section E.4.1.

Proposition C.18 (Entropy sensitivity)**.** *Under Equation* (15)*,*

$$\begin{aligned} R_H(t) &= \mathrm{Cov}_{\nu_t}\left(H_t(x), \sum_{j<h(x)}\zeta_t(x_j, a_j)\right), \\ D_H(t) &= \mathbb{E}_{\nu_t}[\chi_{H,t}(x)^\top v_t(x)], \\ \chi_{H,t}(x) &:= \nabla_z H_t(x) = -p_t(\cdot \mid x) \odot [\log p_t(\cdot \mid x) + H_t(x)\mathbf{1}]. \end{aligned} \tag{22}$$

At a fixed receiver,

$$\dot{H}_t(x) = -\,\mathrm{Cov}_{a\sim p_t(\cdot|x)}(\log p_t(a \mid x), v_t(x, a)), \qquad |\dot{H}_t(x)| \le \sqrt{\mathrm{Var}_{p_t}(\log p_t)\,\mathrm{Var}_{p_t}(v_t)}. \tag{23}$$

The parameter-space entropy gradient is $\Phi_t(x)^\top\chi_{H,t}(x)$. These identities hold with the actual logit velocity; the raw specialization is $v_t = K_t\Gamma_t$.

Proof. Fix x and suppress it in the notation. Softmax differentiation gives $\dot{p}_a = p_a(v_a - \bar{v}_p)$, where $\bar{v}_p = \sum_a p_a v_a$. Since $\sum_a \dot{p}_a = 0$,

$$\begin{aligned}\dot{H} &= -\sum_a \dot{p}_a(\log p_a + 1) = -\sum_a p_a(v_a - \bar{v}_p)\log p_a \\ &= -\operatorname{Cov}_p(\log p, v) = -\sum_a p_a(\log p_a + H)v_a.\end{aligned}$$

The last equality uses $\sum_a p_a \log p_a = -H$. Since the directional identity holds for every logit direction v, its coefficient vector is $\chi_H = \nabla_z H$. Weighted Cauchy–Schwarz applied to the centered vectors $\log p - \mathbb{E}_p \log p$ and $v - \mathbb{E}_p v$ proves the variance bound. The chain rule gives the parameter gradient $\Phi^\top \chi_H$. Finally substitute the local derivative into Proposition C.11; its ancestral score gives the stated R_H and its fixed-prefix derivative gives D_H. □

Proposition C.19 (Ancestral exposure without first-order entropy increase)**.** *Consider a fixed path of $h \geq 1$ binary continue/stop decisions. At every gate, let the conditional continuation probability be a differentiable $c_t \in (0,1)$, and hold the suffix executor fixed. Conditional on the path's starting node, its end is reached with probability c_t^h. With $H_{\mathrm{bin}}(c) = -c\log c - (1-c)\log(1-c)$,*

$$\partial_t \log \Pr_t(\textit{reach depth } h) = h\dot{c}_t/c_t, \qquad \partial_t H_{\mathrm{bin}}(c_t) = \dot{c}_t \log[(1-c_t)/c_t]. \tag{24}$$

Proof. The probability of reaching the end is the product of the h conditional gate probabilities, regardless of any independence interpretation. Its logarithm is $h\log c_t$, giving the first derivative. Direct differentiation yields $H'_{\mathrm{bin}}(c) = -\log c - 1 + \log(1-c) + 1 = \log[(1-c)/c]$; the chain rule gives the second. □

For example, at $c_t = 1/2$ with $\dot{c}_t > 0$, the reach derivative is positive and each gate's entropy derivative is zero. Thus a pre-existing uncertain suffix can receive more exposure in the rollout distribution *without* its predictions becoming flatter.

Checking propensity and teacher-relative discrepancy. For the fixed WAIT family $\mathcal{W}$, let $w_t = p_t(\mathcal{W} \mid x) \in (0,1)$ at a fixed prefix, $b_t = \log[w_t/(1-w_t)]$, and $q_W = q(\mathcal{W} \mid x)$ be fixed. Define

$$c_W(t) = D_{\mathrm{KL}}(\mathrm{Ber}(q_W) \| \mathrm{Ber}(w_t)).$$

With the convention $0\log 0 = 0$ at $q_W \in \{0,1\}$,

$$\partial_t c_W = (w_t - q_W)\partial_t b_t.$$

Indeed, writing out the two Bernoulli terms gives $\partial_w c_W = -q_W/w + (1-q_W)/(1-w) = (w - q_W)/[w(1-w)]$. Also $\partial_t b_t = \dot{w}_t/[w_t(1-w_t)]$. Multiplication proves the identity, including the endpoint teacher probabilities by the same formula. A decrease in propensity is corrective exactly when its product with the current signed teacher discrepancy is negative; an initially above-teacher class need not remain above the teacher throughout training.

D CONVERSION: PROTOCOLS AND ADDITIONAL MEASUREMENTS

This appendix supports Section 3.2: pairings and banks first, semantic request locations next, and then collective estimates and geometric adaptation. The full conversion and frozen-field derivations are in Sections C.3 and C.4; they are not repeated here. The observation populations and uncertainty conventions are specified below. All intervals in this appendix resample rollouts at fixed checkpoints and so exclude independent training-seed variation. Each OPD step samples 256 rollouts of up to 8,192 tokens for 200 steps and scores every prefix with the teacher, sequentially, so one run can reach roughly 0.4B generated tokens; we therefore used our computational budget for breadth across three model families, two scales, and four preparation doses instead of more repetitions of healthy runs.

D.1 PAIRINGS, EVALUATION, AND FIXED PROBE BANKS

The 1.5B student is DeepSeek-R1-Distill-Qwen-1.5B; its teachers are JustRL-1.5B and DeepSeek-R1-Distill-Qwen-7B. The 7B student is DeepSeek-R1-Distill-Qwen-7B; its teachers are AceReason-Nemotron-7B and DeepSeek-R1-Distill-Qwen-14B. The AIME comparison uses 30 AIME 2024 problems, temperature .6, and 16 samples per problem at 1.5B or four at 7B. The stalled 7B endpoint is step 150; the other endpoints are step 200. Its small headline decline is sensitive to truncation, so the distinction concerns absence of the healthy run's gain.

Per scale, eleven initial rollouts supply the anchor bank. They appeared in the first batch once and are not subsequently trained on again. The initial support contains the student's 16 highest-probability tokens at each selected site. Sentence annotations divide sites into CONTROL (planning, checking, reconsideration and related management) and TECHNICAL (calculation and derivation). Within each class and teacher, 80 sites are sampled from the largest 5% of initial request norms. These teacher-dependent strata are fixed across checkpoints. Additional 1.5B bulk probes sample 200 sites per class; no 7B bulk kernel measurement is used here.

Each local kernel uses gradients of its 16 support logits with respect to all parameters. The fp32 diagonal and directional contractions are computed separately; off-diagonal Gram entries use bf16-stored gradients before symmetrization. Thus "full-parameter" refers to parameter coverage, and Gram entries mix fp32 and bf16 precision. The top-stratum Rayleigh quotients agree with the exact directional calculations to relative error below 1.3×10^{-4}. The eigendecomposition is inexpensive once the gradients are available, but obtaining the self-kernels is a sampled research audit; Ξ is the cheap routine monitor.

D.2 SEMANTIC CONTEXT OF THE REQUEST POPULATIONS

There are 36,416 annotated positions at 1.5B and 38,168 at 7B; each teacher's strongest 1% contains 364 and 381 selected positions, respectively. For category c, enrichment is

$$E_q(c) = \frac{\Pr(c \mid \text{strongest } 1\% \text{ under teacher } q)}{\Pr(c)}.$$

Bootstrap intervals resample rollouts, as specified in the annotation protocol below. They are marginal per teacher; the enrichment difference between teachers has no interval. Figure 2 selects twelve rows, Figure 7 shows all categories, and Table 1 records the main figure's point estimates.

Categories, sentence positions, and token kinds are separate partitions of the same positions, so shares combine only within a partition. Unlike this strongest-1% semantic population, the kernel measurements use the teacher-specific strongest-5% strata.

Annotation protocol and category definitions. The typed-tail analysis of Figure 7 assigns every labeled body position exactly one semantic category and one token kind. The rubric (v1.2) gives each sentence a list of fine act labels and, independently, a coarse class (CONTROL, TECHNICAL, BOTH, OTHER); 11–13$\%$ of labeled sentences carry more than one act, almost always within one class (computation with case verdicts or setup, derivation with setup, planning with assessment). We therefore coarsen within class: *derivation* absorbs setup, computation and case verdicts; *planning* absorbs assessment; check initiation, check conclusion, doubt flags, reframing, recap, repair and commitment remain separate, as does recall. A sentence that still lists acts from both classes is assigned by its annotated coarse class (a CONTROL sentence keeps its control act, a TECHNICAL sentence its technical act); sentences annotated BOTH form the *mixed* category and OTHER sentences are dropped. The residual same-class multi-category sentences are resolved to the rarer category and amount to 1.1% (7B) and 1.4% (1.5B) of positions; positions whose act list contradicts the annotated class (0.1%) go to the mixed category. Token kind partitions tokens into wait-family, digit-bearing and other, and sentence position into initial and interior. Over-representation is the share of a category among the teacher's strongest 1% of positions divided by its share of all positions; intervals resample the 11 labeled rollouts. The one-label-per-sentence coarsening leaves every headline row of Figure 7 within 0.1 of its value under the multi-label count..

Category-wise comparisons. At 1.5B, planning is enriched $2.06\times$ under the sibling versus $0.81\times$ under the larger teacher, and sentence-initial positions $2.91\times$ versus $1.46\times$. At 7B, the sibling stands

out at reframing, checking, and doubt positions; wait-family tokens are enriched $4.50\times$ versus $1.23\times$. The separation is category-dependent: both 7B teachers emphasize planning ($1.39\times$ and $1.58\times$), and derivation enrichment is similar ($0.84\times$ and $0.83\times$). Digit-bearing positions are relatively more prominent under the larger teacher at both scales, although below their corpus prevalence. These enrichment ratios locate strong requests, independent of whether those requests promote or suppress the corresponding behavior.

	1.5B student		7B student	
Position category	Sibling	Larger	Sibling	Larger
Reframe	1.63	0.70	2.61	0.65
Check-conclude	1.52	0.91	1.86	0.62
Doubt flag	1.15	1.84	2.95	0.68
Check-initiate	1.37	1.20	2.64	1.32
Planning	2.06	0.81	1.39	1.58
Recap	1.48	1.11	0.00	2.31
Derivation	0.78	0.93	0.84	0.83
Recall	2.04	1.89	0.50	2.18
Mixed	0.98	1.57	1.25	1.25
Wait-family token	1.74	0.43	4.50	1.23
Sentence-initial token	2.91	1.46	1.07	0.46
Digit-bearing token	0.65	0.85	0.40	0.79

Table 1: **Numerical enrichment comparisons for the main semantic figure.** Ratios to corpus prevalence. Marginal bootstrap intervals are plotted in Figure 7.

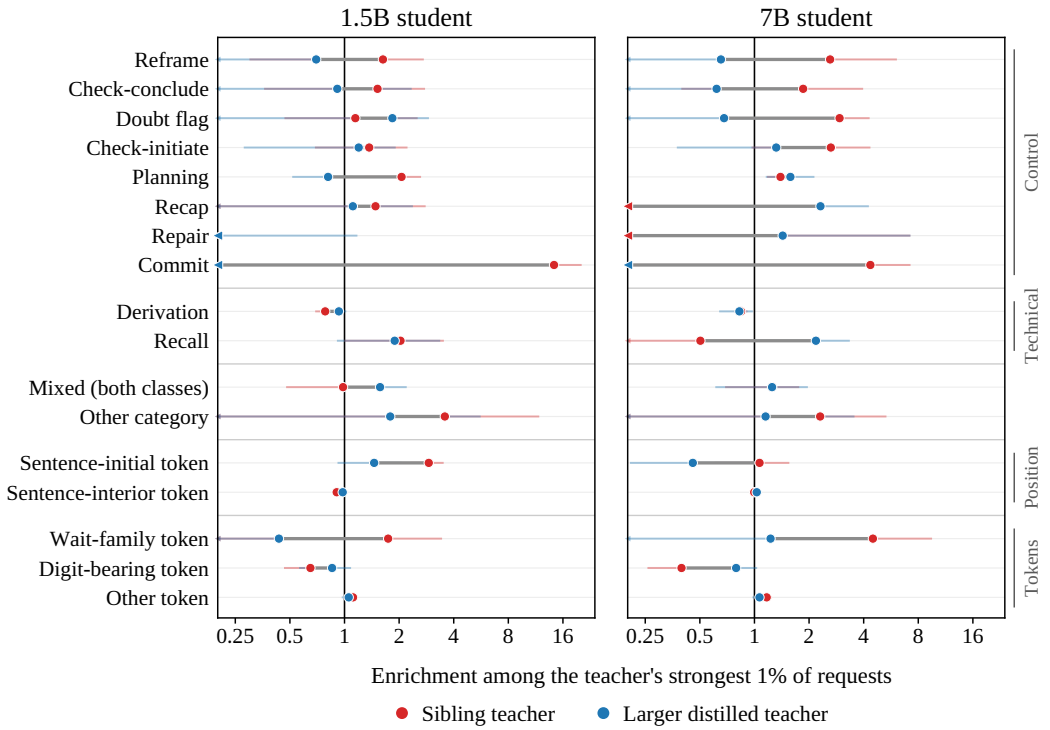


Figure 7: **Complete typed-position enrichment.** All categories and the separate sentence-position and token-kind partitions in the annotated position population. A ratio of one denotes corpus prevalence; faint arrowheads mark intervals extending beyond the plotted range, and solid left-pointing markers denote zero estimates. The main figure selects twelve rows from this plot.

