SELECTION–DEVELOPMENT LOCK-IN: ACCESS STARVATION IN ON-POLICY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy learning changes not only the current policy but also the distribution of data available for future learning, a feedback particularly relevant to modern generative-model fine-tuning. A behavioral region may initially perform poorly yet improve with continued exploration, while policy optimization can reduce its future sampling opportunities to negligible levels before its competence develops sufficiently. We call this selection–development lock-in (SDL). In a canonical environment, we formalize and analyze SDL and show that conditional development requires only polynomial effective learning exposure, yet acquiring that exposure through on-policy sampling takes exponential time as initial competence decreases. We extend the analysis to general on-policy dynamics through two quantities: the support lost per unit of conditional improvement and the dependence of development on sampling access. Under explicit regularity conditions, sufficiently steep support loss combined with access-dependent development produces an exponential barrier. We then conduct controlled shared-parameter neural experiments that reproduce the predicted support valley and delayed development beyond the exactly solvable model. These results identify how locally reward-improving selection can make attainable future improvement prohibitively costly within a training budget.

1 INTRODUCTION

On-policy optimization is widely used for reinforcement learning with verifiable rewards (RLVR) in large language models and for reward-based fine-tuning of text-to-image generators (Shao et al., 2024; Guo et al., 2025; Black et al., 2024; Liu et al., 2025). In these methods, each update depends on samples from the current policy and therefore changes both model behavior and the distribution of future training data. These applications motivate a general question: can locally reward-improving selection remove the experience needed for future learning? Behaviors that lose probability become less available to subsequent learning, coupling policy improvement to the supply of training experience.

Existing research has examined several consequences of this loop. High-reward but low-probability trajectories can be overlooked by finite rollout groups or receive less reinforcement than already common successful trajectories (He et al., 2025; Plyusov et al., 2026). Other studies investigate diversity and mode collapse, including cases where improved single-sample accuracy accompanies worse success under repeated sampling, as measured by pass@k (Yue et al., 2025; Li et al., 2026a).

Unlike prior analyses that primarily study the disappearance of already useful behaviors, our work concerns a different failure mode: a currently weak behavioral region whose conditional competence can improve with sufficient training is deprived of the access required for that improvement. A further question is therefore how changes in the sampling distribution affect what a learner can improve next.

Most descriptions of support contraction do not separately track entry frequency and conditional competence. This distinction matters when competence can develop through learning: a currently weak region may become a better alternative if it continues receiving useful updates. Losing probability mass therefore removes not only behavioral diversity, but also the experience needed for future improvement. This possibility exposes a tension in on-policy learning: selection can improve current reward by reducing the probability of a weaker region, while simultaneously removing the experience that would help that region improve. We call the resulting feedback *selection–development lock-in* (SDL). Its developmental cost arises when a region loses access before realizing an attainable advantage. Although recovery may remain mathematically possible, we refer to this outcome as

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.

lock-in when recovery requires more than the available training budget. Our central question is whether locally reward-improving selection can make such future improvement prohibitively costly.

We first answer this question in a canonical finite-group system. Starting from low conditional competence, useful learning requires only polynomial effective exposure, but the total on-policy sampling time needed to acquire that exposure is exponential. The region initially loses support faster than it develops, which makes later useful updates increasingly expensive. We then extend the analysis to general on-policy learners. Two scaling properties determine the result: how sharply the support cost of conditional improvement grows at low competence, and how strongly useful development depends on the remaining access. Under the stated power-law and regularity assumptions, an exponential barrier arises when this support cost grows faster than inverse competence and development remains positively dependent on access. Sustained learning that does not depend on natural access can avoid the delay.

The contributions of this work are threefold: (1) we identify selection–development lock-in as a distinct access-starvation mechanism in on-policy learning, where currently weak but improvable regions lose future learning opportunities; (2) we theoretically separate the amount of effective learning exposure from the on-policy sampling time required to obtain that exposure, revealing when an exponential access barrier emerges; (3) we provide controlled shared-neural experiments that demonstrate the predicted support valley and delayed development beyond the exactly solvable model. The experiments are designed to isolate the mechanism under shared parameters rather than to estimate its prevalence in any particular large-scale application.

2 Related Work

On-policy optimization and support loss. On-policy optimization changes future training distributions because each update depends on samples generated from the current policy (Schulman et al., 2017; Shao et al., 2024). Recent studies investigate support contraction, verifier-induced reshaping, and collapse under expected return or regularized objectives (GX-Chen et al., 2026; Wei et al., 2026; Sinha et al., 2026). These analyses show that probability mass can disappear during optimization, but they generally do not distinguish the sampling access to a behavioral region from the conditional competence within that region. SDL studies this distinction by asking how losing access affects a region that can still improve through future learning.

RL for language models. Finite rollout groups and probability-weighted reinforcement can restrict learning from rare alternatives in RL for language models (He et al., 2025; Plyusov et al., 2026; Huang et al., 2026; Li et al., 2026b). Related studies analyze pass@k degradation, reasoning-boundary contraction, and diversity loss during RLVR training (Yue et al., 2025; Nguyen et al., 2025; Zhou, 2026; Li et al., 2026c). Much of this literature focuses on behaviors that already contain useful signal, such as correct trajectories or rare behaviors that should be preserved. In contrast, SDL focuses on a region that is initially weak but can become valuable through future conditional learning, and studies how sampling access and conditional competence evolve together.

RL for text-to-image generation. Reward-based fine-tuning has extended policy optimization to diffusion and flow-based generative models (Black et al., 2024; Fan et al., 2023; Liu et al., 2025; Xue et al., 2025). Existing studies investigate reward improvement, reward hacking, and diversity preservation during generative reinforcement learning. These directions motivate studying how policy-induced changes in sampling distributions affect future learning opportunities.

Related theoretical barriers. Policy-gradient analyses study convergence rates and optimization barriers under different assumptions (Mei et al., 2020; Agarwal et al., 2021; Li et al., 2023). Exploration and coverage theory studies how learning algorithms acquire informative experiences (Cai et al., 2020; Jin et al., 2020; Xie et al., 2023; Foster et al., 2022). Rising-bandit formulations also consider alternatives that improve through selection (Metelli et al., 2022). However, these analyses do not separate the amount of effective learning experience required for improvement from the time needed for an evolving policy to obtain that experience. SDL builds on these perspectives by showing how support loss and access-dependent development can jointly create an exponential sampling barrier despite polynomial effective exposure.

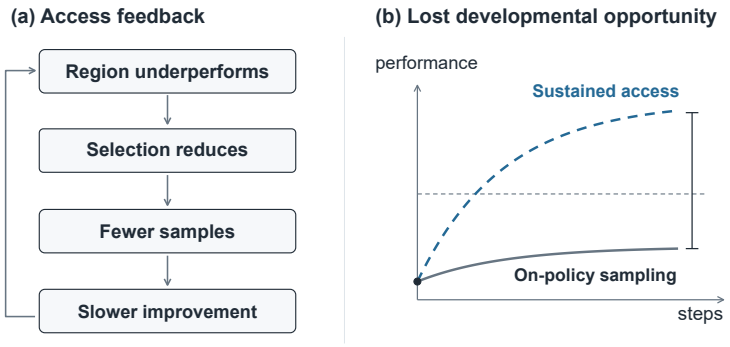


Figure 1: **Selection–development lock-in mechanism.** (a) Selection reduces access probability p for an initially weak region X, slowing conditional improvement in q and reinforcing support loss. (b) Sustained access (blue dashed) allows X to surpass a competitor, whereas on-policy sampling (gray solid) delays development.

3 Problem Formulation

3.1 Fixed regions and conditional success

Let $\mathcal{Z}$ be the space of complete trajectories or generated outputs. Fix a behavioral region $X \subseteq \mathcal{Z}$ and a successful subset $\mathcal{G}_X \subseteq X$ before examining training. Define

$$p_\theta = \mathbb{P}_\theta(Z \in X), \qquad L_\theta = \operatorname{logit} p_\theta, \qquad q_\theta = \mathbb{P}_\theta(Z \in \mathcal{G}_X \mid Z \in X), \tag{1}$$

with $0 < p_\theta < 1$. Selection changes the probability of entering X and development changes the probability of success conditional on entering it. The region's *successful natural support* is

$$s_\theta = \mathbb{P}_\theta(Z \in \mathcal{G}_X) = p_\theta q_\theta. \tag{2}$$

For a continuous reward model or verifier R, a fixed threshold $r_\star$ defines success as $\mathcal{G}_X = \{Z \in X : R(Z) \geq r_\star\}$.

3.2 SDL feedback and developmental opportunity loss

Definition 1 (Selection–development lock-in)**.** ***(a) Feedback mechanism.*** *Selection favors currently better-performing behavior, increasing its relative access to subsequent samples and updates, while access-dependent learning can then sustain or amplify these advantages, reinforcing selection.* ***(b) Developmental consequence.*** *When a suppressed region could improve enough through further learning to become a better alternative, the feedback can deprive it of the updates needed to realize that advantage within a finite training budget. We call this consequence developmental opportunity loss.*

Figure 1 connects the access feedback to the lost opportunity in Definition 1. We quantify this loss by comparing the effective learning exposure required for improvement with the on-policy sampling time needed to obtain it.