D.3 Numerical audit field and class-resolved collective estimates

Arm	Stratum	P_0	$A_0/10^3$	C_0	$\Xi_0/10^3$	$\|u_0\|$
7B healthy	Control	0.192	119.8	0.081	9.70	43.1
7B healthy	Technical	0.151	106.3	0.055	5.85	29.7
7B stalled	Control	0.521	148.1	0.025	3.70	43.9
7B stalled	Technical	0.599	128.8	0.020	2.58	39.3
1.5B healthy	Control	6.024	39.6	0.041	1.62	98.9
1.5B healthy	Technical	2.569	54.5	0.064	3.49	94.7
1.5B stalled	Control	0.824	41.3	0.040	1.65	36.9
1.5B stalled	Technical	1.110	53.8	0.039	2.10	48.2

Table 2: **Initial collective diagnostics.** Each row uses one teacher-specific top stratum ($n = 80$). A and P use per-site quantities; C is a rounded, bias-corrected estimate, defined below. $\Xi = AC$ and $\|u\| = \sqrt{P\Xi}$ are approximate plug-in point estimates. Compare within scale and parameterization.

Arm	Stratum	Step	P_t/P_0	A_t/A_0	C_t/C_0	Ξ_t/Ξ_0	$A_t[B_0]/A_0[B_0]$
7B healthy	Control	200	0.005	0.983	0.012	0.012	1.414
7B healthy	Technical	200	0.007	2.159	0.036	0.079	1.232
7B stalled	Control	150	0.740	0.827	0.520	0.430	0.810
7B stalled	Technical	150	0.740	0.884	0.500	0.442	0.889
1.5B healthy	Control	200	0.009	1.216	0.341	0.415	2.829
1.5B healthy	Technical	200	0.015	3.628	0.031	0.113	5.399
1.5B stalled	Control	200	0.602	0.836	0.725	0.606	0.831
1.5B stalled	Technical	200	0.790	0.793	0.564	0.447	0.789

Table 3: **Remaining requests, live conversion, and frozen accessibility.** A_t, C_t, Ξ_t in the middle columns use the live field B_t; the final column uses the frozen field B_0. Live and frozen columns belong to separate identities. Coherence estimates are rounded, especially at healthy endpoints. Whether a small healthy residual reflects the same bottleneck as persistent stalled pressure remains open.

Gradient evaluation and uncertainty. Each pullback is obtained by differentiating $\sum_a B_a z_a$ at the exact prefix, with B detached. The context ends immediately before the scored token, so the selected logits predict that token rather than its successor. Live student scores are checked against the corresponding fixed-prefix scores before evaluating gradients.

For $n > 1$ sampled site gradients, write $\bar{g} = n^{-1} \sum_i g_i$ and $\widehat{D}_g = n^{-1} \sum_i \|g_i\|^2$. The finite-sample correction used for the coherence estimates is

$$\widehat{C}_{\mathrm{corr}} = \frac{\left(\|\bar{g}\|^2 - [\widehat{D}_g - \|\bar{g}\|^2]/(n-1) \right)_+}{\widehat{D}_g}, \qquad \widehat{D}_g > 0,$$

where $(r)_+ = \max(r, 0)$. Under independent sampling, the untruncated numerator is unbiased for the squared population-mean gradient norm (Proposition C.9); the positive-part operation and division do not yield an unbiased coherence estimator. Sampling without replacement requires the corresponding finite-population correction for an unbiased norm estimate. The tabulated values retain the stated clipped correction and rounding.

Products $\Xi = AC$ and $\|u\| = \sqrt{P\Xi}$ are approximate plug-in point estimates; confidence intervals for them cannot be recovered from the self-kernels alone. Comparisons use a fixed parameterization, output support, and numerical-field convention. These standardized diagnostics apply to the raw gradient; the complete native optimizer step lies outside their scope.

D.4 Geometric changes on the same initial requests

For each site, let $V_{t,3}(s)$ contain the orthonormal eigenvectors of the three largest eigenvalues of its symmetrized self-kernel. Main Figure 3b plots two paired quantities:

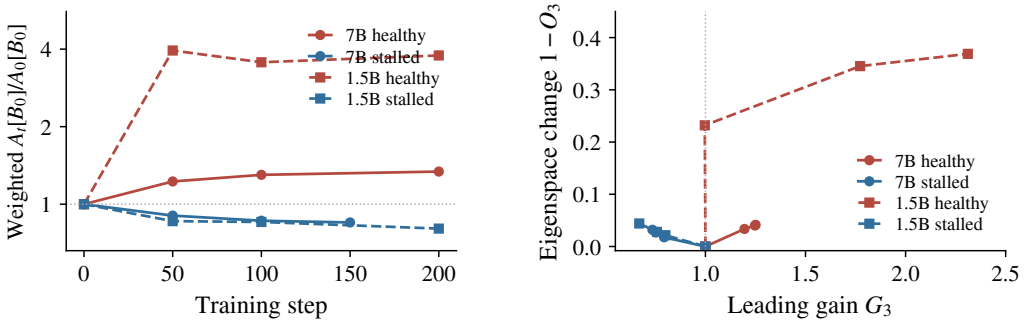
$$G_3(s,t) = \frac{\sum_{j=1}^{3} \lambda_{j,t}(s)}{\sum_{j=1}^{3} \lambda_{j,0}(s)}, \qquad O_3(s,t) = \frac{\|V_{0,3}(s)^\top V_{t,3}(s)\|_F^2}{3}.$$

O_3 is the mean squared cosine of the three principal angles: 1 means identical subspaces, 0 orthogonal. It is invariant to signs and within-subspace basis rotations, provided the selected subspaces are held fixed. A strict gap between the third and fourth eigenvalues makes the leading subspace unique; a tie at that boundary permits different selections, and a crossing can change the selected subspace. The main panel uses median G_3 and O_3 over 160 sites; the table below retains $1 - O_3$.

Overlap tracks subspace movement from the kernels alone; alignment with the teacher is measured by the frozen-request share $w_{1:3}(s,t) = \sum_{j\le 3} |v_{j,t}^\top b_0|^2$ and its paired change $\Delta w_{1:3}$. Figure 3c compares this change between the two fields in identical kernels.

Arm	Step	G_3	$1 - O_3$	$\Delta w_{1:3}$ (pp)	Median a_t/a_0	$A_t[B_0]/A_0[B_0]$
7B healthy	50	1.19	0.034	-0.23	1.04	1.23
7B healthy	200	1.25	0.041	-0.36	1.07	1.34
7B stalled	50	0.79	0.018	+1.12	0.93	0.90
7B stalled	150	0.74	0.032	+1.51	0.91	0.85
1.5B healthy	50	1.00	0.232	+5.49	1.19	3.95
1.5B healthy	200	2.31	0.369	+5.07	2.08	3.78
1.5B stalled	50	0.80	0.022	+1.06	0.91	0.86
1.5B stalled	200	0.67	0.044	+0.89	0.83	0.80

Table 4: **Pooled geometric summaries on 160 sites per arm.** All columns except the final one are site medians, with pairwise ratios/differences taken before aggregation. The final column is the ratio of request-energy-weighted population accessibilities. These estimands can differ considerably: a small number of high-energy sites affects the population statistic more.



(a) The accessibility factor on frozen requests.

(b) Leading-subspace change of the kernel alone.

Figure 8: **Kernel motion and its response to a fixed probe are distinct.** Left: exact population $A_t[B_0]/A_0[B_0]$ on the pooled balanced site sample. Right: per-site median leading strength and eigenspace change. Red/blue denote healthy/stalled; circles/squares denote 7B/1.5B.

D.5 Teacher-relative reorientation and spectral checks

To test whether a moving eigenspace turns preferentially toward the trained field, we use the other teacher's numerical B_0 at the same site and support as a control, and compare both projection changes in the identical checkpoint kernel. The paired advantage is

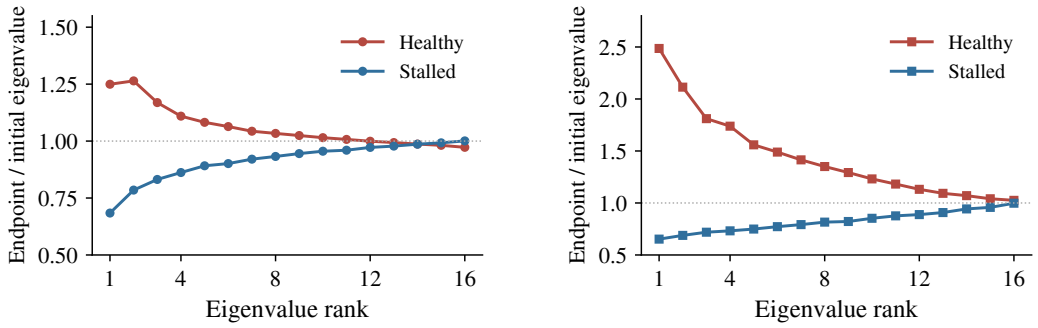
$$d_s(t) = \Delta w_{1:3}^{\text{trained}}(s,t) - \Delta w_{1:3}^{\text{other}}(s,t).$$

Figure 3c uses a matched step-50 checkpoint for all four arms. A positive value favors the trained field, while generic drift shared by both fields cancels in the comparison. The fields can be correlated, so the other field can also gain projection.

The crossed-field result supports preferential local adaptation under healthy 1.5B; broad amplification can also improve other fields. The measured growth is not preferentially concentrated in initial directions with larger request energy, so we do not attribute all of it to targeted learning of the most strongly requested directions.

Arm	Trained gain (pp)	Other gain (pp)	Paired advantage (pp)	95% interval
7B healthy	-0.23	+0.07	-0.16	$[-0.33, +0.27]$
7B stalled	+1.12	+1.25	+0.00	$[-0.53, +0.60]$
1.5B healthy	+5.49	+2.42	+1.79	$[+0.94, +3.13]$
1.5B stalled	+1.06	+1.00	-0.16	$[-0.29, +0.05]$

Table 5: **Matched step-50 field comparison.** Site medians on 160 pooled sites. The paired advantage is the median of per-site differences; it need not equal the difference between the two marginal medians. Intervals resample the eleven anchor rollouts 3,000 times with the trained runs fixed, retaining all sampled sites within each selected rollout and pairing the fields.



(a) 7B student.

(b) 1.5B student.

Figure 9: **Endpoint eigenvalue profiles.** Median paired ratios at each sorted rank, pooling 160 sites per arm. This figure compares spectral strengths; Figure 3b separates leading-subspace overlap from gain, and Figure 3c tests the teacher-relative projection change. Red/blue: healthy/stalled. Endpoints are step 200 except stalled 7B at 150.

Scope of the collective measurements. The crossed-field result leaves open where any reasoning circuit lies. Inferring a near-nullspace would need evidence beyond lower-rank request occupancy. The local kernels do not determine intermediate live coherence, a frozen-field collective-gain trajectory, or the between-class aggregate-gradient products needed for pooled Ξ. These quantities are not interpolated from self-kernel ladders. In particular, frozen accessibility and live coherence refer to different numerical fields and cannot be multiplied to estimate one collective gain.

E BROADCAST: ROUTING, CONDITIONAL LEARNING, AND TRAJECTORY AUDITS

This appendix supports Section 3.3. The routing/drift, source-renewal, endpoint, outcome, and entropy derivations are collected in Sections C.5 to C.9. Here we give the corresponding instruments, measurements, and empirical limitations.

E.1 MODELS, BANKS, AND READOUT DEFINITIONS

Three learners under one teacher. The teacher is JustRL-1.5B. The strong learner is DeepSeek-R1-Distill-Qwen-1.5B; the unprepared learner is Qwen2.5-1.5B-Instruct; the prepared learner is the latter after four passes of offline distillation over the half-corpus stream specified in Section G.2. Each is one training realization. The Qwen runs have 200 updates, 256 rollouts per update, temperature one, an 8,192-token generation cap, and the top-16 probability-weighted recipe below. DeepSeek's training cap is 4,096. The retrospective observation cap is 4,096 for all three. The preparation-dose and held-out evaluation protocols are in Section G.2.

Frozen trajectory banks. The grid uses 32 rollouts from each recorded batch $\ell \in \{1, 21, 41, \ldots, 161, 181, 196\}$. Batch ℓ is generated after $\ell - 1$ updates. The routing/drift figures use

update numbers, so batch 41 is plotted at update 40. Checkpoint 0 always denotes the untrained or prepared starting learner. Every evaluator is run on the same exact prompt and response tokens. Pooled loss/entropy means weight positions up to the observation cap; "front" restricts to the first 512 response tokens. The retrospective depth grid uses exact available generating checkpoints, and its depth means condition on positions observed in each bin. The main-text heatmaps instead use the separate full-batch measurements to pair student surprisal with teacher top-1 confidence on identical source histories; their conventions are specified below and in Section E.6.1.

Checkpoint substitutions and changed prompt composition. DeepSeek has the exact evaluators for the displayed 0, 20, …, 160 grid. The unprepared batch-61 bank uses checkpoint 50 instead of 60. Prepared batches 41 and 61 both use checkpoint 50 in the dense-window accounting. The resulting zero drift between those prepared banks is an artifact of the shared evaluator; the early negative drift is described as present *by checkpoint 50*. Open plot markers identify these substitutions. The sensitivity analysis uses coarser windows retaining only exactly matched generating checkpoints (Table 9). Banks 181 and 196 have a changed prompt-language composition and are excluded from the primary trajectories. Earlier batches can still differ in prompt composition and finite-sample content; the empirical routing component is precisely bank replacement and is interpreted as policy routing subject to those sampling controls.

Local training readout. For each evaluated prefix, $S_t(x)$ is the evaluating student's own top-16 token set and $\widetilde{p}_t$ renormalizes p_t over that set. The teacher is scored on every evaluator's selected tokens. The training-batch curve uses the stored generating-policy scores; the grid recomputes each evaluator's scores. Their populations and numerical values differ, and each is reported separately. This statistic is the reported local discrepancy underlying the recipe's rewards. It is restricted to the selected tokens and can be negative, and its ordinary derivative can differ from the recipe's detached-coefficient policy-gradient update. A low value can still accompany a sizable numerical request or raw gradient. We use the term local loss for this diagnostic in the main text. The smooth identity in Proposition C.11 applies between support changes; Section C.6.2 gives the corresponding boundary qualification. Finite endpoint accounting remains exact.

Entropy and teacher uncertainty. Student entropy is full-vocabulary $-\sum_a p_t(a \mid x) \log p_t(a \mid x)$ on the fixed receiver prefixes. Teacher entropy is the same quantity for q on those histories; their difference compares uncertainty levels, independent of which tokens carry the mass. These grid values are distinct from the full-batch heatmaps of sampled-token surprisal and teacher top-1 probability in Section E.6.1.

Checking instruments. The fixed family $\mathcal{W}$ consists of the 13 vocabulary tokens whose decoded form, after trimming whitespace and lowercasing, is "wait". A candidate sentence start is response position zero or a position whose preceding token ends with a period, question mark, exclamation mark, colon, or newline. Let $w_t(x) = p_t(\mathcal{W} \mid x)$ and $q_W(x) = q(\mathcal{W} \mid x)$. The pooled propensity instrument uses $\log w_t$ at candidate starts. The two-sided instrument uses $b_t = \log[w_t/(1 - w_t)]$; the teacher reference is $b_q = \log[q_W/(1-q_W)]$. Classes are frozen using checkpoint 0 for DeepSeek and checkpoint 20 for the Qwen learners: eligibility requires $\max(w_r, q_W) \geq 10^{-3}$ and $|b_r - b_q| \geq 0.5$; the sign defines initially above- or below-teacher contexts.

Realized WAIT-family tokens per thousand response tokens count actual emissions. This emission rate, the class density per thousand tokens, the class's normalized population, and the mean log-propensity are four separate quantities. We also measure the arithmetic mean w_t on candidate starts. Taking logs and averaging do not commute, so the mean log-propensity and the arithmetic mean can change with different signs, and the log-propensity decomposition does not sum to the count-rate change. Teacher-relative checking discrepancy is $c_W = D_{\mathrm{KL}}(\mathrm{Ber}(q_W) \,\|\, \mathrm{Ber}(w_t))$. At a fixed prefix, $\partial_t c_W = (w_t - q_W)\partial_t b_t$: whether lowering propensity is corrective depends on the current sign of $w_t - q_W$, which can differ from the class's initial side.

Source–kernel probes. A crossed response uses the kernel at checkpoint i and the complete numerical request population from a recorded batch ℓ, denoted $K_i\Gamma^{[\ell]}$. The bracketed index is a batch index: batch ℓ is generated by checkpoint $\ell-1$. Source positions are averaged with their original token weights, and their numerical coefficients are held fixed when differentiating the logits at checkpoint i.

The field includes the implemented normalization contributions, unlike the noncentered diagnostic field in Section D.3. Directional entropy, taken-token log-probability, and stopping responses are evaluated by central finite differences along the normalized gradient direction, then rescaled by its norm. Reported responses are first-order rates per unit learning rate.

Receiver banks contain exact, fixed response histories. The reach probes below use the 32-rollout banks from batches 41 and 61. The coherent-front probe uses the initial student's responses on 50 held-out MATH-500 prompts, four samples per prompt, at a 2,048-token cap; "front" means its first 512 tokens. Source-depth splits retain each source's full prefix context but restrict the differentiated positions to depths below 1,024 or at least 1,024. Each part is multiplied by its fraction of all source positions, so the two contributions sum to the full raw response.

Uncertainty and scope. The grid figures show single-run point estimates; token positions are not treated as independent replications. Token-weighted means and averages of per-rollout means are different estimands. The WAIT request audit has its own 95% intervals, and regression controls use rollout-cluster uncertainty; both reflect sampling within each run. Fixed-bank class summaries use the complete pooled trajectory bank, including its late batches, with the same receiver membership at every evaluator. In contrast, moving-bank claims stop at batch 161 because later batches change prompt-language composition. The held-out banks contain 200 rollouts on 50 MATH prompts with a 2,048-token cap; they probe transfer to those generated texts only. The checking readouts in this appendix measure propensities, emission rates, and teacher-relative discrepancy; they do not grade individual checks, which Section F.2 does for the DeepSeek run.

E.2 DeepSeek: conditional refinement and selectivity

Table 6 summarizes the cumulative loss and entropy accounting. In DeepSeek the local loss decreases mainly through fixed-prefix change. The main checking plot uses filled circles for over-use and open triangles for under-use. These classes are fixed by the reference checkpoint and teacher on each candidate sentence-start prefix, independent of the token subsequently emitted. Checking is adjusted in both directions in the front window: the above-teacher class has $\sum \widehat{R} = 0.036$, $\sum \widehat{D} = -1.824$; the below-teacher class has 0.505 and 1.918. Routing terms can be sizable in individual windows and their net can reflect cancellation, so the result concerns drift on a moving source.

Table 7 records the fixed-bank discrepancy at the reference and final checkpoints. DeepSeek's pooled above-teacher discrepancy falls from 0.0789 to 0.0032, and below-teacher from 0.1915 to 0.0149. On held-out prepared-Qwen text, the corresponding values fall from about 0.039 to 0.014 and 0.293 to 0.039. The fine-grained fit is strongest on its own trajectory contexts: at checkpoint 200 the teacher coefficient in the two-predictor log-odds-displacement regression is about 0.90, the reference-student coefficient about -0.88, and R^2 about 0.94. The same detailed regression transfers much less strongly to the held-out Qwen bank. The claim is therefore conditional refinement on the declared populations.

	Local loss				Conditional entropy			
Learner	Initial	Final	$\sum \widehat{R}$	$\sum \widehat{D}$	Initial	Final	$\sum \widehat{R}$	$\sum \widehat{D}$
DeepSeek	0.263	0.015	-0.033	-0.214	0.664	0.707	-0.049	0.093
Unprepared	0.579	0.073	0.474	-0.980	0.650	3.988	3.666	-0.327
Prepared	0.707	0.362	0.108	-0.453	2.171	0.842	-1.183	-0.146

Table 6: Dense-window cumulative accounting through bank 161, including the declared substituted evaluators. Values are rounded independently; totals telescope at full precision.

E.3 Unprepared Qwen: source motion and the value of fitting

The full primary loss decline is drift-driven, not routing-driven: $0.579 \to 0.073$ decomposes as $\sum \widehat{R}_L = +0.474$ and $\sum \widehat{D}_L = -0.980$. In the first crossing the loss actually rises ($+0.268$), because routing $+0.455$ exceeds drift -0.187. Later, the 141–161 loss drop is mostly population change (-0.115 versus drift -0.024). These phase-specific effects leave the complete decline drift-driven. By contrast, the large entropy increase is predominantly routing, as stated in the main text.

Learner	Reference class	Sites	c_W at reference	c_W at 200	Δ log-odds
DeepSeek	Above teacher	33,010	0.0789	0.0032	-2.073
DeepSeek	Below teacher	11,336	0.1915	0.0149	+1.381
Prepared Qwen	Above teacher	25,794	0.0276	0.0580	-0.606
Prepared Qwen	Below teacher	18,211	0.5627	0.3813	+0.721

Table 7: Fixed pooled-bank checking discrepancy and propensity change. Reference: checkpoint 0 for DeepSeek, checkpoint 20 for prepared Qwen. These are fixed-bank endpoints; the main figures instead sum drift over changing banks.

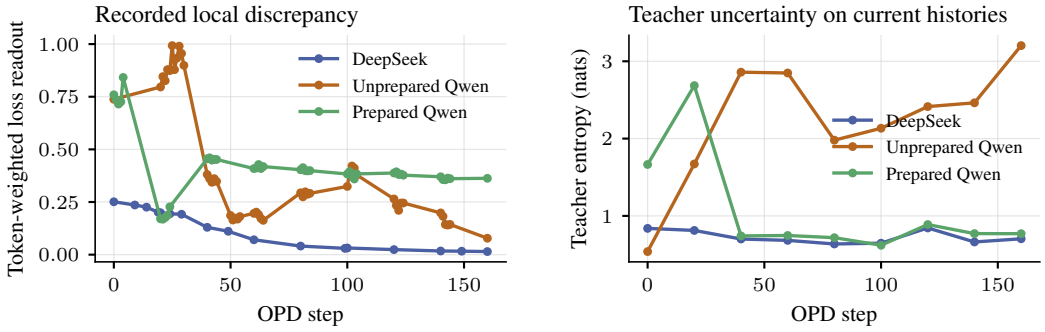


Figure 10: **Local agreement and the uncertainty of supervision.** Left: token-weighted local discrepancy on the full recorded rollout batches through the fixed-language period. Right: teacher entropy on the 32-rollout frozen trajectory-bank texts at their generation steps. Colors match Figures 4 and 11. These curves are single-run point estimates.

Figure 4a,b shows $\sum \widehat{R}_Y$, $\sum \widehat{D}_Y$, and their exact sum $M_{tt} - M_{00}$ for each readout. The total is the change from the initial cross-scored grid bank. The vertical line at update 20 is a descriptive onset reference; routing and fitting can overlap in time. The more densely sampled training-batch loss continues to rise around steps 26–29 before declining; that peak falls between points of the 20-step crossing grid.

The initial evaluator assigns entropy 4.99 nats to batch-41 histories; checkpoint 160 assigns 3.99. On those same histories the local discrepancy falls from about 0.79 to 0.06. Thus substantial fitting can coexist with persistently uncertain continuations, and the learner actually sharpens this particular receiver bank. The failure is not explained by uniformly entropy-increasing drift. The stored fixed batch-41 grid contains these means at every available checkpoint through 160; the initial-evaluator values are 0.788908 for loss and 4.987408 for entropy. The corresponding endpoint values are 0.056665 and 3.988516. Both curves use stored values on the same recorded prefixes, with the same denominator, at every evaluation. The large difference in their fractional changes therefore shows that the loss and entropy readouts remain distinct even on the identical receiver bank.

The source–kernel measurements supply a directional routing probe. Through the initial kernel, fields from batches 1, 21, and 41 suppress end-of-turn log probability at tested stopping sites and increase cumulative taken-token log-probability response along deep prefixes. The batch-21 field's reach response at depth 2,048 is about $+124$k and $+141$k per unit learning rate on the batch-41 and batch-61 receiver banks, respectively. Stopped rollouts in the 32-sample bank decline from 31 initially to 6 at batches 41 and 61. Fields from batches 61 and 101 partly reverse these tendencies, accompanying a temporary reduction of generated entropy. Later uncertainty rises again, so the push toward deeper continuation can be nonmonotone and partly reversed.

The flatness check evaluates the initial, step-20, and step-30 learners on identical deep batch-21/31 histories. Their entropy differs little relative to the large between-history contrast. These tests link the population component to changes in decisions that favor deeper continuation, while showing that much of the resulting uncertainty pre-exists in the base learner. The expanded histories then form the next training source. Teacher entropy on the sampled banks grows from 0.54 to 3.20 nats by batch 161.

Late front deterioration: mostly routing, with no demonstrated deep-to-front leak. On the initial front bank, entropy increases from 0.70 at checkpoint 140 to 0.96 at 160. In the contemporaneous front-bank crossing, however, the total increase of about 2.44 nats splits into routing $+2.18$ and drift $+0.25$. Later fronts were already uncertain under earlier evaluators. Moreover, the source-weighted depth split of $K_0\Gamma^{[141]}$ on the coherent front is about $+16.0 = +15.6$ from shallow sources plus $+0.4$ from deep sources. These measurements therefore do not support a blanket causal narrative in which entropy spreads backward from deep requests. The comparison concerns the measured raw direction on this receiver bank; the native optimizer step may distribute the change differently.

E.4 Continuation value at the unprepared student's answer boundaries

Receivers: the initial student's own rollouts on the 50 held-out MATH-500 problems (251–300), samples 0 and 1 of each problem, cut at the token boundary before the final answer marker (`\boxed` or `Answer:`); 86 sites (prefix positions). Each evaluator (initial student, updates 20–200 of the unprepared run, the JustRL teacher) is scored on the identical prefixes: the next-token mass on the answer class at the deployment temperature, and six continuations from each of three branches at the deployment protocol ($T = 0.6$, top-p 0.95, $16{,}384$ new tokens max): the answer token, the evaluator's most probable non-answer token, and the natural continuation. At each site the change of the natural value relative to the initial student is split into *steering*, $\Delta\pi\,(\bar{Q}_{\mathrm{ans}} - \bar{Q}_{\mathrm{cont}})$, and *installation*, $\bar{\pi}\,\Delta Q_{\mathrm{ans}} + (1-\bar{\pi})\,\Delta Q_{\mathrm{cont}}$ (π the answer mass, bars midpoints); intervals are $2{,}000$-draw bootstraps over sites. Single run.

	log-odds of	correctness			change vs. initial student		
Evaluator	answering	answer	continue	natural	ΔV	steering	installation
initial student	$+2.4$	0.30	0.22	0.27	—	—	—
update 20	-3.9	0.30	0.29	0.29	$+0.01$	-0.03	$+0.05$
update 30	-4.9	0.33	0.32	0.31	$+0.04$	-0.03	$+0.07$
update 40	-6.1	0.29	0.27	0.26	-0.01	-0.04	$+0.03$
update 50	-6.7	0.31	0.17	0.21	-0.09	-0.08	-0.02
update 100	-7.5	0.34	0.03	0.06	-0.24	-0.15	-0.09
update 200	-7.4	0.39	0.13	0.17	-0.15	-0.13	-0.02
teacher	-3.5	0.42	0.58	0.60	$+0.31$	$+0.04$	$+0.27$

Table 8: Two dynamics at the 86 frozen answer boundaries: the mean log-odds of the answer branch, and the correctness value of the answer branch, the continue branch and the natural continuation, with the steering/installation split of the natural value's change. At update 100 the intervals are ΔV $[-0.32, -0.16]$, steering $[-0.20, -0.10]$, installation $[-0.14, -0.04]$. The teacher also moves mass off answering but its continue branch is worth more than answering.