3.3 A canonical construction

We use a minimal exactly solvable system to isolate the feedback between selection, lost access, and delayed development, and to expose the quantities needed for the general analysis. The smallest discrete realization of Eq. (1) takes $X = \{X^+, X^-\}$ and $\mathcal{G}_X = \{X^+\}$. It separates entry into a developing region from success within it, with Y as the competing region:

$$\mathbb{P}(X^+) = pq, \qquad \mathbb{P}(X^-) = p(1-q), \qquad \mathbb{P}(Y) = 1 - p, \tag{3}$$

where $p = \sigma(\ell)$ and $q = \sigma(h)$, so $L = \ell$. Rewards are $1, 0$, and $1/2$, respectively. We call $q_m = 1/2$ the *maturity threshold*: conditional reward then matches the competitor, and the theorem below shows that selection reverses there. Write $T_{\text{rev}} = \inf\{t : q(t) \geq q_m\}$ for the time to this point.

Indeed, $J(p, q) = 1/2 + p(q - 1/2)$. Before maturity, both decreasing p and increasing q improve current reward. Nevertheless, the former reduces the chance of obtaining subsequent X-containing training groups. The outcome X^+ has reward 1, above the reward of Y, and the conditional dynamics below reach maturity with finite effective exposure.

For G IID rollouts, let $\hat{A}_i = (r_i - \bar{r})/s_r$, where s_r is the population standard deviation. A zero-variance group receives no update. For exact analysis, we study the one-step on-policy GRPO score update

$$g_\theta = \frac{1}{G} \sum_{i=1}^{G} \widehat{A}_i \nabla_\theta \log \pi_\theta(z_i), \tag{4}$$

following Shao et al. (2024). We use effective learning rates η_s, η_d and write $\rho = \eta_s/\eta_d$.

Write g_ℓ and g_h for the selection and conditional-development components of the group update. Increasing the selection logit raises the probability of both X outcomes together, while increasing the conditional logit shifts probability from X^- to X^+. Because standardized advantages sum to zero, the explicit selection-score dependence on p cancels. In a group containing both X and Y, selection favors X exactly when its successes outnumber its failures. For development, a success has a nonnegative advantage and a positive conditional score, whereas a failure has a nonpositive advantage and a negative conditional score. Both therefore contribute toward increasing q when the group is informative. Appendix A derives the exact group statistics and update formulas.

Proposition 1 (Pathwise development)**.** *Every non-degenerate group containing at least one sample from X satisfies $g_h > 0$.*

In particular, even if a group contains no X^+ samples, it still increases q as long as it contains an X^- sample and induces a nonzero gradient. The negative advantage of X^- multiplies its negative conditional-failure score. The algebra and a corresponding identity for general conditional policies appear in Appendix A.

Theorem 1 (Exact finite-group sign)**.** *For finite $G \geq 2$, $0 < p < 1$, and $0 < q < 1$, the complete expected gradients satisfy*

$$\operatorname{sign} \bar{g}_\ell^{(G)}(p, q) = \operatorname{sign}(q - q_m), \qquad \bar{g}_h^{(G)}(p, q) > 0. \tag{5}$$

Thus support decreases before maturity and increases afterward, while conditional competence continues to develop.

The proof conditions on the number n of X samples. For a group that also contains Y, the selection update is a strictly increasing odd function of the number of successes minus failures. At maturity these counts are symmetric, so expected selection is zero. Increasing conditional competence shifts the count distribution toward successes and makes selection positive. Groups drawn entirely from one region do not change selection. Averaging over compositions gives the theorem (Appendix B).

3.4 Self-induced rarity

Write $\bar{g}_\ell^{(G)}$ and $\bar{g}_h^{(G)}$ for the gradients averaged over all group compositions. Define $\mathsf{S}_G(p) = -\bar{g}_\ell^{(G)}(p, 0)$, the selection pressure when all X outcomes fail, and $\mathsf{D}_G(p) = \partial_q \bar{g}_h^{(G)}(p, q)\Big|_{q=0}$, the initial slope of the expected conditional-logit update. Both are positive for $0 < p < 1$. To obtain them, condition on $N_X \sim \mathrm{Binomial}(G, p)$ and Taylor-expand the finite composition sum at $q = 0$. All-failure groups determine the zeroth-order selection term and together with one-success groups determine the first-order development term. Finally, the chain rule $\dot{q} = \eta_d q(1 - q)\bar{g}_h^{(G)}$ introduces one more factor of q. Thus, as $q \to 0$,

$$-\bar{g}_\ell^{(G)}(p, q) \sim \mathsf{S}_G(p), \qquad \dot{q} \sim \eta_d \mathsf{D}_G(p) q^2. \tag{6}$$

Here $f \sim g$ means $f/g \to 1$. These asymptotic equivalences are uniform on compact interior support ranges. Appendix C gives the explicit finite sums and error bounds, including the extension to rare support. Selection has a nonzero low-competence limit, while probability-space development is quadratic. Dividing the two drifts removes time:

$$-\frac{d\ell}{dq} \sim \rho R_G(p) q^{-2}, \qquad R_G(p) = \frac{\mathsf{S}_G(p)}{\mathsf{D}_G(p)}. \tag{7}$$

As $p \to 0$, the leading informative groups contain exactly one X sample. Their contributions give $\mathsf{S}_G(p) \sim \sqrt{G-1}\,p$ and $\mathsf{D}_G(p) \sim 2\sqrt{G-1}\,p$, hence $R_G(p) \to 1/2$. The phase cost is therefore asymptotic to $(\rho/2)q^{-2}$ in the joint low-competence, rare-support regime.

Theorem 2 (Support valley and reversal time)**.** *Fix finite $G \geq 2$, $\rho > 0$, and any interior $p_0 \in (0,1)$. As $q_0 \to 0$, the expected-update trajectory, with $\dot{\ell} = \eta_s \bar{g}_\ell$ and $\dot{h} = \eta_d \bar{g}_h$, reaches every fixed lower support threshold while $q/q_0 \to 1$. Its minimum support at maturity and the corresponding reversal time obey*

$$-\log p_{\min} \sim \frac{\rho}{2q_0}, \qquad \log T_{\mathrm{rev}} \sim \frac{\rho}{2q_0}. \tag{8}$$

The proof first follows the trajectory to $p = q_0$, where competence still satisfies $q/q_0 \to 1$. Beyond that point, $R_G(p) \to 1/2$, so integrating the phase cost $(\rho/2)q^{-2}$ gives the leading support-logit loss $\rho/(2q_0)$. Support is then rare, so its logit and logarithm have the same leading behavior. Integrating $dt = dq/\dot{q}$ through this valley introduces the inverse support factor. The deepest part of the valley determines the leading log time. Appendix D supplies the uniform bounds and controls the entry and remaining parts of the trajectory.

To express the cost of acquiring learning opportunities, introduce the effective-exposure clock

$$\tau(t) = \int_0^t p(s)(1-p(s))^{G-1}\,ds, \tag{9}$$

the rare-support dynamics give $dq/d\tau \sim 2\eta_d\sqrt{G-1}\,q^2(1-q)^2$ as $p \to 0$. Integrating its reciprocal from low initial competence to maturity requires exposure of order $1/q_0$. Let $\tau_{\mathrm{rev}} = \tau(T_{\mathrm{rev}})$. The complexity of reaching maturity is therefore

$$\boxed{\begin{aligned} &\text{Effective exposure:} \quad \tau_{\mathrm{rev}} = \Theta(q_0^{-1}), \\ &\text{Natural sampling:} \quad T_{\mathrm{rev}} = \exp\left[\Theta(q_0^{-1})\right]. \end{aligned}} \tag{10}$$

The notation Θ means upper and lower bounds by positive constant multiples of its argument and Eq. (8) retains the sharper coefficient $\rho/2$ in log time. The natural sampling cost is exponential even though conditional development requires only polynomial exposure. In particular, counting rollouts multiplies the sampling time by G and preserves its order.

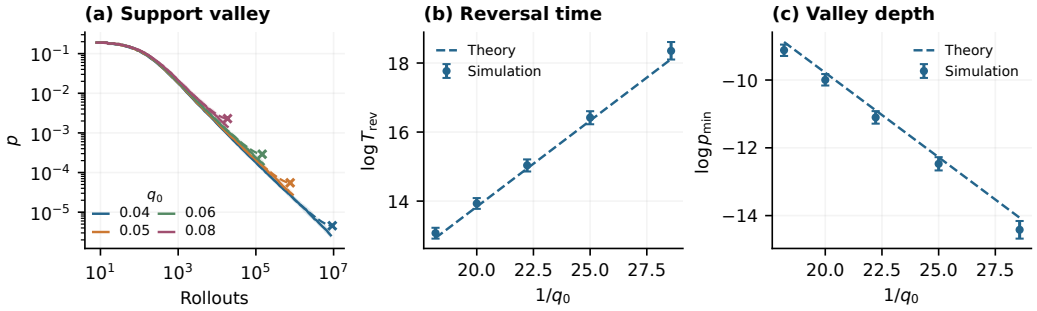


Figure 2: **Canonical support valley and exponential complexity.** Left: support trajectories for $G = 8, p_0 = 0.2$. Solid trajectories summarize simulations and crosses mark the theoretical maturity point. Middle and right: reversal time and valley depth for $G = 16, p_0 = 0.2$. Markers summarize simulations with error bars showing 95% bootstrap intervals, and dashed curves show theoretical values calculated from the finite-group dynamics. Time counts rollouts.