E.4.1 Fixed-bank entropy and loss: measured sharpening

The entropy-sensitivity identity and illustrative examples are in Section C.9. This subsection examines the fixed batch-41 receiver bank.

The measured middle phase includes sharpening. On the batch-41 bank, teacher entropy is 2.859 nats, below both the initial student's 4.987 and checkpoint 160's 3.989. The same-bank loss falls from 0.789 to 0.057. Thus the teacher is uncertain relative to coherent initial text but still less entropic than the student on these histories: fitting here accompanies sharpening, not continued uniform flattening. This sign is observed in the fixed-bank scores; for the native optimizer, it depends on more than the lower teacher entropy. The source–kernel probes likewise show sharpening from the batch-31/41 fields on the tested flat receivers, after the onset field's transient entropy-raising effect. Other fields and receivers can have the opposite sign, so the directional probes and finite-checkpoint readouts can differ in sign on a given bank or window.

Loss, entropy, and terminal score therefore have complementary roles. Loss establishes conditional fitting; entropy identifies the large excursion with changed exposure; Equation (21) explains why neither establishes useful continuation advantage. The argument holds even where the failed policy sharpens at some fixed prefixes or updates.

E.5 PREPARED QWEN: EXPANSION, REQUESTS, AND MODERATION

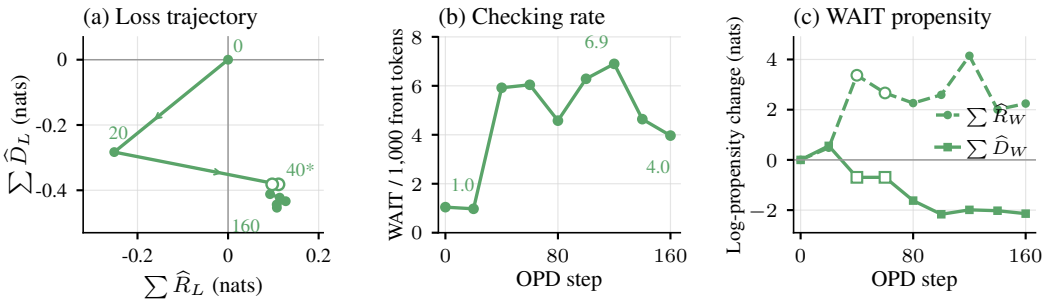


Figure 11: **Full prepared-run phase account.** Left: cumulative loss contributions. Middle: emitted WAIT frequency in the first 512 tokens. Right: cumulative routing (dashed) and drift (solid) of mean log-WAIT propensity on candidate front sentence starts. The last panel decomposes log-propensity, a readout separate from emitted frequency. Open markers indicate the evaluator substitutions stated in Section E.1.

The prepared run has distinct early events. Its token-weighted conditional entropy initially increases from 2.17 to 4.33, with routing $+1.98$ and drift $+0.18$, while its local loss declines from 0.707 to 0.173. The subsequent loss rebound to 0.437 has routing $+0.363$ and drift -0.099 under the batch-21/checkpoint-20 to batch-41/checkpoint-50 crossing. Entropy on the latter bank is instead low, about 0.70, and remains approximately 0.6–0.9 on the later primary banks. Loss rebound and the initial entropy excursion therefore describe different intervals.

The first-512-token emitted WAIT rates at batches 1, 21, 41, 61, 81, 101, 121, 141, and 161 are 1.04, 0.98, 5.92, 6.04, 4.58, 6.29, 6.90, 4.64, and 3.97 per thousand tokens. They fluctuate during the elevated phase rather than forming a smooth single peak. The claim of moderation is relative to that elevated phase, with substantial checking retained above the start.

The pooled mean log-propensity on candidate front starts has positive drift in the first crossing ($+0.56$) and negative drift by checkpoint 50 (-1.25), despite positive routing ($+2.89$) in the expansion crossing. Further conditional decreases accumulate through checkpoint 100. Later fluctuations, including much of the 121–141 decline, are predominantly population change in this readout. Hence only some of the changes in emitted frequency are attributed to drift.

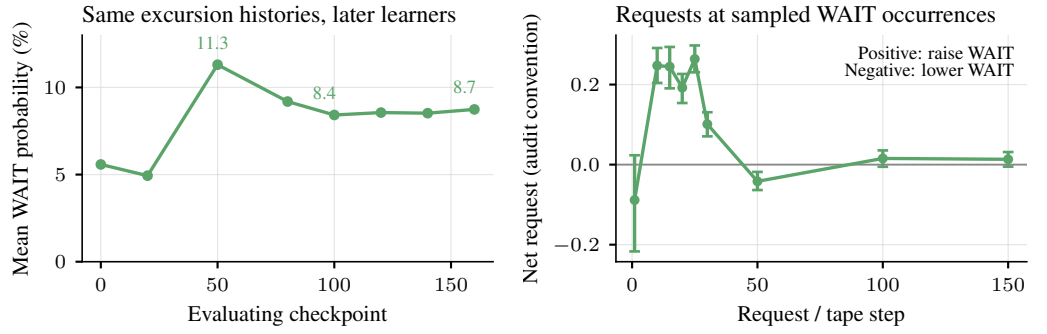


Figure 12: **Conditional moderation and changing requests in the prepared run.** Left: arithmetic mean WAIT probability on the same batch-41 candidate front contexts, rescored by each learner. Right: the separate pre-answer sampled-WAIT audit, with its reported 95% intervals; positive requests raise the sampled token and negative requests lower it in the audit convention. The two panels use different receiver/source populations.

The arithmetic probability grid supplies a complementary fixed-text check. On batch-41 contexts, mean WAIT probability is 0.0494 under checkpoint 20, 0.1130 under checkpoint 50, 0.0842 under

checkpoint 100, and 0.0874 under checkpoint 160. Thus an elevated conditional probability is subsequently reduced without returning to the earlier level. On some other banks arithmetic-probability and log-propensity changes have different signs, so each readout is reported separately.

Request audit versus delivered response. The separate request audit samples pre-answer WAIT occurrences from full training batches. Its mean local request is about $+0.19$ to $+0.26$ at batches 10–25, with the reported intervals above zero, and -0.0418 at batch 50 (95% interval $[-0.0635, -0.0182]$); the negative fraction there is 0.528. Later audited means are nearer zero. This is direct evidence that the request population at sampled WAIT occurrences changes during expansion. The crossed first-order response $K_i\Gamma^{[\ell]}$ also depends on the kernel and the whole batch. In particular, a field can suppress its recorded WAIT placements while promoting the family elsewhere, so source requests and delivered responses can differ in sign.

Selectivity with imperfect calibration. The prepared response retains dependence on teacher propensity after conditioning on the reference student's propensity, including on held-out contexts. However, its fixed-bank above-teacher binary discrepancy increases from 0.0276 to 0.0580, while the below-teacher discrepancy decreases from 0.5627 to 0.3813. The split by teacher probability is particularly informative: on above-teacher contexts with $q_W \geq 0.05$, mean student probability rises approximately $0.33 \to 0.50$ on own text and $0.29 \to 0.46$ on held-out text, against teacher probability about 0.15. These rows show residual over-use alongside oversuppression. The main claim is phased, teacher-related behavioral adjustment, with calibration improved only in part.

Scope of the feedback interpretation. The rate excursion, fixed-context moderation, changing request signs, and later held-out gain support a productive feedback interpretation. The evidence shows a different ensuing interaction after preparation. A complete mediation analysis of the accuracy improvement remains open, including the contribution of each newly exposed source subset to subsequent correction and whether any particular subset is necessary for the gain.

E.6 Checkpoint-alignment sensitivity and depth-resolved readouts

Learner	Readout	Dense $\sum \widehat{R}$	Dense $\sum \widehat{D}$	Exact-only $\sum \widehat{R}$	Exact-only $\sum \widehat{D}$
DeepSeek	Loss	-0.033	-0.214	-0.033	-0.214
DeepSeek	Entropy	-0.049	+0.093	-0.049	+0.093
Unprepared	Loss	+0.474	-0.980	+0.440	-0.946
Unprepared	Entropy	+3.666	-0.327	+3.730	-0.392
Prepared	Loss	+0.108	-0.453	+0.063	-0.408
Prepared	Entropy	-1.183	-0.146	-1.038	-0.291

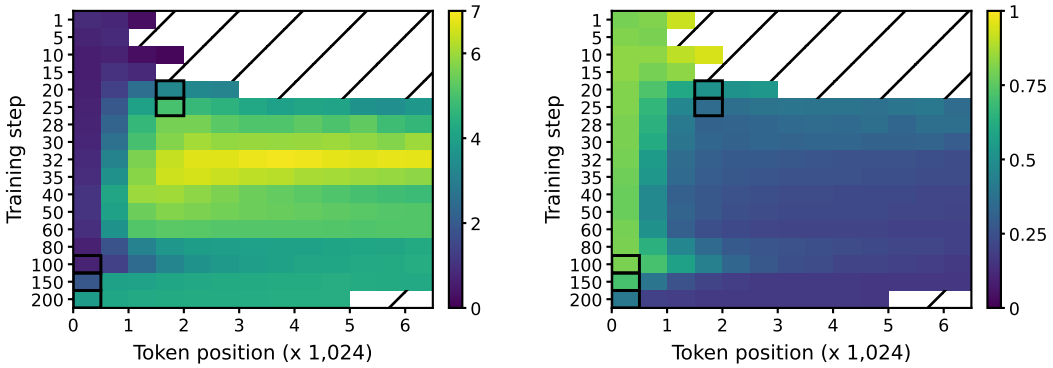
Table 9: Sensitivity to checkpoint substitutions. Exact-only accounting retains available true generating checkpoints and recomputes the four-cell split over wider windows. All rows start at initialization and end at checkpoint 160.

The coarse sensitivity in Table 9 keeps only banks with an exactly matched generating checkpoint. Cumulative allocations change with these wider windows, but the primary distinctions remain: DeepSeek loss is predominantly conditional refinement, failed-run entropy is predominantly population change, and the prepared loss rebound remains a population contribution opposed by conditional improvement. These alternative allocations are computed from measured checkpoints only.

Figure 13 gives the depth-resolved onset chronology also shown in Figure 4c,d. These full-batch measurements have finer temporal sampling near onset than the 32-rollout retrospective grid. They measure sampled student surprisal and teacher top-1 probability on the same source histories. Each map has 171 measured and 50 unreported cells. The row labels use batch indices, including the "tape step" labels in the main-text heatmaps. Later rows are descriptive; the primary accounting runs through batch 161. The definitions and coverage are detailed below.

E.6.1 Heatmap quantities, coverage, and scoring

The following conventions accompany the full-batch matrices used in Figure 4.



(a) Sampled student surprisal.

(b) Teacher top-1 probability.

Figure 13: **Depth-resolved full-batch onset readouts.** Both quantities are read on student-generated histories, in 512-token bins. Hatched cells mark unreported values. The chronology is descriptive.

	Student surprisal (nats)		Teacher top-1 probability	
Batch	First 512	1536–2047	First 512	1536–2047
1	.85	—	.79	—
5	.69	—	.81	—
10	.58	.12	.83	.94
15	.58	—	.83	—
20	.62	3.24	.81	.51
25	.66	4.97	.81	.35
28	.70	5.55	.80	.31
30	.74	5.71	.80	.33
32	.83	6.38	.79	.30
35	.90	6.47	.78	.29
40	1.04	6.00	.77	.27
50	1.05	5.69	.77	.27
60	.95	5.20	.79	.24
80	.65	3.70	.80	.34
100	.66	3.19	.80	.43
150	1.89	4.14	.71	.20
200	3.85	4.26	.40	.17

Table 10: Two position windows from the heatmaps. The first column indexes rollout batches; batch ℓ is generated after $\ell - 1$ updates. Dashes mean unreported. Student values are sampled-token surprisal.

Training samples. The heatmaps describe the unprepared Qwen run's own training samples: 256 rollouts per update on DAPO prompts, sampled at temperature one without truncation and with an 8,192-token cap.

Heatmap quantities and coverage. The heatmaps use 17 rollout-batch indices: 1, 5, 10, 15, 20, 25, 28, 30, 32, 35, 40, 50, 60, 80, 100, 150, and 200. Equally spaced rows denote selected batches at unequal training-time intervals. Columns are 512-token windows from $[0, 512)$ through $[6144, 6656)$; the generation cap remains 8,192. Color scales are fixed within each quantity across all checkpoints. Outlines descriptively mark the extension at batches 20/25 and the first-window sequence at batches 100/150/200.

The student quantity is mean sampled-token surprisal $-\log p_t(a_j \mid x_j)$ on temperature-one, untruncated rollouts. Teacher top-1 is the mean $\max_a q(a \mid x_j)$ on those same student histories, independent of the sampled token. The teacher remains fixed at every exact prefix. The two matrices have 171 reported and 50 unreported cells each; unreported cells can still contain visited tokens. Cell colors encode means; per-cell sample counts are omitted. Each deep-window average conditions on its tokens being observed there, and window contents differ across checkpoints.

At batches 20 and 25, the reported fractions reaching windows beginning at tokens 0, 512, 1024, 1536, 2048, and 2560 are respectively $(1.00, .92, .46, .22, .14, .07)$ and $(1.00, .98, .68, .41, .29, .23)$. At batch 25, the fractions at starts 3072, 3584, and 4096 are $.18, .14, .12$. These reach fractions complement the conditional depth-bin averages. A reach fraction is not itself a normalized training-source weight, and a window-level change is not an estimate of Λ for a fixed exact prefix.

F THE BROADCAST IS COARSE: READOUTS OF AVERAGE AND CONDITIONAL BEHAVIOR

This extension complements Section 3.3 by asking which components of behavior receive an update. We first derive the matched-pair response identities, then examine reasoning and refusal readouts and their measurement protocols. The motivating empirical pattern is that average behavior can change more readily than context-specific discrimination. The identities identify the relevant feature factors; whether a given update shows this ordering is an empirical question.

The reasoning measures below concern emitted marker rates grouped by problem difficulty and execution following a move. They complement the teacher-relative WAIT probabilities on fixed sentence-start contexts in Sections 3.3 and E.2, and either conditional readout can gain on its own.

F.1 THEORY: AVERAGE AND CONDITIONAL CHANNELS OF THE BROADCAST

We give the proofs here so this extension is self-contained. Let $p_\theta(\cdot \mid x) = \mathrm{softmax}(z_\theta(x))$ on a fixed finite vocabulary. All probe states and readout weights are fixed. Assume $z_\theta(x)$ is C^2 on an open neighborhood of each parameter segment or compact path considered. Write $\Phi_\theta(x) \in \mathbb{R}^{|\mathcal{V}| \times d}$ for the logit Jacobian, with row a equal to $\nabla_\theta z_\theta(x, a)^\top$, and $\mathsf{A}^p_\theta(x) = I - \mathbf{1} p_\theta(\cdot \mid x)^\top$. Gradients of scalar readouts are parameter-space column vectors.

Definition F.1 (Centered readout and readout feature)**.** For $w \in \mathbb{R}^{|\mathcal{V}|}$ with $\mathbf{1}^\top w = 0$, define

$$f_{w,\theta}(x) = w^\top \log p_\theta(\cdot \mid x), \qquad \psi_\theta(x; w) = \nabla_\theta f_{w,\theta}(x).$$

Since $\log p_\theta = z_\theta - \log \sum_a e^{z_\theta(a)} \mathbf{1}$, $f_{w,\theta}(x) = w^\top z_\theta(x)$ and $\psi_\theta(x; w) = \Phi_\theta(x)^\top w$.

Proposition F.2 (Single-step broadcast with a bounded remainder)**.** *Let $\theta^+ = \theta + \Delta\theta$, and write $\delta f_w(x) = f_{w,\theta^+}(x) - f_{w,\theta}(x)$. Then*

$$\delta f_w(x) = \int_0^1 \psi_{\theta + r\Delta\theta}(x; w)^\top \Delta\theta \, dr.$$

In particular, for $\Delta\theta = \eta u$ with a fixed direction u,

$$\delta f_w(x) = \eta \langle \psi_\theta(x; w), u \rangle + R_w(x), \qquad |R_w(x)| \le \tfrac{1}{2} \eta^2 M_w(x) \|u\|^2, \tag{25}$$

where $M_w(x) = \sup_{0 \le r \le 1} \|\nabla^2_\theta f_{w,\theta + r\eta u}(x)\|_{\mathrm{op}} < \infty$.

Proof. By Definition F.1, $\psi = \Phi^\top w$ is the exact gradient. For $F(r) = f_{w,\theta + r\Delta\theta}(x)$, the chain rule gives $F'(r) = \psi_{\theta + r\Delta\theta}(x; w)^\top \Delta\theta$. The fundamental theorem of calculus proves the integral. Differentiating again gives $F''(r) = \Delta\theta^\top \nabla^2_\theta f_{w,\theta + r\Delta\theta}(x) \Delta\theta$. The integral Taylor remainder is $\int_0^1 (1 - r) F''(r) \, dr$. Bounding the quadratic form by $M_w(x) \|\Delta\theta\|^2$ proves Equation (25). The supremum is finite by continuity on the compact segment. On a fixed compact neighborhood and for bounded u, it can be chosen uniformly for sufficiently small η, yielding the usual $O(\eta^2)$ notation. □

Definition F.3 (Average and conditional pair components)**.** Let $(x_i^H, x_i^B)_{i=1}^n$ be fixed ordered pairs. For any pair-indexed vector $v = (v_i^H, v_i^B)_{i=1}^n$, let S swap the members of each pair. Define

$$v_{\mathrm{avg}} = \tfrac{1}{2}(I + S)v, \qquad v_{\mathrm{cond}} = \tfrac{1}{2}(I - S)v, \qquad v = v_{\mathrm{avg}} + v_{\mathrm{cond}}. \tag{26}$$

Writing $\bar{v}_i = (v_i^H + v_i^B)/2$ and $\Delta v_i = (v_i^H - v_i^B)/2$, the full average component has pair $(\bar{v}_i, \bar{v}_i)$ and the full conditional component has pair $(\Delta v_i, -\Delta v_i)$. Below, $f_{\mathrm{avg},i}$ and $f_{\mathrm{cond},i}$ denote the representative half-sum and signed half-difference, respectively. This definition applies to pair-indexed scores or source contributions, not to an aggregate parameter vector without a pair index.

Lemma F.4 (Orthogonality under matched pair weights)**.** *Suppose the two members of pair i have the same positive weight α_i. Under $\langle v, w\rangle_\alpha = \sum_i \alpha_i[(v_i^H)^\top w_i^H + (v_i^B)^\top w_i^B]$, the two maps in Equation* (26) *are complementary orthogonal projections. Hence $\|v\|_\alpha^2 = \|v_{\mathrm{avg}}\|_\alpha^2 + \|v_{\mathrm{cond}}\|_\alpha^2$.*

Proof. Swapping twice is the identity, so $S^2 = I$. Equal weights within pairs give $\langle Sv, w\rangle_\alpha = \langle v, Sw\rangle_\alpha$, hence S is self-adjoint. Therefore $P_\pm = (I \pm S)/2$ satisfy $P_\pm^2 = P_\pm$, $P_\pm^* = P_\pm$, $P_+P_- = 0$, and $P_+ + P_- = I$. In particular, $\langle P_+v, P_-v\rangle_\alpha = \langle v, P_+P_-v\rangle_\alpha = 0$. Expanding the squared norm of their sum proves the identity. Zero-weight pairs can be omitted. With unequal weights, the algebraic split still holds but need not be orthogonal in that weighted inner product. □

Proposition F.5 (Expression and emission of conditional structure)**.** *At a fixed checkpoint, define $\bar\Phi_i = (\Phi(x_i^H) + \Phi(x_i^B))/2$, $\Delta\Phi_i = (\Phi(x_i^H) - \Phi(x_i^B))/2$, and analogously $\bar B_i, \Delta B_i$ for the numerical requests. For one centered readout vector w_i shared by both members of a probe pair,*

$$\begin{aligned} &\psi_{\mathrm{avg},i} = \bar\Phi_i^\top w_i, \qquad \psi_{\mathrm{cond},i} = \Delta\Phi_i^\top w_i, \\ &u_{\mathrm{pair},i} := \Phi(x_i^H)^\top B_i^H + \Phi(x_i^B)^\top B_i^B = 2(\bar\Phi_i^\top \bar B_i + \Delta\Phi_i^\top \Delta B_i). \end{aligned} \tag{27}$$

The first line concerns expression at a probe. The second is the unnormalized sum of the two training pullbacks. For arbitrary member weights α_i^H, α_i^B, the weighted sum instead equals

$$(\alpha_i^H + \alpha_i^B)(\bar\Phi_i^\top \bar B_i + \Delta\Phi_i^\top \Delta B_i) + (\alpha_i^H - \alpha_i^B)(\bar\Phi_i^\top \Delta B_i + \Delta\Phi_i^\top \bar B_i).$$

Proof. Since the same fixed w_i is used on both members, differentiation commutes with their half-sum and half-difference. Substituting $\nabla_\theta f_{w_i}(x) = \Phi(x)^\top w_i$ gives the first line of Equation (27). For emission, substitute $\Phi^H = \bar\Phi + \Delta\Phi$, $\Phi^B = \bar\Phi - \Delta\Phi$, $B^H = \bar B + \Delta B$, and $B^B = \bar B - \Delta B$:

$$\begin{aligned} (\Phi^H)^\top B^H &= \bar\Phi^\top \bar B + \bar\Phi^\top \Delta B + \Delta\Phi^\top \bar B + \Delta\Phi^\top \Delta B, \\ (\Phi^B)^\top B^B &= \bar\Phi^\top \bar B - \bar\Phi^\top \Delta B - \Delta\Phi^\top \bar B + \Delta\Phi^\top \Delta B. \end{aligned}$$

Adding cancels the two odd cross terms and gives the factor of two. Multiplying the first line by α_i^H and the second by α_i^B before adding proves the weighted expression. □

Theorem F.6 (Weighted broadcast block structure)**.** *Suppose the entire raw update is formed from finitely many source pairs j, with weight $\alpha_j \geq 0$ for each member and $2\sum_j \alpha_j = 1$. Thus $u = \sum_j \alpha_j u_{\mathrm{pair},j}$. Probe pair i with one fixed centered w_i shared by its members, and take $\Delta\theta = \eta u$. Define $\bar K_{ij} = \bar\Phi_i \bar\Phi_j^\top$ and $\Delta K_{ij} = \Delta\Phi_i \Delta\Phi_j^\top$, all at the initial checkpoint. Then*

$$\begin{aligned} \delta f_{\mathrm{avg},i} &= 2\eta \sum_j \alpha_j w_i^\top [\bar K_{ij} \bar B_j + \bar\Phi_i \Delta\Phi_j^\top \Delta B_j] + R_{\mathrm{avg},i}, \\ \delta f_{\mathrm{cond},i} &= 2\eta \sum_j \alpha_j w_i^\top [\Delta\Phi_i \bar\Phi_j^\top \bar B_j + \Delta K_{ij} \Delta B_j] + R_{\mathrm{cond},i}. \end{aligned} \tag{28}$$

For either component $\sigma \in \{\mathrm{avg}, \mathrm{cond}\}$, $|R_{\sigma,i}| \leq \eta^2 M_{\sigma,i} \|u\|^2/2$, where $M_{\sigma,i} = \sup_{0\leq r\leq 1} \|\nabla_\theta^2 f_{\sigma,i}(\theta + r\eta u)\|_{\mathrm{op}}$.

Proof. Each component readout is C^2 because it is a half-sum or half-difference of two C^2 readouts. The scalar integral Taylor formula from Proposition F.2 applies directly to it. By Proposition F.5, the first-order average term is

$$\begin{aligned} \eta\langle \psi_{\mathrm{avg},i}, u\rangle &= 2\eta \sum_j \alpha_j \langle \bar\Phi_i^\top w_i, \bar\Phi_j^\top \bar B_j + \Delta\Phi_j^\top \Delta B_j\rangle \\ &= 2\eta \sum_j \alpha_j w_i^\top [\bar\Phi_i \bar\Phi_j^\top \bar B_j + \bar\Phi_i \Delta\Phi_j^\top \Delta B_j]. \end{aligned}$$

For the conditional readout its gradient is $\Delta\Phi_i^\top w_i$, so the same inner-product calculation gives $2\eta \sum_j \alpha_j w_i^\top [\Delta\Phi_i \bar\Phi_j^\top \bar B_j + \Delta\Phi_i \Delta\Phi_j^\top \Delta B_j]$. These are exactly the two leading terms. The Taylor bound applied to each component proves the stated remainders; no kernel is assumed constant on the segment. □

For n uniformly weighted pairs, $\alpha_j = 1/(2n)$; thus each pair carries total weight $2\alpha_j = 1/n$ in the normalized update, so the update averages over the $2n$ sites. For matched weights the source pressure splits as

$$P = P_{\text{avg}} + P_{\text{cond}}, \quad P_{\text{avg}} = 2\sum_j \alpha_j \|\bar{B}_j\|^2, \quad P_{\text{cond}} = 2\sum_j \alpha_j \|\Delta B_j\|^2,$$

by Lemma F.4. Nonzero P_{cond} does not imply nonzero conditional emission, because $\Delta\Phi_j^\top \Delta B_j$ can vanish or cancel across sites. A harmful-only pool has unequal pair weights and therefore retains the additional terms in Proposition F.5.

The conditional-to-conditional block ΔK_{ij} contains feature differences at both ends, whereas $\bar{K}_{ij}$ contains neither. The average-to-conditional term $\Delta\Phi_i \bar{\Phi}_j^\top \bar{B}_j$ can also move a conditional readout. A shared component of $\Delta\Phi_i$ can generate a shared contribution; signs across pairs and the ordering of channel magnitudes depend on the particular features, readouts, and requests. An additional source direction outside the paired population contributes $\eta\langle\psi_{\sigma,i}, u_{\text{extra}}\rangle$ at first order; a native optimizer step must use its actual direction rather than silently identifying it with the paired raw sum.

Definition F.7 (Feature resolution)**.** For a shared centered pair readout with $\|\psi_{\text{avg},i}\| > 0$, define $\rho_i(w_i) = \|\psi_{\text{cond},i}\|/\|\psi_{\text{avg},i}\|$. If the denominator vanishes, use $\|\psi_{\text{cond},i}\|$ directly; the ratio is undefined.