We test these predictions with exact finite-group simulations. Figure 2 shows the support valley and the predicted linear relations of $\log T_{\mathrm{rev}}$ and $\log p_{\min}$ to $1/q_0$, with leading slopes $\rho/2 = 0.5$ and $-\rho/2 = -0.5$. The full study contains 48 conditions and 9,600 trajectories. Additional group-size and initial-support results, together with the simulation settings, appear in Appendix I. The experimental results for the minimum support and the corresponding reversal time closely match the theoretically predicted exponential relationship, as shown in the middle and right panels of Figure 2.

4 THEORETICAL ANALYSIS

The canonical derivation identifies two quantities that control the barrier: support lost per unit competence gain, and the dependence of development on access. We now formulate them directly for a general learner, using the fixed $X, \mathcal{G}_X, q$ of Section 3.1. Its complete mean update $v(\theta, z)$ induces

$$\dot{L} = \nabla L^\top v(\theta, z), \qquad \dot{q} = \nabla q^\top v(\theta, z). \tag{11}$$

The auxiliary state z collects the behavior and reference policies, optimizer moments, retained data, and training schedule. Clipping, regularization, and repeated optimization enter through v and z (Appendix H). Shared parameters allow updates in other regions to affect competence in X, while replay provides learning opportunities from stored samples.

4.1 THE PHASE EXPONENT δ: FORMATION OF THE SUPPORT VALLEY

On a segment with $\dot{L} < 0$ and $\dot{q} > 0$, define

$$C(L, q, z) = -\frac{dL}{dq} > 0, \qquad \delta_{\text{eff}}(q) = -\frac{d \log C}{d \log q}. \tag{12}$$

Here C is the support-logit cost per unit increase in the fixed conditional-success probability and δ_{eff} is its log–log slope along the trajectory. A common positive rescaling of time cancels from this ratio. Fixing the success event makes the coordinate comparable across checkpoints and methods. Appendix F derives the transformation of phase exponents under coordinate changes and the invariance of the phase integral. Its exact identity is

$$L(q_I) = L(q_0) - \int_{q_0}^{q_I} C(q)\, dq. \tag{13}$$

The additional hypothesis $C(q) \sim \kappa q^{-\delta}$, $\kappa > 0$, specifies a low-competence regime. The exponent characterizes the learner, task, region, and update protocol. Finite leading drift orders provide a sufficient route to this law: if adverse selection is of order q^α and development of order $q^{\beta_{\text{ph}}}$ under a common positive clock, then $\delta = \beta_{\text{ph}} - \alpha$ (Appendix G.1).

We use the access model $\dot{q} = p^\gamma d(q, z)$. Throughout the analysis, $\gamma \geq 0$ is fixed as $q_0 \to 0$. Let T denote time to the specified competence endpoint. When this endpoint is the selection-reversal point, $T = T_{\text{rev}}$.

Proposition 2 (Conditional support and time classification)**.** *Consider a family of trajectories with fixed interior initial support and $q_0 \to 0$. Suppose a segment of decreasing support and increasing competence extends to a fixed $q_I > 0$, the power-law approximation to C holds uniformly near zero, and the remaining phase cost is bounded. Then*

$$\begin{aligned} p(q_I) &= \Theta(1), && \delta < 1, \\ -\log p(q_I) &\sim \kappa \log(1/q_0), && \delta = 1, \\ -\log p(q_I) &\sim \frac{\kappa}{\delta - 1} q_0^{-(\delta-1)}, && \delta > 1. \end{aligned} \tag{14}$$

If, additionally, $\dot{q} = p^\gamma d(q, z)$ with $\gamma > 0$ and $d, 1/d$ uniformly bounded by powers of $1/q$, time is at most polynomial for $\delta \leq 1$, while for $\delta > 1$

$$\log T \sim \frac{\gamma \kappa}{\delta - 1} q_0^{-(\delta-1)}. \tag{15}$$

To see the classification, integrate $C(q) \sim \kappa q^{-\delta}$ in Eq. (13): the accumulated logit loss stays bounded for $\delta < 1$, grows logarithmically for $\delta = 1$, and grows as a power of $1/q_0$ for $\delta > 1$. Next write $T = \int_{q_0}^{q_I} p(q)^{-\gamma} d(q, z)^{-1}\, dq$. In the last case, any fixed competence interval away from zero has already accumulated the leading support loss, so the inverse-support factor determines log time. The bounded polynomial factors do not change its leading coefficient. Appendix G makes these upper and lower bounds explicit. Thus, this class of access-dependent on-policy dynamics exhibits an exponential access barrier when $\delta > 1$ and $\gamma > 0$ under the proposition's assumptions. We call this a *supercritical phase barrier*. Its time scale is exponential in $q_0^{-(\delta-1)}$. Proposition 7 in Appendix G.4 extends this growth order to two-sided power bounds. The canonical exposure comparison in Eq. (10) exhibits the resulting developmental cost.

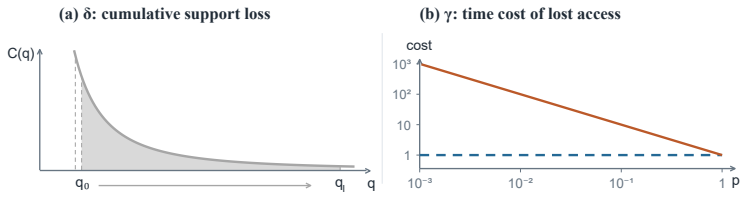


Figure 3: **How δ and γ produce an access barrier.** (a) $C(q) = -dL/dq$ is support-logit loss per unit competence gain. Consider variants of the canonical learner that retain its selection update and the same frequency of learning opportunities in X. With outcome-only feedback, failures provide weak conditional updates and successes are rare, giving $\delta = 2$. If a teacher instead supplies a correct X answer at each opportunity and the conditional policy learns from its log likelihood, learning no longer has to wait for self-generated success, reducing δ to 1. Conversely, suppose each opportunity draws two independent candidates from the conditional policy within X, and conditional training on their log likelihoods occurs only if both succeed. This extra acceptance requirement makes useful updates rarer at low competence, increasing δ to 3. The shaded integral is $L(q_0) - L(q_I)$ (shown: $\delta = 2$). (b) Under $\dot{q} = p^{\gamma} d$, the time cost of a fixed small competence gain scales as $p^{-\gamma}$. If useful learning requires a group containing X, with a nonvanishing conditional contribution at rare support, then $D \propto p$ and $\gamma_{\text{eff}} \to 1$. Transfer or replay can sustain development despite falling natural access, and if D has a positive, smooth limit as $p \to 0$, then $\gamma_{\text{eff}} \to 0$. Orange solid: $\gamma = 1$; blue dashed: $\gamma = 0$.

Successful support can fall while competence rises. For $s = pq$, the phase identity gives

$$\frac{d \log s}{dq} = \frac{1}{q} - (1-p)C(q). \tag{16}$$

Thus successful natural support decreases whenever $(1-p)qC(q) > 1$, even though conditional success is improving. This holds in the rare-support, sufficiently low-competence part of a supercritical phase. At any fixed endpoint q_I, $\log s(q_I) = \log p(q_I) + \log q_I$, so the leading support barrier is also present in the actual probability of successful trajectories.

4.2 The access exponent γ: starvation of development

The local support elasticity is

$$\gamma_{\text{eff}}(p, q, z) = \left. \frac{\partial \log D(p, q, z)}{\partial \log p} \right|_{q,z}, \qquad D = \dot{q} > 0. \tag{17}$$

For G IID trajectories from a fixed distribution,

$$\mathbb{P}(N_X \geq 1) = 1 - (1-p)^G \sim Gp \quad (p \to 0). \tag{18}$$

Direct opportunities to learn from X therefore vanish linearly at rare support. Equation (17) measures how the total development rate responds to access at fixed competence and auxiliary state. A uniformly positive elasticity makes development decline with support. If development has a positive, smooth limit as $p \to 0$, then $\gamma_{\text{eff}} \to 0$.

The condition for an exponential access barrier. Table 1 summarizes the classification under the trajectory, uniform power-law, and polynomial-prefactor assumptions of Proposition 2.

Table 1: **Support loss and the cost of development.** The phase exponent δ determines the support valley while the access exponent γ determines whether that valley slows learning. Orders refer to $q_0 \to 0$ at a fixed competence endpoint, under the stated uniform assumptions.

Phase	Access	Support at endpoint	Time to endpoint
$\delta < 1$	$\gamma \geq 0$	Bounded away from zero	At most polynomial
$\delta = 1$	$\gamma \geq 0$	Polynomially small	At most polynomial
$\delta > 1$	$\gamma > 0$	Exponentially small	Exponential
$\delta > 1$	$\gamma = 0$	Exponentially small	At most polynomial

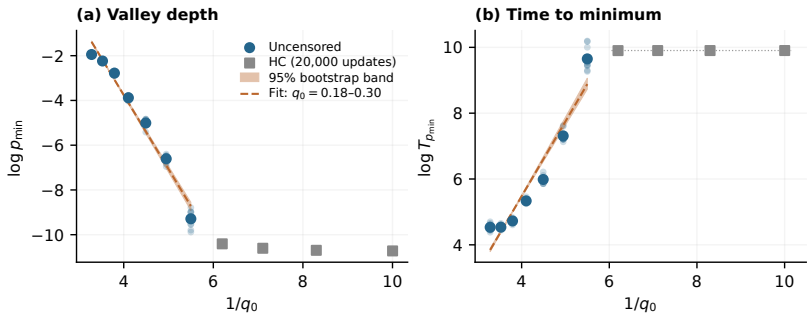


Figure 4: **Neural realization of the SDL support valley.** Blue circles: $q_0 = 0.18$–0.30. Gray squares: *horizon-censored* (HC) conditions, $q_0 = 0.10$–0.16, whose support is still declining near 20,000 updates. Large markers show condition means in log space; faint markers show ten individual seeds. Dashed lines fit all seven uncensored conditions. Their slopes are -3.311 (95% CI: $[-3.408, -3.221]$) for depth and 2.274 (95% CI: $[2.181, 2.375]$) for time. Shading shows pointwise 95% confidence bands from 5,000 paired-seed bootstrap resamples.

In particular, the exponential-access condition and its complexity are

$$\boxed{\delta > 1 \text{ and } \gamma > 0 \quad \Longrightarrow \quad T = \exp\left[\Theta\left(q_0^{-(\delta-1)}\right)\right].} \tag{19}$$

This states the growth order and Eq. (15) gives the precise leading coefficient in $\log T$. For $\gamma = 0$, the same polynomial bounds on d and $1/d$ imply at most polynomial time even inside an exponentially deep valley. The canonical environment has $\delta = 2$, $\gamma = 1$, and $\kappa = \rho/2$, recovering $\log T_{\text{rev}} \sim \rho/(2q_0)$.

Figure 3 shows how δ controls the support lost per unit conditional improvement at low competence and how γ controls the slowdown at matched competence and auxiliary state. In the canonical environment, selection remains finite while development is quadratic, yielding $\delta = 2$, while informative groups contain one X trajectory at leading order, so their probability is $O(p)$ and hence $D \propto p$, yielding $\gamma = 1$.

5 Experiments

To isolate the mechanism from architecture- and task-specific effects, we use a shared neural policy on MNIST (LeCun et al., 1998) that selects digit actions $X(o)$ and a fallback Y, rewarded 1, 0, and 0.5 for a correct digit, an incorrect digit, and fallback, respectively. This controlled shared-parameter setting tests whether SDL persists beyond the exactly solvable construction, rather than estimating its prevalence in any particular large-scale application.

Let ν denote the fixed uniform distribution over the evaluation set. For each context o, let p_o denote the probability of selecting a digit action and q_o the probability of selecting the correct digit conditional on selecting a digit action:

$$\begin{aligned} p_o &= \sum_{k \in X(o)} \pi_\theta(k \mid o), \quad q_o = \frac{\sum_{k \in G_X(o)} \pi_\theta(k \mid o)}{p_o}, \\ p &= \mathbb{E}_{o\sim\nu}[p_o], \qquad\qquad q = \mathbb{E}_{o\sim\nu}[q_o]. \end{aligned} \tag{20}$$

Here $G_X(o) \subseteq X(o)$ contains the correct digit action for context o. The successful natural support is $\mathbb{E}_{o\sim\nu}[p_o q_o]$. We vary $q_0 = 0.10$–0.30 with ten seeds per setting and $p_0 \approx 0.8$. Appendix J specifies the architecture and training protocol.

Figure 4 reports the observed support depths and times to the recorded minimum. Detailed competence changes are provided in Appendix J.2. At $q_0 = 0.20$, mean support falls to 1.37×10^{-3} while mean competence increases by 0.0187, illustrating support loss during conditional improvement.

The fits over all seven uncensored conditions, $q_0 = 0.18$–0.30, give $R^2 = 0.976$ for depth and 0.917 for time. Coefficients and 95% bootstrap intervals are in Appendix J.3. They characterize the resolved range $q_0 = 0.18$–0.30. Conditions $q_0 = 0.10$–0.16 remain HC and are excluded from both fits.

Forced-access control. At the $q_0 = 0.20$ initialization, sampling only digit actions lets all ten seeds reach $q \geq 0.8$ in 2,400–2,450 updates. This confirms that the conditional task can develop well beyond the fallback reward. With fixed random targets, support still contracts to 1.96×10^{-5}, while held-out competence changes by only 1.40×10^{-4}. The forced-access control demonstrates further learnability of the suppressed region, while the random-target control shows support contraction with little held-out development.

5.1 Ablations

We vary competitor reward, group size, and initial support separately at target $q_0 = 0.20$. The common baseline uses $p_0 = 0.8$, fallback reward $b = 0.5$, and $G = 8$. Table 2 reports arithmetic means and sample standard deviations across ten seeds, using the same convention as Table 4.

Table 2: **One-factor neural ablations.** Arithmetic mean $\pm$ sample standard deviation over ten seeds. Times count updates. HC marks a minimum unresolved within 20,000 updates.

Condition	$p_{\min}$	$T_{p_{\min}}$
Baseline	$(1.37 \pm 0.23) \times 10^{-3}$	$1,515 \pm 284$
$b = 0.3$	$(6.21 \pm 1.29) \times 10^{-3}$	475 ± 54
$b = 0.7$	$(3.36 \pm 0.92) \times 10^{-4}$	$5,080 \pm 1,668$
$G = 4$	$(1.36 \pm 0.27) \times 10^{-3}$	$2,115 \pm 467$
$G = 16$	$(1.34 \pm 0.35) \times 10^{-3}$	$1,070 \pm 273$
$p_0 = 0.2$ (HC)	$(5.81 \pm 0.47) \times 10^{-5}$	$19,910 \pm 187$
$p_0 = 0.5$	$(1.68 \pm 0.36) \times 10^{-4}$	$10,355 \pm 2,388$

Competitor reward: a higher fallback reward makes abstention more attractive while digit competence is still low. The resulting deeper and later support minimum is consistent with stronger selection against the developing region, reducing the subsequent supply of digit samples. Competence at the minimum also decreases with fallback reward (Appendix J.5), so the longer delay does not translate into greater conditional development.

Group size: larger groups reach the minimum in fewer updates but leave its depth nearly unchanged. Increasing G raises the chance of including a digit action at rare support, yet also spends more rollouts per update. With 256 contexts per update, the geometric-mean rollout costs for $G = 4, 8, 16$ are 2.12, 3.06, and 4.26 million actions. Thus faster progress in update count does not imply better sampling efficiency or removal of the support bottleneck.

Initial support: lower p_0 restricts access from the outset and produces a deeper, later observed valley. The geometric-mean fractions of initial support retained at $p_0 = 0.2, 0.5, 0.8$ are approximately 0.029%, 0.033%, and 0.169%. Hence the effect is more than a proportional rescaling of initial access; even an initially common digit region can lose most of its support. All ten $p_0 = 0.2$ runs are HC, so their observed minima are upper bounds on the full-trajectory minima and do not establish recovery. Appendix J.5 provides the remaining ablation statistics and uncertainty estimates.

6 Conclusion

SDL links selection-driven learning access to the loss of attainable development. The exponential laws and their empirical patterns establish the access-cost component. This connection between the price of learning opportunities and what those opportunities could achieve gives the feedback its optimization significance. More generally, the access barrier is characterized by two properties of the complete learner: the support lost per unit of conditional improvement and the dependence of development on sampling access.

Limitations and future work. The present study is designed to establish and isolate SDL analytically and in controlled shared-neural experiments; it does not aim to quantify its prevalence in large-scale generative systems. Extending the analysis to LLM and text-to-image training, and measuring the resulting practical performance cost, remain important directions for future work.

AI USE STATEMENT

Generative AI was used to aid and polish writing, support retrieval and discovery of relevant literature, draft sections of the paper, and assist in proving mathematical claims. It also assisted with translation and reconciliation of experimental summaries. The authors are responsible for verifying the derivations, experimental results, citations, and final content, and take responsibility for the submission.

ETHICS STATEMENT

We use the publicly available MNIST dataset for research in accordance with its terms of use. This work involves no collection of personal data.