Corollary F.8 (Necessary and sufficient condition for first-order conditional motion)**.** *For any fixed parameter direction u,*

$$\frac{d}{d\eta} f_{\text{cond},i}(\theta + \eta u)\bigg|_{\eta=0} = \langle \psi_{\text{cond},i}, u\rangle.$$

This derivative vanishes for every u if and only if $\psi_{\text{cond},i} = 0$. For a specified u, it is nonzero if and only if $\langle\psi_{\text{cond},i}, u\rangle \neq 0$. When ρ_i is defined, existence of some direction with nonzero derivative is equivalent to $\rho_i > 0$. Moreover, for $\eta \geq 0$,

$$|\delta f_{\text{cond},i}| \leq \eta\|\psi_{\text{cond},i}\|\|u\| + \tfrac{1}{2}\eta^2 M_{\text{cond},i}\|u\|^2.$$

Proof. The derivative is the directional derivative of the differentiable scalar readout. If its gradient is zero, its inner product with every u is zero. Conversely, if the gradient is nonzero, choosing $u = \psi_{\text{cond},i}$ makes the derivative $\|\psi_{\text{cond},i}\|^2 > 0$. The claim for a specified direction is the displayed equality itself, and the claim about ρ_i follows from its positive denominator. The finite-step bound follows from Proposition F.2 and $|\langle\psi, u\rangle| \leq \|\psi\|\|u\|$. □

This is a first-order statement: a zero initial derivative still allows a nonzero second-order finite-step change. When both feature norms are nonzero and the average derivative is nonzero, the ratio of first-order magnitudes is $\rho_i|\cos(\psi_{\text{cond},i}, u)|/|\cos(\psi_{\text{avg},i}, u)|$. Thus ρ_i alone does not bound that ratio; the average alignment can be arbitrarily small.

Proposition F.9 (Accumulated conditional response)**.** *Let $I = [a, b]$ and let $\theta(t)$ be a continuous piecewise C^1 path with finitely many pieces. With the probe pair and weights fixed,*

$$|\Delta_I f_{\text{cond},i}| \leq \int_a^b \|\psi_{\text{cond},i}(\theta(t))\|\|\dot{\theta}(t)\|\, dt \leq \sup_{t\in I}\|\psi_{\text{cond},i}(\theta(t))\|\,\text{len}_I(\theta), \tag{29}$$

where $\text{len}_I(\theta) = \int_a^b \|\dot{\theta}(t)\|\, dt$ and $\Delta_I f_{\text{cond},i} = f_{\text{cond},i}(\theta(b)) - f_{\text{cond},i}(\theta(a))$.

Proof. On each smooth piece, the chain rule gives $\frac{d}{dt} f_{\text{cond},i}(\theta(t)) = \langle\psi_{\text{cond},i}(\theta(t)), \dot{\theta}(t)\rangle$. Integrating separately over the finitely many pieces and telescoping their continuous boundary values gives the exact integral for $\Delta_I f_{\text{cond},i}$. The triangle inequality and pointwise Cauchy–Schwarz give the first bound. Continuity of $\psi_{\text{cond},i}$ on the compact path makes its supremum finite; moving it outside the integral gives the second. □

If the conditional feature vanishes throughout the path, the readout cannot change. If it is small but nonzero, a longer path can accumulate a larger change even without feature growth. Relative resolution must be accompanied by the scale of the average feature. Small conditional requests, weak conditional emission, limited receiver sensitivity, and update misalignment are distinct explanations to test; the bound is compatible with all four and, by itself, constrains the conditional readout rather than behavioral AUC.

Scope of prompt-specific and uncentered readouts. The shared-centered-weight hypothesis is essential for the expression $\psi_{\mathrm{cond},i} = \Delta\Phi_i^\top w_i$. For arbitrary fixed vectors w_i^H, w_i^B, define $f_i^\sigma = (w_i^\sigma)^\top \log p(x_i^\sigma)$ for $\sigma \in \{H, B\}$. Differentiating log-softmax gives

$$\psi_i^\sigma = (\Phi_i^\sigma)^\top \chi_i^\sigma, \qquad \chi_i^\sigma = (\mathsf{A}_i^{p,\sigma})^\top w_i^\sigma = w_i^\sigma - p_i^\sigma(\mathbf{1}^\top w_i^\sigma).$$

Let $\bar{\chi}_i = (\chi_i^H + \chi_i^B)/2$ and $\Delta\chi_i = (\chi_i^H - \chi_i^B)/2$. Directly expanding the half-sum and half-difference of these two gradients proves

$$\psi_{\mathrm{avg},i} = \bar{\Phi}_i^\top \bar{\chi}_i + \Delta\Phi_i^\top \Delta\chi_i, \qquad \psi_{\mathrm{cond},i} = \Delta\Phi_i^\top \bar{\chi}_i + \bar{\Phi}_i^\top \Delta\chi_i. \tag{30}$$

For example, the half-difference is $\frac{1}{2}[(\bar{\Phi}_i + \Delta\Phi_i)^\top(\bar{\chi}_i + \Delta\chi_i) - (\bar{\Phi}_i - \Delta\Phi_i)^\top(\bar{\chi}_i - \Delta\chi_i)]$; the even terms cancel and leave the conditional formula. The half-sum instead cancels the odd terms. Common centered weights give $\chi_i^H = \chi_i^B = w_i$, recovering Equation (27). With unequal or uncentered weights, the additional $\bar{\Phi}_i^\top \Delta\chi_i$ term need not vanish. The exact path identity still holds using the full gradient in Equation (30).

The safety gap projections in Section F.5 fall under this general case, with a prompt-specific direction for each twin. Their pair decomposition is exact algebra and must be interpreted through Equation (30); the pure shared-readout blocks describe the common centered case. Family log-mass scores and behavioral AUC are also different, nonlinear observables. They motivate the comparison with the theory and remain distinct from ρ_i.

F.2 Reasoning readouts: initiation and execution

Reasoning behavior at initiation and execution. We return to DeepSeek-R1-Distill-Qwen-1.5B trained under JustRL-1.5B. We distinguish a control move's initiation, such as a "wait, let me check" opening, from the execution that follows. For initiation, we measure the overall marker rate and its dependence on problem difficulty. For execution, we measure whether an initiated check or plan is carried through and used. These behavioral summaries separate average behavior from context dependence, as the theory motivates.

The AIME 2024 evaluation uses 30 problems, the initial learner and five training checkpoints, and $n = 8$ samples per problem at temperature 0.6. Difficulty bands are fixed by the base policy's own pass rate. Accuracy increases from .34 to .46 on this set. Sampling and annotation protocols are in Section F.4.

Checking: execution and verdict use improve. Execution moves the other way. Conditional on a visible continuation, the fraction of checks that reach a verdict the trajectory then uses rises from 0.71 to 0.97, execution rises similarly ($0.75 \to 0.90$), and the product, used-verdict checks per ten thousand characters, rises from 8.6 to 13.7 while responses halve in length. These are blinded-panel statistics: 635 sites on eleven problems, binary functional criteria, a supporting quote required for every judgment, gold calibration anchors passed at 11 of 11 (protocol and reliability in Section F.4). A crossed continuation probe (Section F.4) also shows that the base model can complete such checks in bounded windows. These behavioral results support the execution-side interpretation.

Planning: stable initiation rates and improved follow-through. Figure 14 repeats the measurement for planning, and the contrast carries the information. The per-token planning rate does not inflate ($1.4 \to 1.3$ markers per thousand tokens on easy problems, $2.9 \to 1.8$ on hard). The difficulty coupling of the rate survives, holding near 0.4 at every checkpoint. And the execution-conditional statistic improves as checking's does: follow-through rises ($0.64 \to 0.73$), with specificity and grounding flat. Across these two modes, the increasing checking rate loses its difficulty dependence, whereas the non-increasing planning rate retains it. The comparison leaves open whether checking calibration was weakly requested, lost during learning, or obscured by rate saturation.

Reframing is retained as a descriptive control. The panel's per-move criteria show an apparent loss there (the justified fraction falls while the aptness of new frames stays flat; Figure 18), but reframing is a trajectory-scoped steering operation: whether a switch was right depends on the fate of the directions it leaves and adopts, so per-move criteria attribute a policy-level property to single moves. We treat those series as descriptive and retain their instrument limitations in Section F.4.

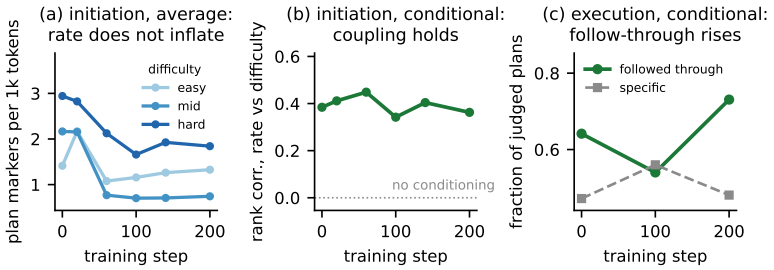


Figure 14: The planning mode in the same terms (panels (a), (b) are lexical marker rates over six checkpoints; panel (c) is the blinded panel at checkpoints 0, 100, 200). Key trends to observe: 1.) the per-token rate does not inflate; 2.) the difficulty coupling of the rate holds near 0.4 at every checkpoint; 3.) follow-through rises while specificity is flat.

The reasoning results show different changes at initiation and execution scope: checking becomes more frequent but less difficulty-dependent, while verdict use and planning follow-through improve. This pattern is compatible with distinct readout sensitivities.

Scope. All reasoning readouts are behavioral analogues of the average and conditional components in Equation (27). They bear only indirectly on the Jacobian feature norms and per-axis resolution in Corollary F.8 and on the feature factors and update alignments in Equation (28). Panel (c) of **??** scores verdict use rather than verdict correctness; development examples, excluded from the main annotation statistics, include checks that accept three-decimal numerical agreement as exact, including at step 200. The marker and probe limits in Section F.4 bound the reading further: suppressing a check's opening leaves eventual correctness unchanged, and the policy re-emits an opening within the continuation. Whether the rise in verdict use reflects more accurate adjudication remains open.

F.3 Refusal readouts: criterion and discrimination

A second domain: refusal distillation, where the dissociation is pronounced. The refusal study provides fixed prompt pairs on which to compare average response shifts and discrimination. Its log-probability instrument is evaluated at response position zero, conditioned on the prompt alone, so these probe states stay fixed across training. Behavioral refusal is scored separately on completed sampled responses. A 1.5B student is distilled on-policy from safety-tuned teachers on a fixed prompt pool (moderator and scoring details in Section F.5). The twin pairs of Definition F.3 are XSTest's matched harmful/benign prompts, scored by an LLM judge at $n = 8$ samples per prompt. The behavioral summaries are the refusal criterion c and matched-twin AUC. They measure operating point and discrimination, respectively; the exact linear average/conditional decomposition in Definition F.3 is applied separately to the fixed-support log-probability readouts.

Criterion shifts without a resolved discrimination gain. Figure 15 shows the dissociation. The untrained student has a high false-alarm rate: its criterion is $c = -0.83$, with a false-alarm rate of .41 on the safe twins, and its AUC is .860 [.829, .893]. After 93 steps of OPD from a 32B teacher at a 1024-token cap, the criterion has moved to $c = +0.03$, three quarters of the way to the teacher's $+0.38$. The AUC estimate is .849 [.814, .884], with an interval that broadly overlaps the initial interval. This is evidence for a changed operating point without a resolved behavioral AUC improvement. The pattern extends beyond this recipe: across cap, pool, teacher, and an offline-distillation variant, criteria span -0.37 to $+0.19$ while every arm's AUC stays inside the baseline band (Mann–Whitney AUC with bootstrap confidence intervals; moderator comparisons are in Section F.5).

A nearer teacher produces a larger criterion shift. The two teachers produce different criterion shifts (Figure 15). At matched cap and pool, the near 7B teacher moves the criterion to $+0.05$ while the far 32B teacher leaves it at -0.37, even though the 7B teacher's own criterion is the lower of the two ($+0.24$ against $+0.38$). The reported onset request drains 73% over training for the near teacher, against 40% for the far teacher. This drain tracks the change in the request; accessibility and collective gain, as defined in Section 3.2, measure its conversion separately. The AUC estimates of

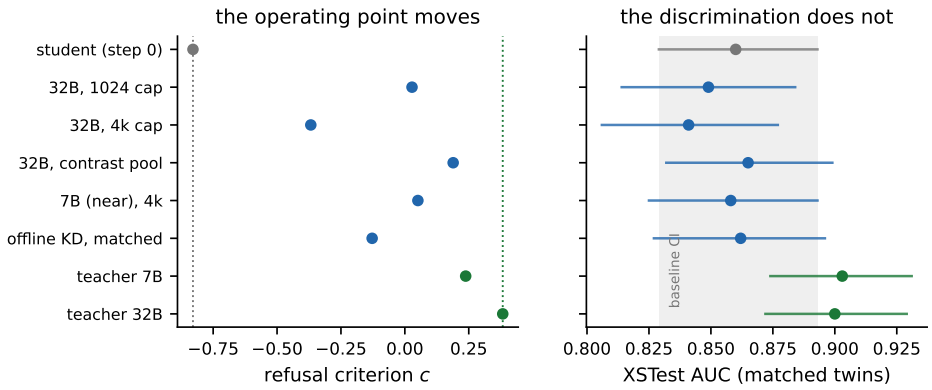


Figure 15: The safety dissociation (dots are point estimates and bars are bootstrap confidence intervals). Key trends to observe: 1.) the refusal criterion c moves most of the way from the base -0.83 toward the teachers' operating points, and at matched cap it moves much further for the near teacher than for the far one; 2.) the matched-twin AUC of every trained arm stays inside the baseline confidence band, far below the teachers' $\approx .90$. The criterion changes substantially, whereas no trained arm shows a clear increase in the displayed AUC estimates. All eight entries are shown, including the contrast-pool variant and offline control.

both arms remain within the baseline confidence band, so the larger criterion shift comes without a resolved gain in behavioral discrimination.