REPRODUCIBILITY STATEMENT

The assumptions and definitions underlying the theoretical results are stated in Sections 3 and 4, with detailed derivations and proofs in Appendices A–H. Appendix I specifies the canonical simulator, parameter grids, random-seed construction, computational clocks, and uncertainty estimation. Appendix J documents the MNIST data splits, model architecture, initialization and calibration procedures, optimization settings, random seeds, evaluation schedules, and training horizons.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. URL http://jmlr.org/papers/v22/19-736.html.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably Efficient Exploration in Policy Optimization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1283–1294. PMLR, 13–18 Jul 2020.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885. Curran Associates, Inc., 2023. doi: 10.52202/075280-3497.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline Reinforcement Learning: Fundamental Barriers for Value Function Approximation. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3489–3489. PMLR, 02–05 Jul 2022.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Anthony GX-Chen, Jatin Prakash, Jeff Guo, Rob Fergus, and Rajesh Ranganath. KL-Regularized Reinforcement Learning for Generative Modelling is Designed to Mode Collapse. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 50665–50693, 2026.

Andre Wang He, Daniel Fried, and Sean Welleck. Rewarding the Unlikely: Lifting GRPO Beyond Distribution Sharpening. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25548–25560, Suzhou, China, November 2025. Association for Computational

Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1298. URL https://aclanthology.org/2025.emnlp-main.1298/.

Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. Low-probability Tokens Sustain Exploration in Reinforcement Learning with Verifiable Reward. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24158–24188, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1209. URL https://aclanthology.org/2026.findings-acl.1209/.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-Free Exploration for Reinforcement Learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4870–4879. PMLR, 13–18 Jul 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. *Mathematical Programming*, 201:707–802, 2023. doi: 10.1007/s10107-022-01920-6.

Long Li, Zhijian Zhou, Jiaran Hao, Jason Liu, Yanting Miao, Wei Pang, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, Chao Qu, and Yuan Qi. The Choice of Divergence: A Neglected Key to Mitigating Diversity Collapse in Reinforcement Learning with Verifiable Reward. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 64742–64768, 2026a.

Pengyi Li, Elizaveta Goncharova, Andrey Kuznetsov, and Ivan Oseledets. Back to Basics: Revisiting Exploration in Reinforcement Learning for LLM Reasoning via Generative Probabilities. arXiv preprint arXiv:2602.05281, 2026b.

Xiaozhe Li, Yang Li, Xinyu Fang, Shengyuan Ding, Peiji Li, Yongkang Chen, Yichuan Ma, TianYi Lyu, Linyang Li, Dahua Lin, Qipeng Guo, Qingwen Liu, and Kai Chen. Beyond Mode Collapse: Distribution Matching for Diverse Reasoning. In *Forty-third International Conference on Machine Learning*, 2026c. URL https://openreview.net/forum?id=FEmu30g4ts.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training Flow Matching Models via Online RL. 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. 40783–40818. Curran Associates, Inc., 2025. doi: 10.52202/085713-1362.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the Global Convergence Rates of Softmax Policy Gradient Methods. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 13–18 Jul 2020.

Alberto Maria Metelli, Francesco Trovò, Matteo Pirola, and Marcello Restelli. Stochastic Rising Bandits. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15421–15457. PMLR, 17–23 Jul 2022.

Phuc Minh Nguyen, Chinh D. La, Duy M. H. Nguyen, Nitesh V. Chawla, Binh T. Nguyen, and Khoa D. Doan. The Reasoning Boundary Paradox: How Reinforcement Learning Constrains Language Models. arXiv preprint arXiv:2510.02230, 2025.

Daniil Plyusov, Alexey Gorbatovski, Boris Shaposhnikov, Viacheslav Sinii, Alexey Malakhov, Daria Korotyshova, and Daniil Gavrilov. F-GRPO: Don't Let Your Policy Learn the Obvious and Forget the Rare. arXiv preprint arXiv:2602.06717, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint arXiv:2402.03300, 2024.

Abhijeet Sinha, Sundari Elango, and Dianbo Liu. Expected Return Causes Outcome-Level Mode Collapse in Reinforcement Learning and How to Fix It with Inverse Probability Scaling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Vr0kjNERRf.

Shaohang Wei, Zikun Su, Feifan Song, Wen Luo, Wei Li, Guangyue Peng, and Houfeng Wang. Verifier-Induced Support Reshaping in On-Policy Optimization. arXiv preprint arXiv:2608.00220, 2026.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The Role of Coverage in Online Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on Visual Generation. arXiv preprint arXiv:2505.07818, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? 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. 57654–57689. Curran Associates, Inc., 2025. doi: 10.52202/085713-1933.

Todd Zhou. When RLVR Shrinks the Reasoning Boundary: Diagnosing Pass@k Inversion. arXiv preprint arXiv:2607.20543, 2026.

A Exact Group Statistics and Gradients

Let a group have counts (n_+, n_-, n_Y), where $n = n_+ + n_-$, $n_Y = G - n$, and $d = n_+ - n_-$. With rewards $1, 0, 1/2$, respectively,

$$\bar{r} = \frac{1}{2} + \frac{d}{2G}, \qquad \sum_i (r_i - \bar{r})^2 = \frac{nG - d^2}{4G}, \qquad s_r^2 = \frac{nG - d^2}{4G^2}. \tag{21}$$

The group is active iff $nG - d^2 > 0$. Its standardized X-advantages are

$$A_+ = \frac{G - d}{\sqrt{nG - d^2}}, \qquad A_- = -\frac{G + d}{\sqrt{nG - d^2}}. \tag{22}$$

For either X outcome, $\partial_\ell \log P(X^\pm) = 1 - p$, whereas $\partial_\ell \log P(Y) = -p$. Since standardized advantages sum to zero,

$$\begin{aligned} g_\ell &= \frac{1}{G}\left[(1-p)(n_+ A_+ + n_- A_-) - p n_Y A_Y\right] \\ &= \frac{1}{G}(n_+ A_+ + n_- A_-) = \frac{d(G-n)}{G\sqrt{nG - d^2}}. \end{aligned} \tag{23}$$

For development, the scores are $1 - q$ for X^+, $-q$ for X^-, and zero for Y. Hence

$$g_h = \frac{1}{G}\left[n_+ A_+ (1-q) - n_- A_- q\right] = \frac{n_+(G-d)(1-q) + n_-(G+d)q}{G\sqrt{nG - d^2}}. \tag{24}$$

Every numerator term is nonnegative. If the active group contains X, at least one is strictly positive, proving Proposition 1.

The failure-score identity extends to general conditional policies. Let $\pi_\theta(y \mid X)$ be any conditional policy, let $\mathcal{G}_X \subseteq X$ denote successes, and let $q(\theta) = P_\theta(Z \in \mathcal{G}_X \mid X)$. The score identity and zero-mean conditional score give

$$\nabla q = \mathbb{E}[\mathbf{1}_{\mathcal{G}_X}(Z) \nabla \log \pi_\theta(Z \mid X)], \qquad \mathbb{E}[-\nabla \log \pi_\theta(Z \mid X) \mid Z \notin \mathcal{G}_X] = \frac{\nabla q}{1 - q}. \tag{25}$$

Thus, when within-X failures receive approximately homogeneous negative weights, their expected conditional direction aligns with increasing success probability.

B Proof of the Full-Composition Sign

Condition on $N_X = n$. For $n = 0$ or $n = G$, the selection contribution is zero. For $1 \le n \le G - 1$, let $K = N_+ \sim \mathrm{Binomial}(n, q)$, so $d = 2K - n$. Equation (23) becomes

$$g_\ell = \frac{G - n}{G} \phi_{G,n}(2K - n), \qquad \phi_{G,n}(d) = \frac{d}{\sqrt{nG - d^2}}. \tag{26}$$

On the admissible interval,

$$\phi'_{G,n}(d) = \frac{nG}{(nG - d^2)^{3/2}} > 0. \tag{27}$$

At $q = 1/2$, the distribution of $2K - n$ is symmetric and ϕ is odd, so its expectation is zero. For $q_1 < q_2$, a monotone coupling of binomial variables gives $K(q_1) \le K(q_2)$ almost surely, with strict inequality with positive probability. Strict monotonicity of ϕ makes the conditional expected selection gradient strictly increasing in q. It is therefore negative below $1/2$ and positive above it. Averaging over the positive probabilities of $1 \le N_X \le G - 1$ proves the selection claim in Theorem 1; Eq. (24) proves the development claim.

C Low-Competence Expansion

For fixed finite G, each composition probability is the finite polynomial

$$\Pi(n_+, n_-; p, q) = \frac{G!}{n_+! n_-! n_Y!} (pq)^{n_+} [p(1-q)]^{n_-} (1-p)^{n_Y}. \tag{28}$$

At $q = 0$, every X sample is X^-. If a group contains $n \in \{1, \dots, G-1\}$ such samples, $g_\ell = -\sqrt{n(G-n)}/G$. Define

$$\mathsf{S}_G(p) = \sum_{n=1}^{G-1} \binom{G}{n} p^n (1-p)^{G-n} \frac{\sqrt{n(G-n)}}{G}. \tag{29}$$

Then $\bar{g}_\ell^{(G)}(p, q) = -\mathsf{S}_G(p)[1 + O(q)]$ uniformly for p in a compact subset of $(0, 1)$.