The conditional component is requested. The teacher-derived request instrument itself has conditional structure. The teacher's position-zero requests separate harmful from benign twins at a per-step sensitivity of $d' \approx 1.4$–2.3 (per-step counts are small, so these summaries aggregate observations), commensurate with the teacher's own behavioral $d' = 2.49$. The request is also shaped like a decision: over training the onset request halves while the body request stays flat (position-graded absorption in Figure 19).

The block structure, one window at a time. Theorem F.6 describes first-order responses. We examine the related empirical average/conditional pattern in finite inter-checkpoint windows on the 1024-cap arm's full ladder: base and nine checkpoints at ten-step spacing, 124 twin pairs, position-zero log-probabilities on a fixed support. In each window the request is the teacher-minus-base gap, the broadcast movement is the checkpoint-pair change projected on it, and both are split into the components of Equation (26) (instrument in Section F.5). Figure 16 gives the result as three statements.

1. *Magnitude.* The request field is six-to-one average to conditional (87.2 against 14.2); the broadcast movement is eleven-to-one (30.5 against 2.8). Per unit of the corresponding request statistic, the average response is larger by a factor of about 1.8, beginning in the first window.
2. *Structure.* The conditional movement includes a shared signed component. A sign-aligned mean clears $z = 3$ in three of nine windows, while the per-pair magnitude sits on the shuffled-pair noise floor in every window (3.73 against a floor of 3.74 in the largest). The signed shift is compatible with a shared contribution; the prompt-specific readout directions also contribute through Equation (30).
3. *Resolution.* The missing discrimination is missing from the logits too: a position-zero logit refusal score separates the twins no better at step 90 than at base (AUC $0.931 \to 0.936$; the teacher's is 0.979). The last-layer feature proxy also shows little differential change: within-pair and cross-pair twin similarity fall in lockstep, with no differential separation (Figure 16c).

The request has conditional structure, but the observed discrimination changes little and the last-layer proxy shows no differential twin separation. This is consistent with limited conditional delivery.

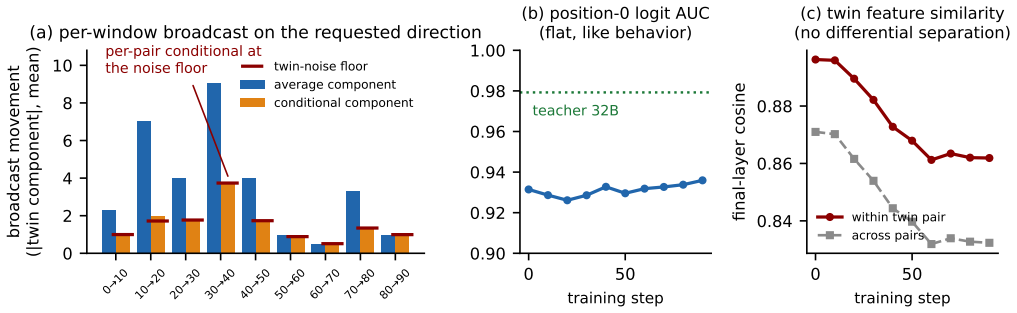


Figure 16: The per-window twin accounting on the safety 1024-cap arm (124 XSTest twin pairs, position-zero log-probabilities on a fixed support, windows of ten steps). Key trends to observe: 1.) panel (a), the average component of broadcast movement is clear of its noise scale in every window from the first, while the per-pair magnitude of the conditional component lies on the shuffled-pair noise floor in every window (a small pair-uniform conditional mean survives sign alignment; see text); 2.) panel (b), a position-zero logit-level refusal score separates the twins no better after ninety steps than at base, far below the teacher, in agreement with the flat behavioral AUC; 3.) panel (c), the resolution proxy: within-pair and cross-pair feature similarity fall in lockstep, so this last-layer proxy shows no differential separation of the twins.

Together, the two domains separate changes in average behavior from changes in context dependence. The healthy JustRL reasoning run also improves by $+0.10$ on MATH-500 while changing from 0.39 to 0.30 on MMLU-Pro. This is an off-training-distribution trade-off; its link to any particular broadcast block remains open. The complementary question of how successive updates change the population that supplies new requests is studied in Sections 3.3 and E.

Scope. Each arm is a single training run, and its bootstrap intervals are computed within that run. The dissociation concerns the operating point and the summary AUC; the full ROC curve can still change, and individual conditional log-probability contrasts can still be nonzero. The gap-projection scores use prompt-specific directions, so they relate to Theorem F.6 through Equation (30) in Section F.1. The window accounting compares readouts and is exact for these declared scores; the contribution of each kernel block, including either conditional-response block, remains open. Likewise, the request's conditional structure leaves open whether the update carries nonzero conditional emission, and the teacher comparison bears on the shared-readout condition in Corollary F.8 only indirectly. The feature probe is a last-layer proxy for the full-Jacobian ρ in Definition F.7, and the flat behavioral AUC is an observation rather than a consequence of Corollary F.8 and Equation (29) alone.

F.4 Reasoning instruments and supporting measurements

These protocols define the lexical, graded, and continuation-based measurements in Section F.2.

The difficulty-calibration analysis. The reasoning-domain broadcast measurements of Section F.2 use the 30 problems of AIME 2024, $n = 8$ samples per problem at the evaluation temperature (0.6, top-p 0.95, $24{,}576$ new tokens) at base and five checkpoints of the healthy run, with fixed lexical marker definitions (a "wait"-boundary regex for checking; a fixed phrase list for planning). AIME carries no difficulty labels, so difficulty is empirical: bands from the base policy's own pass count (easy $\geq 6/8$, seven problems; hard $\leq 2/8$, eighteen; mid otherwise). Marker statistics are computed per rollout both raw and per thousand generated tokens, then medianed within problem and within band; calibration is summarized as the hard-to-easy band ratio and the Spearman rank correlation with empirical difficulty, for both the raw and the rate versions, alongside band median lengths and truncation fractions, since raw counts conflate rate with length. The per-response check counts that remain difficulty-ordered at step 200 (hard-to-easy ratio 1.5) are ordered because hard responses are still longer, not because the model checks at a higher rate on them. A separate 40-problem

MATH-500 version of this analysis (8 per level, $n = 6$, raw counts only), where accuracy is flat across checkpoints, is retained as the flat-accuracy control cited in the text.

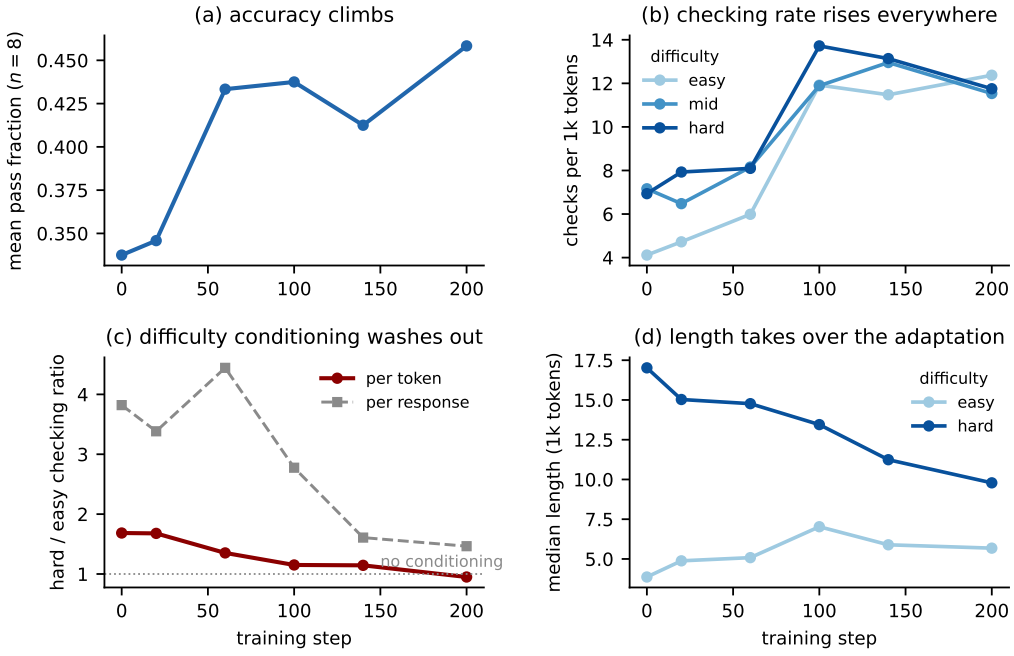


Figure 17: The full difficulty-calibration figure for the checking mode (companion to **??**; 30 AIME problems, base + 5 checkpoints, $n = 8$ samples at temperature 0.6; difficulty bands from base pass rate: easy $\geq 6/8$, hard $\leq 2/8$; lexical markers; medians over problems). Key trends to observe: 1.) accuracy climbs from .34 to .46, so this set discriminates checkpoints; 2.) per-token checking rates rise across difficulty bands and approximately triple on easy problems; 3.) both checking ratios fall toward 1: per-token from 1.7 to 1.0, per-response from 3.8 to 1.5; 4.) response length takes over the case adaptation: hard-band median length falls from 17.0k to 9.8k tokens while easy grows.

The blinded move-quality protocol. Quality of CONTROL moves is graded per move, in two stages by separate model instances. Stage one is a census: every checking, planning, or reframing move is typed by function, independent of keywords, with contentless marker runs logged once as degenerate. Stage two judges a deterministic subsample of moves (evenly spaced through each rollout's move list, up to 6/5/4 per subtype) under binary, functionally defined criteria: a check has warrant, consequence, executed verification, and a used verdict; a plan is specific, grounded, followed through, and exited with a stated reason; a reframe is diagnosed, apt, executed, and non-cycling. Ex-ante criteria see only the problem and the rollout prefix up to the move; follow-up criteria additionally see a bounded window to the end of the current segment. Checkpoint identity is blinded by a fixed permutation with the key held out. Within each selected problem, rollouts enter by a fixed arrival index, independent of correctness. Headers carry no ground truth, and every judgment must include a supporting quote of at most fifteen words or is invalid. The main annotation panel spans eleven AIME problems: the correctness transition, the full mid band, hard-band improvers, and two stable-hard controls. A development analysis on one MATH problem is excluded from these statistics. Calibration anchors include a clear pass and a clear fail for each criterion, and a subset is independently graded twice. Uncertainty is assessed at the problem level, which accounts for dependence among moves from the same response. Reliability: a declared subsample of ten batches (66 sites) was independently regraded by second instances blind to the first grades. Cohen's κ per criterion: warrant 0.79, execution 0.93, verdict use 1.00; plan specificity 1.00, grounding 0.65, follow-through 0.91, principled exit 0.82; reframe diagnosis 1.00, aptness 1.00, execution 0.86, non-cycling 0.84. The consequence criterion has $\kappa = 0.47$ despite raw agreement 0.93, with a pass prevalence of 97%; our conclusions rest on the other criteria. To prevent final-answer leakage into move-quality judgments, the primary analysis

excludes 40 of the 635 sites whose follow-up windows exposed the final answer. These exclusions are distributed across checkpoints (9–11 per checkpoint).

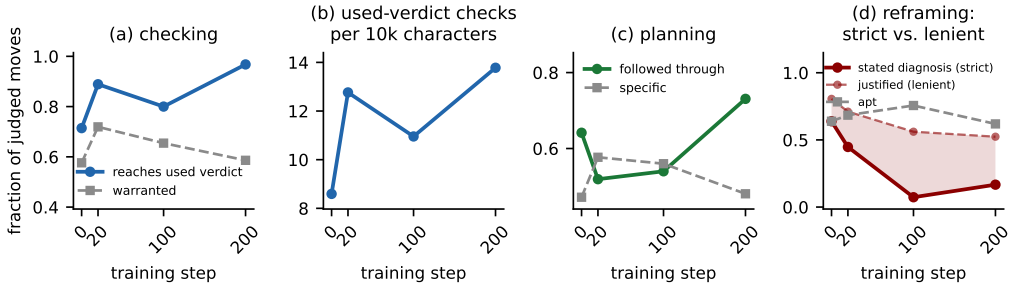


Figure 18: Appendix companion to panels (c) of **??** and Figure 14: the graded quality series over the full checkpoint set $\{0, 20, 100, 200\}$, with the strict reframing criterion. In panel (d) the strict criterion requires a stated diagnosis before the frame switch and gives $0.64 \to 0.45 \to 0.07 \to 0.17$; the lenient reading, which also credits an evident but unstated warrant, gives $0.81 \to 0.71 \to 0.56 \to 0.52$. The shaded band between them is deliberate-looking switching that goes unexplained, widest at step 100. The reasoning-readout subsection reports checkpoints $\{0, 100, 200\}$; step 20 is a transient on several axes (an overshoot in verdict use and the used-verdict rate, an undershoot in planning follow-through) and is retained here for completeness.

The reframing construct. Per-move criteria for reframing (diagnosis, aptness, execution, non-cycling) attribute a policy-level property, the allocation of persistence across solution approaches, to single moves, and they required two leniency amendments during grading (credit for evident but unstated warrants; a purposeful-restart exception to non-cycling). We therefore report the reframing series in Figure 18 descriptively; our quality claims rest on checking and planning. A trajectory-level measure of successful reframing remains to be validated.

The counterfactual move probe. Marked CHECK and PLAN sites are sampled from the middle 15–85% of AIME rollouts (regex openings, difficulty bands from base pass rate). For each site the rollout is truncated immediately before the move and continued $n = 6$ times under three conditions: *started* (the move's opening words are forced), *suppressed* (the opening tokens are banned for the continuation), and *free*. Continuations are scored for eventual correctness, and the continuation head is scanned for re-emitted move openings, which measures whether suppression actually removes the move or only rewords it. Results quoted in the text: eventual correctness is unchanged in every difficulty band at both probed checkpoints (steps 20 and 200), and under suppression the policy re-emits a move opening within the continuation head one to two times per continuation.