The first-order development coefficient receives one term from differentiating all-failure compositions and another from one-success compositions. Define

$$\mathsf{D}_G(p) = \sum_{n=1}^{G} \binom{G}{n} p^n (1-p)^{G-n} b_{G,n}, \tag{30}$$

$$b_{G,n} = \mathbf{1}_{n<G} \frac{\sqrt{n(G-n)}}{G} + \frac{n(G+n-2)}{G\sqrt{nG - (n-2)^2}}. \tag{31}$$

Then $\bar{g}_h^{(G)}(p, q) = q\mathsf{D}_G(p)[1 + O(q)]$. Since $dq/dh = q(1-q)$,

$$\dot{q} = \eta_d \mathsf{D}_G(p) q^2 [1 + O(q)]. \tag{32}$$

As $p \to 0$, groups with exactly one X dominate:

$$\mathsf{S}_G(p) = \sqrt{G-1}\, p + O(p^2), \qquad \mathsf{D}_G(p) = 2\sqrt{G-1}\, p + O(p^2), \tag{33}$$

so $R_G(p) = \mathsf{S}_G(p)/\mathsf{D}_G(p) = 1/2 + O(p)$. As $p \to 1$, selection requires at least one Y whereas development can occur in all-X groups; direct expansion gives $R_G(p) = 1 - p + O((1-p)^2)$.

D Self-Induced Rarity and Matched Asymptotics

D.1 Fixed-threshold rare entry

On any compact support interval $[p_r, p_0] \subset (0, 1)$, R_G is continuous and bounded between positive constants. Equation (7) therefore supplies two-sided bounds proportional to q^{-2}. Crossing the fixed logit distance $\Delta\ell = \operatorname{logit} p_0 - \operatorname{logit} p_r$ requires

$$\Delta\ell = \Theta\left(\int_{q_0}^{q_e} q^{-2}\, dq \right), \tag{34}$$

which implies

$$q_e - q_0 = \Theta(q_0^2), \qquad q_e/q_0 \to 1. \tag{35}$$

Thus every fixed interior initialization becomes rare before competence changes by a comparable fraction.

D.2 Moving-threshold match

To recover the rare-support coefficient, use the moving threshold $p_e(q_0) = q_0$. The required logit loss is $\log(1/q_0) + O(1)$, so the same bounds imply

$$\frac{1}{q_0} - \frac{1}{q_e} = \Theta\left(\log \frac{1}{q_0} \right), \qquad q_e - q_0 = O\left(q_0^2 \log \frac{1}{q_0} \right). \tag{36}$$

After entry and before reversal, $p(q) \le q_0$. Using Eq. (33),

$$-\frac{d\ell}{dq} = \frac{\rho}{2} q^{-2} [1 + O(p) + O(q)]. \tag{37}$$

Integrating to a fixed small $\bar{q} < 1/2$ produces the leading term $-\rho/(2q_0)$; accumulated approximation error is at most logarithmic. The remainder from $\bar{q}$ to $1/2$ is $O(1)$. More precisely,

$$\log p_{\min} = -\frac{\rho}{2q_0} + O\left(\log \frac{1}{q_0}\right),$$

which gives the support equivalent in Theorem 2.

In the rare phase, the leading dynamics are

$$\dot{\ell} = 2\eta_s\sqrt{G-1}\, p(1-p)^{G-1}(q-1/2), \tag{38}$$

$$\dot{q} = 2\eta_d\sqrt{G-1}\, p(1-p)^{G-1}q^2(1-q)^2. \tag{39}$$

Eliminating time yields

$$\frac{d\ell}{dq} = \rho\frac{q-1/2}{q^2(1-q)^2}, \qquad \ell(q) = \ell_e + \rho\left[\frac{1}{2q(1-q)} - \frac{1}{2q_e(1-q_e)}\right]. \tag{40}$$

The time integral is

$$T_{\mathrm{rev}} = T_e + \int_{q_e}^{1/2} \frac{dq}{2\eta_d\sqrt{G-1}\, p(q)(1-p(q))^{G-1}q^2(1-q)^2}[1+o(1)]. \tag{41}$$

Here T_e is the time to the moving entry surface, which is at most polynomial in $1/q_0$. The post-entry integral is dominated exponentially by the deepest support region. All other factors are subexponential, hence $\log T_{\mathrm{rev}} = \rho/(2q_0) + o(1/q_0)$.

D.3 Interior initialization

The support and time coefficients hold for each fixed $p_0 \in (0,1)$ as $q_0 \to 0$, with entry-estimate constants depending on p_0. The neural initial-support ablation likewise compares fixed interior initializations.

E General Intermediate Competitor Reward

To extend the analysis to other intermediate rewards, set $r(Y) = b \in (0,1)$, keeping $r(X^+) = 1$ and $r(X^-) = 0$. Positive affine invariance of standardized advantages covers any three ordered reward levels.

Proposition 3 (Intermediate competitor rewards)**.** *For fixed finite $G \geq 2$, $0 < p, q < 1$, and $0 < b < 1$, every informative X-containing group has $g_h > 0$. Expected selection is strictly increasing in q with a unique zero $q_c^{(G)}(p,b)$. Moreover,*

$$q_c^{(G)}(p,b) = \tfrac{1}{2} + O(p), \qquad \frac{\mathsf{S}_G(p)}{\mathsf{D}_{G,b}(p)} = \tfrac{1}{2} + O(p) \quad (p \to 0), \tag{42}$$

where $\dot{q} = \eta_d \mathsf{D}_{G,b}(p)q^2[1+O(q)]$ at low competence. For fixed b, p_0, ρ, the first selection reversal has the same leading support-depth and time coefficients as Theorem 2 as $q_0 \to 0$.

E.1 Exact group gradients and pathwise development

Let the rewards be $(1, 0, b)$ with fixed $0 < b < 1$. For a group with counts (k, m, c), write $n = k+m$ and $G = n + c$. The pairwise variance identity gives

$$V_b = km + kc(1-b)^2 + mcb^2, \qquad s_r^2 = V_b/G^2, \qquad \bar{r} = (k+bc)/G. \tag{43}$$

For $V_b > 0$, centering of the advantages and the same score functions as in Appendix A imply

$$g_\ell = \frac{c(k-bn)}{G\sqrt{V_b}}, \tag{44}$$

$$g_h = \frac{k(G-k-bc)(1-q) + m(k+bc)q}{G\sqrt{V_b}}. \tag{45}$$

Zero-variance groups receive zero update. Each term in the development numerator is nonnegative, since $G - k - bc = m + (1-b)c$. For an informative group containing X, at least one is positive. Thus $g_h > 0$. Substituting $b = 1/2$ recovers Eqs. (23)–(24) exactly. Standardization is invariant to a positive affine reward transform, so $b = (r_Y - r_-)/(r_+ - r_-)$ covers all fixed $r_- < r_Y < r_+$.

E.2 Unique selection threshold

Condition on $1 \le n \le G-1$ and put $x = k - bn$. Then

$$V_b = b(1-b)nG + G(1-2b)x - x^2, \qquad -bn \le x \le (1-b)n. \tag{46}$$

The derivative of the selection factor satisfies

$$\frac{d}{dx}\frac{x}{\sqrt{V_b}} = \frac{b(1-b)nG + \frac{G}{2}(1-2b)x}{V_b^{3/2}} > 0. \tag{47}$$

The numerator is bounded below by $Gn\min(b, 1-b)/2$ on the admissible interval. Since $k \sim \mathrm{Binomial}(n, q)$, monotone coupling makes the conditional expected selection strictly increasing in q. The selection contributions for $n = 0$ and $n = G$ vanish. Averaging over n preserves strict monotonicity for $0 < p < 1$, and the endpoint expectations are $-\mathsf{S}_G(p)$ and $+\mathsf{S}_G(p)$. Hence there is a unique zero $q_c^{(G)}(p, b) \in (0, 1)$.

For $n = 1$, standardization removes the magnitude of the nonzero reward gap. A success has $g_\ell = +\sqrt{G-1}/G$, and a failure has its negative. Consequently,

$$\bar{g}_\ell(p, q; b) = \sqrt{G-1}\, p(1-p)^{G-1}(2q-1) + O(p^2), \tag{48}$$

$$\bar{g}_h(p, q; b) = 2\sqrt{G-1}\, p(1-p)^{G-1}q(1-q) + O(p^2 q(1-q)). \tag{49}$$

The remainders are uniform in $q \in [0, 1]$ for fixed G, b; the factors $q(1-q)$ follow from the finite polynomial expectation vanishing at both endpoints. Near $q = 1/2$, the implicit-function argument gives $q_c^{(G)}(p, b) = 1/2 + O(p)$. By contrast, reward equivalence is exactly $q = b$.