The crossed follow-up probe. Prefixes are cut at CHECK and PLAN openings in rollouts from two checkpoints of the healthy run (steps 20 and 200), forming two banks. Base, step-20, and step-200 policies each continue every bank ($n = 6$, 1024 new tokens). Structural proxies score each continuation: verification executed (density of mathematical content following the opening), verdict reached within 800 characters, degenerate marker run, plan instantiated, and plan persisted. Crossing the prefix-generating checkpoint with the continuing checkpoint separates changes on fixed contexts from differences between generated context banks. Results summarized in the text: training raises the fraction of continuations that execute a verification ($0.57 \to 0.66$ on early-era prefixes) while halving the fraction that reach a verdict within the probe window ($0.37 \to 0.19$ early-era, $0.45 \to 0.27$ late-era); degenerate marker runs rise ($0.04 \to 0.12$ and $0.05 \to 0.24$). Within the bounded window, the base policy reaches a verdict more often on these check prefixes than the trained policy does, and the prefix-generating checkpoint matters much less than the identity of the continuing checkpoint.

Scope. The lexical markers are behavioral proxies for a semantic classification of moves. The counterfactual probe's suppression arm bans surface forms rather than applying a splice-grade lesion to the functional move, so the probe bounds substitutability rather than per-move future advantage. The crossed probe's bounded-window result must be read alongside the full-segment panel: probe

verdicts are scored within 800 characters, a window that checks outgrow as they lengthen, while full-segment verdict use rises.

F.5 Safety instruments, moderators, and scope

The refusal study uses Qwen2.5-1.5B-Instruct student and the Qwen2.5 7B and 32B Instruct teachers. The comparisons vary the generation cap, prompt pool, teacher, and offline versus on-policy training. Each behavioral evaluation samples eight responses per prompt and applies the same refusal judge.

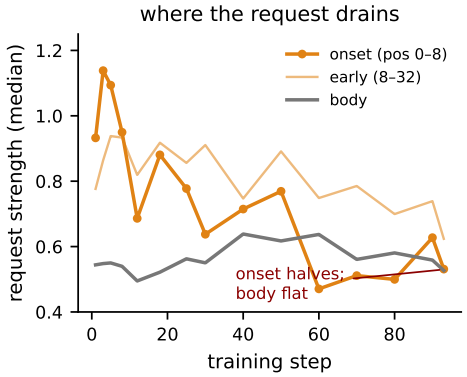


Figure 19: Position-graded absorption in the safety run (32B teacher, 1024 cap). The onset request (positions 0–8) halves over training ($0.93 \to 0.53$, medians) while the body request is flat ($0.54 \to 0.53$): the decision window absorbs, the bulk does not, and a pooled scalar, dominated by body mass, reads "nothing happening" throughout.

Generation cap, prompt pool, and teacher comparisons. Criterion values at endpoint checkpoints: $c = +0.03$ at the 1024-token cap against -0.37 at 4096 (cap effect); $c = +0.19$ with benign-contrast prompts against -0.37 without, at matched cap (pool effect); $c = +0.05$ for the near 7B teacher against -0.37 for the far 32B at matched cap and pool, with onset-field drain 73% against 40% (teacher-distance effect). The offline forward-KL control at matched prompts, steps, and learning rate: $c = -0.13$, AUC .862 (baseline .860). Changing the cap changes the relative weight of response-onset positions in the source, while benign contrast prompts add supervision absent from a harmful-only pool. These are possible contributors to the criterion differences; isolating their causal effects remains open. The distinction between matched and unequal source weights is made explicit in Proposition F.5. The teacher comparison is discussed in Section F.3. The offline control shows no resolved discrimination gain at this dose; on its own, it leaves open whether feature learning or state coverage is the bottleneck. For the 1024-cap arm, the approximately flat pooled training signal decomposes into routing $+0.17$ and fixed-prefix drift -0.18; the two contributions largely cancel.

The per-window twin accounting (safety). The instrument behind Figure 16. Twin pairs are XSTest's own contrast design: an unsafe prompt of type contrast-T paired with the safe prompt of type T sharing its focus word (124 pairs). For the teacher and each checkpoint, one forward pass per chat-templated prompt yields the position-zero log-softmax on a fixed per-prompt support (union of teacher and base top-64 tokens) and the final-layer state at the last prompt position. The requested direction per prompt is the normalized teacher-minus-base log-probability gap on that support; a window's broadcast movement is the inner product of the checkpoint-pair log-probability change with it. Requested magnitudes and broadcast movements are split into pair-symmetric (average) and pair-antisymmetric (conditional) parts. The noise floor for the conditional component is the same statistic under 200 random re-pairings of unsafe members; a signed variant aligns each pair's conditional broadcast with the sign of its conditional request before averaging. The logit-level refusal score is the log-mass difference between a refusal-family and a comply-family token list (fixed token-family lists), and its AUC over twins is Mann–Whitney. The feature probe is the within-pair cosine of final-layer states against the mean cross-pair cosine.

Student	Offline-control steps	Steps in the OPD arm	Sequence presentations per arm
Qwen 1.5B	200 offline	200 OPD	51,200
Qwen 7B	100 offline	100 OPD	25,600
Llama 8B	200 offline	200 OPD	51,200

Table 11: Comparison budgets from the same prepared checkpoint within each model pair, after 512 OMI2 preparation steps. Sequence presentations include repetitions.

G PREPARATION: PROTOCOLS AND ADDITIONAL MEASUREMENTS

G.1 THREE-FAMILY AIME COMPARISON

Students and teachers. The pairs are Qwen2.5-1.5B-Instruct / JustRL-1.5B, Qwen2.5-7B-Instruct / AceReason-Nemotron-7B, and Llama-3.1-8B-Instruct / Llama-3.1-Nemotron-Nano-8B-v1. Each pair shares a tokenizer vocabulary. The fixed teacher supplies the preparation text, the offline-control targets, and the on-policy log-probabilities. The Llama teacher uses its reasoning mode with the system prompt "detailed thinking on."

Stage 1: offline preparation on OMI2. Each teacher generates solutions to problems drawn from OpenMathInstruct-2 (Toshniwal et al., 2024), at temperature one with an 8,192-token response cap. The student uses token-level forward KL on the teacher's recorded top-16 next-token support. The plotted cold start uses 512 optimizer steps of 256 sequences, or 131,072 sequence presentations. The Qwen 1.5B, 7B and Llama corpora each contain 32,768 teacher solutions, so this is four corpus sweeps.

The offline settings are AdamW, learning rate 10^{-6}, momentum coefficients $(0.9, 0.95)$, no weight decay, gradient clipping at 1.0. The cold-start sequence limit is 8,192 tokens.

Stage 2: DAPO comparisons. The prompt pool is DAPO-Math-17k (Yu et al., 2025), recorded as 17,398 unique problems. Training uses consecutive blocks of 6,400 rows in the released ordering. Each block contains 6,235–6,318 unique problems.

OPD samples four student responses for each of 64 prompts per step, giving 256 responses and one optimizer step per update. Sampling is at temperature one, capped at 8,192 tokens for Qwen and 14,336 for Llama.

The offline continuation instead uses four teacher-generated solutions for each prompt row in the same blocks: 25,600 sequences per 6,400-row block. It uses the teacher forward-KL objective and optimizer settings as preparation, at 256 sequences per step. The sequence limits are 8,192 for Qwen and 15,360 for Llama, the latter retaining the reported teacher solutions generated with a 14,336-token response cap. Thus the comparison matches the implemented offline and on-policy recipes in prompt blocks, sequence counts, and total optimizer steps.

Evaluation and sampled coverage. The evaluation pools AIME 2024 and 2025, 30 problems each. Each checkpoint produces 64 responses per problem (two unseeded batches of 32 with identical settings; for three of the 24 checkpoint–year cells the second batch was generated on a different machine with the same driver and library versions) using vLLM, temperature 0.6, top-$p = 0.95$, and a 16,384-new-token cap. Lenient extraction uses a boxed answer when present, otherwise an explicit answer marker, otherwise the last line. The strict MATH and in-loop commitment readouts in Section 3.3 use their own scoring rules.

Following Chen et al. (2021), if c_i of the 64 responses solve problem i, we estimate

$$\widehat{\text{pass@}k} = \frac{1}{60}\sum_{i=1}^{60}\left[1 - \frac{\binom{64-c_i}{k}}{\binom{64}{k}}\right], \qquad 1 \le k \le 64, \tag{31}$$

with a zero numerator when $64 - c_i < k$. In particular, pass@1 is the average observed success rate, and pass@64 is the fraction solved at least once in the sample. A problem never solved in 64 attempts can still lie within the model's support, and one solved on all 64 attempts can still have success probability below one.

Student	Arm	pass@1	pass@64	Coverage	$\geq 1/2$	All 64
Qwen 1.5B	Base	0.0117	0.1500	9	0	0
	Cold start	0.0094	0.1833	11	0	0
	+ offline	0.0076	0.1333	8	0	0
	+ OPD	0.0255	0.2167	13	0	0
Qwen 7B	Base	0.0964	0.3333	20	4	0
	Cold start	0.1307	0.4333	26	7	0
	+ offline	0.1906	0.4333	26	10	1
	+ OPD	0.2711	0.6167	37	15	5
Llama 8B	Base	0.0219	0.1833	11	1	0
	Cold start	0.0339	0.2667	16	0	0
	+ offline	0.0398	0.2833	17	0	0
	+ OPD	0.0714	0.4000	24	1	0

Table 12: Pooled AIME endpoints and observed reliability at 64 samples per problem. Coverage, at-least-half, and all-64 columns count problems out of 60. Both Llama comparison arms start from the same prepared checkpoint and use the budgets in Table 11.

Student	Year	Δpass@1 [95% interval]	Δpass@64 [95% interval]
Qwen 1.5B	2024	$+.028$ $[+.004, +.055]$	$+.200$ $[+.067, +.367]$
	2025	$+.008$ $[-.005, +.029]$	$-.033$ $[-.133, +.067]$
Qwen 7B	2024	$+.092$ $[+.052, +.135]$	$+.133$ $[+.033, +.267]$
	2025	$+.069$ $[+.033, +.114]$	$+.233$ $[+.067, +.400]$
Llama 8B	2024	$+.040$ $[+.006, +.081]$	$+.133$ $[+.033, +.267]$
	2025	$+.023$ $[-.004, +.054]$	$+.100$ $[+.000, +.200]$

Table 13: OPD minus offline control at 64 samples per problem, with paired problem-bootstrap intervals. Values are in probability units.

The lower panels plot the empirical survival curve $60^{-1}\sum_i \mathbf{1}\{c_i/64 \geq r\}$ for $r > 0$. Its limit as $r \downarrow 0$ is observed coverage and its area on $[0, 1]$ is the average solve rate. At exactly $r = 0$, the literal weak-inequality expression equals one; the plotted left endpoint represents the positive-threshold limit. These are marginal distributions across problems.

Endpoints and paired problem changes. Table 12 gives the pooled pass@k endpoints and the number of problems at each empirical reliability threshold.

Among problems solved at least once by both the OPD and offline endpoints, mean observed success rates are $0.159/0.062$ for Qwen 1.5B (6 problems), $0.610/0.455$ for Qwen 7B (25), and $0.222/0.141$ for Llama (17). These intersections are selected using both arms' samples, so their means are descriptive. Against the offline endpoint, OPD gains coverage on 7, 12, and 7 problems and loses it on 2, 1, and 0, respectively. The reported absence of downward crossings of the 50% threshold is compatible with such losses and with decreased solve rates. We therefore describe aggregate coverage and reliability gains.

Uncertainty and differences between AIME years. The figure gives point estimates from 60 problems. Paired problem-bootstrap intervals for each 30-problem year are given in Table 13; they quantify evaluation-problem variability with each arm's trained checkpoint fixed. Resolution per year, at $k = 1$ and $k = 64$, is read from Table 13, where some 2025 intervals include zero; other k have point estimates only.

The Llama base has observed pass@64 0.300 on AIME 2024 but 0.067 on AIME 2025. Its OPD endpoint is 0.333 and 0.467, respectively. Thus almost all of the pooled coverage gain over the base comes from AIME 2025 ($+12$ problems, against $+1$ on AIME 2024). The cause of the year asymmetry, benchmark contamination or otherwise, remains unidentified. Cold start itself raises the pooled pass@1 for Qwen 7B and Llama, while Qwen 1.5B changes little.

Finite-budget interpretation. The two training stages use finite prompt pools, fixed sampled corpora, and a limited number of updates; repeated exposure to the same offline corpus leaves its unique-example count fixed. The comparison uses one corpus size per family, fixed cold-start reference checkpoints, and offline controls run for the recorded budgets. The supported conclusion is that the stated preparation-plus-OPD recipes outperform continued offline distillation at the recorded sequence/update budgets, with positive pooled point estimates across the three pairs.

G.2 Preparation-dose study

We follow the same preparation (cold start) procedure as before, except we vary the number of passes over the cold-start corpus and study how well OPD performs after that. We present the results in Table 14.

Student	Passes	Cold start	After 100 OPD steps
Qwen 1.5B	1	0.459 [0.411, 0.506]	0.459 [0.409, 0.509]
	2	0.457 [0.409, 0.504]	0.530 [0.476, 0.585]
	4	0.471 [0.420, 0.522]	0.562 [0.510, 0.614]
Qwen 7B	1	0.683 [0.638, 0.725]	0.823 [0.782, 0.862]
	2	0.693 [0.649, 0.736]	0.828 [0.791, 0.864]
	4	0.711 [0.669, 0.753]	0.842 [0.805, 0.880]

Table 14: The dose panels of Figure 6: MATH-500 pass@1 (first 250 problems, 4 samples) of the cold-start checkpoint and of the checkpoint after 100 OPD steps from it, at 0, 1, 2 and 4 passes over the preparation corpus.