E.3 Low-competence coefficient and the matched barrier

At $q = 0$, mixed X^-–Y groups give selection $-\sqrt{n(G-n)}/G$, independent of b. Define

$$H_{n,b} = \frac{G - 1 - b(G-n)}{G\sqrt{(n-1) + (G-n)(1-b)^2 + (n-1)(G-n)b^2}}, \quad 1 \le n \le G. \tag{50}$$

This is g_h at $q = 0$ for the composition $(1, n-1, G-n)$. The all-failure compositions contribute $\mathsf{S}_G(p)q$ to $\bar{g}_h$ to first order; the one-success compositions contribute the remaining term. Therefore

$$\mathsf{D}_{G,b}(p) = \mathsf{S}_G(p) + \sum_{n=1}^{G} \binom{G}{n} p^n (1-p)^{G-n} n H_{n,b} > 0, \tag{51}$$

and $\bar{g}_\ell = -\mathsf{S}_G(p)[1 + O(q)]$, $\dot{q} = \eta_d \mathsf{D}_{G,b}(p) q^2 [1 + O(q)]$. Since $H_{1,b} = \sqrt{G-1}/G$,

$$\mathsf{S}_G(p) = \sqrt{G-1}\, p + O(p^2), \qquad \mathsf{D}_{G,b}(p) = 2\sqrt{G-1}\, p + O(p^2). \tag{52}$$

The ratio $R_{G,b} = \mathsf{S}_G/\mathsf{D}_{G,b}$ extends continuously to $p = 0$ with value $1/2$, and is bounded above and below by positive constants on $[0, p_0]$ for every fixed $p_0 < 1$. Low-q errors are uniform there: the finite composition polynomials have a common factor p, and $\mathsf{D}_{G,b}/p$ remains positive.

Choose a small fixed $\bar{q}$ below the minimum of $q_c^{(G)}(p, b)$ on $[0, p_0]$. The moving-threshold proof of Appendix D then applies: entry at $p = q_0$ occurs with $q_e - q_0 = O(q_0^2 \log(1/q_0))$, and support decreases until at least $\bar{q}$. Integrating the phase cost gives

$$\log p(\bar{q}) = -\rho/(2q_0) + O(\log(1/q_0)). \tag{53}$$

On any fixed compact competence interval containing $[\bar{q}, 1/2]$ and lying inside $(0, 1)$, Eqs. (48)–(49) give

$$\frac{d\ell}{dq} = \rho \frac{q - 1/2}{q^2(1-q)^2} + O(p). \tag{54}$$

The remaining phase integral is bounded. The first selection reversal occurs at $q = 1/2 + O(p)$, with exponentially small p, and is a local support minimum. It follows that

$$\log p_{\min} = -\rho/(2q_0) + O(\log(1/q_0)), \qquad \log T_{\mathrm{rev}} = \rho/(2q_0) + o(1/q_0). \tag{55}$$

For the time claim, $\dot{q} = \Theta(pq^2)$ up to a fixed competence below one. The early entry time and competence factors are at most polynomial, and any fixed competence interval immediately before reversal has the displayed exponential support scale. Upper and lower bounds on its time integral give the coefficient. The constants depend on the fixed parameters G, b, p_0.

F Conditional Success, Score Geometry, and Access

F.1 A fixed event and its score identities

Let X and $\mathcal{G}_X \subseteq X$ be fixed measurable sets. Assume $0 < p_\theta < 1, 0 < q_\theta < 1$, and a differentiable policy density with respect to a fixed dominating measure. The following identities require that differentiation can pass under the relevant integrals and that the score expectations exist; locally common support and an integrable bound on the density derivatives suffice.

Write $f_\theta(y) = \pi_\theta(y \mid X)$ and $u_\theta(y) = \nabla \log f_\theta(y)$. Normalization and differentiation of the fixed-event integral give

$$\mathbb{E}[u_\theta(Z) \mid X] = 0,$$

$$\nabla q = \nabla \int_{\mathcal{G}_X} f_\theta(y)\, dy = \mathbb{E}[\mathbf{1}_{\mathcal{G}_X}(Z) u_\theta(Z) \mid X] = q\mu_+, \tag{56}$$

$$\mu_+ = \mathbb{E}[u_\theta(Z) \mid Z \in \mathcal{G}_X], \qquad \mathbb{E}[-u_\theta(Z) \mid Z \in X \setminus \mathcal{G}_X] = \frac{\nabla q}{1-q}. \tag{57}$$

Equivalently, if $w_\theta(y) = \nabla \log \pi_\theta(y)$ is the natural-policy score,

$$\nabla q = \frac{1}{p}\mathbb{E}\big[\big(\mathbf{1}_{\mathcal{G}_X}(Z) - q\mathbf{1}_X(Z)\big) w_\theta(Z)\big]. \tag{58}$$

This follows by differentiating $q = s/p$. The factor $1/p$ normalizes the natural-policy expectation to the conditional distribution.

Equation (57) explains why suppressing conditional failures can improve conditional success even without a successful sample in a particular group. It is an identity about the *expected conditional-score direction*. In a shared policy the full score for an X sample is $\nabla \log p + u_\theta(Z)$, and other regions also update shared parameters.

F.2 Sufficient conditions for quadratic probability-space development

Proposition 4 (Conditional success-gradient geometry)**.** *Suppose the score identities above hold along a low-success family, with constants $0 < c \le \|\mu_+\| \le C < \infty$. For a conditional-learning clock τ, suppose*

$$\frac{d\theta}{d\tau} = M_\theta \nabla q, \qquad mI \preceq M_\theta \preceq MI, \quad 0 < m \le M < \infty, \tag{59}$$

where M_θ is symmetric. Then

$$\|\nabla q\| = \Theta(q), \qquad mc^2 q^2 \le \frac{dq}{d\tau} \le MC^2 q^2. \tag{60}$$

Proof. Equation (56) gives $\|\nabla q\| = q\|\mu_+\|$. Taking the inner product of Eq. (59) with ∇q gives $dq/d\tau = (\nabla q)^\top M_\theta \nabla q$, to which the spectral and score-norm bounds apply. □

The two-sided score bound and positive quadratic form in Proposition 4 jointly determine the quadratic development order. They account for the magnitude of the mean success score and its alignment with the conditional update.

For a finite-action softmax, one concrete sufficient upper bound is: if the conditional logits satisfy $\|\nabla z_y\| \le K$ on the parameter range of interest, then $u_\theta(y) = \nabla z_y - \mathbb{E}[\nabla z_Z \mid X]$ has norm at most $2K$, and therefore $\|\nabla q\| \le 2Kq$. Together with the mean-score lower bound and update alignment in Proposition 4, this yields quadratic development.

The canonical rare-support conditional dynamics satisfy the sufficient geometry: after removing the access clock, $\dot{h}$ is a constant multiple of $q(1-q)$ to leading order. This explains its development order.

F.3 A SUFFICIENT BRIDGE BEYOND FACTORIZED LOGITS

Proposition 5 (Quadratic phase order from projected geometry)**.** *Suppose* $\|\nabla q\| = \Theta(q)$ *and the complete update can be written*

$$v = a(p, z)\{M_\theta \nabla q + w\}, \qquad a(p, z) > 0, \tag{61}$$

with symmetric $mI \preceq M_\theta \preceq MI$*. Suppose, uniformly on a segment with* $\dot{L} < 0$ *and* $\dot{q} > 0$*,*

$$0 < s_- \le -\nabla L^\top (M_\theta \nabla q + w) \le s_+, \qquad |\nabla q^\top w| \le K q^2, \tag{62}$$

and the total projection onto ∇q *is bounded below by* $d_- q^2 > 0$*. Then* $S = \Theta(a)$*,* $D = \Theta(aq^2)$*, and* $C = \Theta(q^{-2})$*.*

Proof. The quadratic form $(\nabla q)^\top M_\theta \nabla q$ is bounded above and below by positive multiples of q^2. The cross-term bound gives the upper bound on D/a; the stipulated non-cancellation gives its positive lower bound. Equation (62) bounds S/a. Taking their ratio proves the result. □

The proposition applies directly to the *projected* complete update in a shared-parameter policy. Its two-sided bounds establish quadratic phase order. If $(S/a)/(D/(aq^2)) \to \kappa$, then $C \sim \kappa q^{-2}$; a local log-slope limit additionally needs derivative regularity. If $a(p, z) \asymp p$ and the residual positive development factor has uniform polynomial bounds, Proposition 7 gives $\log T = \Theta(1/q_0)$. More generally, $a \asymp p^\gamma$ gives the corresponding positive-access order. This supplies a bridge from success-gradient geometry to the general phase argument through the projected complete update.

F.4 COORDINATE CHANGES PRESERVE THE PHASE INTEGRAL

For a differentiable increasing $\widetilde{q} = f(q)$ with $f'(q) > 0$,

$$\widetilde{C}(\widetilde{q}) = \frac{C(q)}{f'(q)}, \qquad \widetilde{C}(\widetilde{q})\, d\widetilde{q} = C(q)\, dq. \tag{63}$$

Thus the accumulated support-logit loss is unchanged, while the exponent attached to the coordinate can change. In particular, if $C(q) \sim \kappa q^{-\delta}$ and $f(q) = q^r$ with $r > 0$,

$$\widetilde{C}(\widetilde{q}) \sim \frac{\kappa}{r} \widetilde{q}^{-\widetilde{\delta}}, \qquad \widetilde{\delta} = 1 + \frac{\delta - 1}{r}. \tag{64}$$

Positive power changes preserve whether δ is above, equal to, or below one, but change its numerical value and the representation of initial difficulty. If instead $f(q) = aq[1+o(1)]$ and $f'(q) = a[1+o(1)]$ with $a > 0$, the asymptotic power remains δ, with coefficient $\kappa a^{\delta-1}$. These statements concern power-law orders; convergence of derivative-based effective exponents requires corresponding derivative regularity. The fixed event-probability coordinate gives the phase exponent a consistent interpretation in terms of conditional success.

F.5 SUCCESSFUL SUPPORT AND ITS TURNING POINT

For $s = pq$ and $L = \operatorname{logit} p$, differentiation yields

$$\frac{d \log s}{dq} = \frac{1}{q} + \frac{1}{p} \frac{dp}{dL} \frac{dL}{dq} = \frac{1}{q} - (1 - p) C(q). \tag{65}$$

Hence a positive competence drift can coexist with a decreasing probability of naturally generating successful trajectories. If $C(q) \sim \kappa q^{-\delta}$ and support is rare, the selection term dominates for $\delta > 1$ at sufficiently low competence. At $\delta = 1$ the leading competition is between κ and one, while equality requires higher-order analysis. For $\delta < 1$ the direct $1/q$ term dominates in this rare-support limit.

At a fixed endpoint $q_I > 0$, $\log s(q_I) = \log p(q_I) + \log q_I$ preserves the leading exponential coefficient. A stationary point of s instead solves $(1 - p)qC(q) = 1$ on a segment with $\dot{L} < 0$ and $\dot{q} > 0$; it need not coincide with a stationary point of p. In particular, at the canonical support minimum, $\dot{p} = 0$ and $\dot{q} > 0$, so $\dot{s} = p\dot{q} > 0$. The theorem's selection-reversal endpoint is therefore not a successful-support minimum.

F.6 Context weighting and neural evaluation

For the neural experiment, define fixed-context competence $q = \mathbb{E}_\nu[q_{\mathbf{o}}]$ and average support $\bar{p} = \mathbb{E}_\nu[p_{\mathbf{o}}]$. These averages use a context law ν fixed throughout training. With fixed $\nu(\mathbf{o})$, let $q_{\mathbf{o}} = \mathbb{P}(\mathcal{G}_X(\mathbf{o}) \mid \mathbf{o}, X(\mathbf{o}))$ and $p_{\mathbf{o}} = \mathbb{P}(X(\mathbf{o}) \mid \mathbf{o})$. The joint natural law $\nu(\mathbf{o})\pi_\theta(a \mid \mathbf{o})$ gives

$$q^{\mathrm{visit}} = \frac{\mathbb{E}_\nu[p_{\mathbf{o}} q_{\mathbf{o}}]}{\bar{p}}, \qquad q^{\mathrm{visit}} - q = \frac{\mathrm{Cov}_\nu(p_{\mathbf{o}}, q_{\mathbf{o}})}{\bar{p}}, \qquad s = \bar{p}q + \mathrm{Cov}_\nu(p_{\mathbf{o}}, q_{\mathbf{o}}). \tag{66}$$

Natural access changes the weighting in q^{visit}, which can rise with no improvement at any fixed context.

The fixed-context protocol is itself a well-defined probability law,

$$\widetilde{\pi}_\theta(\mathbf{o}, a) = \nu(\mathbf{o})\pi_\theta(a \mid \mathbf{o}, X(\mathbf{o})), \qquad q = \mathbb{P}_{\widetilde{\pi}_\theta}(a \in \mathcal{G}_X(\mathbf{o})). \tag{67}$$

For $u_{\mathbf{o}}(a) = \nabla \log \pi_\theta(a \mid \mathbf{o}, X(\mathbf{o}))$, differentiation gives

$$\nabla q = \mathbb{E}_\nu[q_{\mathbf{o}} \mu_{+,\mathbf{o}}] = q\mathbb{E}_{\mathbf{o}\sim\nu_+}[\mu_{+,\mathbf{o}}], \qquad \nu_+(\mathbf{o}) = \frac{\nu(\mathbf{o})q_{\mathbf{o}}}{q}. \tag{68}$$

The score-norm bounds apply to the aggregate vector in Eq. (68). Under the corresponding phase hypotheses, integration for $(\mathrm{logit}\,\bar{p}, q)$ describes aggregate access versus fixed-context learning. Successful natural support is given separately by Eq. (66), including the covariance between per-context access and competence.

F.7 Direct access probability versus total development

Let $g_{\mathcal{B}}$ be a sampled update direction at fixed parameters and E_X the event that its G IID trajectories contain an X sample. For the mean-field competence drift, define

$$d_1 = \mathbb{E}[\nabla q^\top g_{\mathcal{B}} \mid E_X], \qquad d_0 = \mathbb{E}[\nabla q^\top g_{\mathcal{B}} \mid E_X^c]. \tag{69}$$

The law of total expectation gives the exact decomposition

$$D = \big[1 - (1-p)^G\big] d_1 + (1-p)^G d_0. \tag{70}$$

At matched competence and other state, bounded d_1 makes the direct-access contribution $O(p)$. If $d_0 = 0$ and d_1 tends to a strictly positive finite limit, then $D = \Theta(p)$, giving the canonical access order. In contrast, a positive non-vanishing d_0 can sustain development as $p \to 0$; negative transfer or cancellation can make net development slower or nonpositive. Importance weights or score norms that grow with inverse support also require separate analysis. For shared-context groups the IID formula applies conditionally on the context; one averages $1 - (1 - p_{\mathbf{o}})^G$ across contexts using the per-context access probabilities.

The decomposition separates direct learning from transfer through shared parameters. To evaluate the access elasticity, vary support along a specified differentiable perturbation family at fixed competence and auxiliary state.

G General Phase Classification

G.1 Finite-order drifts and the phase exponent

The power-law phase cost follows from finite leading orders of the actual selection and development drifts. Let u be a common positive clock along a trajectory, with $du/dt > 0$, and let ξ collect the remaining state, including L and any optimizer or reference state. This clock may be chosen to factor out a common access dependence.

Proposition 6 (Finite-order drift reduction)**.** *Suppose, uniformly over the relevant low-competence state set $\mathcal{K}$,*

$$-\frac{dL}{du} = a(\xi)q^{\alpha}[1 + r_L(\xi, q)], \qquad \frac{dq}{du} = b(\xi)q^{\beta_{\mathrm{ph}}}[1 + r_q(\xi, q)], \tag{71}$$

where both remainders converge uniformly to zero as $q \to 0$, and $0 < a_- \le a(\xi) \le a_+ < \infty$ and $0 < b_- \le b(\xi) \le b_+ < \infty$. Then, uniformly on $\mathcal{K}$,

$$C(\xi, q) = \frac{a(\xi)}{b(\xi)}q^{-\delta}[1 + o(1)], \qquad \delta = \beta_{\mathrm{ph}} - \alpha. \tag{72}$$

If $a(\xi(q))/b(\xi(q)) \to \kappa_\star \in (0, \infty)$ along the trajectory family, then $C(q) \sim \kappa_\star q^{-\delta}$.

Proof. The common positive clock cancels from the ratio, so

$$C = -\frac{dL/du}{dq/du} = \frac{a(\xi)}{b(\xi)}q^{\alpha-\beta_{\mathrm{ph}}}\frac{1 + r_L(\xi, q)}{1 + r_q(\xi, q)}. \tag{73}$$

Uniform convergence makes the last ratio $1 + o(1)$, proving both statements. □

Two-sided drift bounds suffice for the corresponding phase order: if $c_L^- q^\alpha \le -dL/du \le c_L^+ q^\alpha$ and $c_q^- q^{\beta_{\mathrm{ph}}} \le dq/du \le c_q^+ q^{\beta_{\mathrm{ph}}}$, with positive finite constants, then

$$\frac{c_L^-}{c_q^+}q^{-\delta} \le C(q) \le \frac{c_L^+}{c_q^-}q^{-\delta}. \tag{74}$$

These bounds connect directly to Proposition 7. Thus a phase-order gap $\beta_{\mathrm{ph}} - \alpha > 1$, together with positive access dependence and the stated uniform bounds, gives exponential access complexity. Finite order is a sufficient regularity assumption on the complete drifts. In a finite discrete model, analytic composition-wise update terms and polynomial composition probabilities give analytic expected drifts; a nonzero analytic drift has a finite leading order. Uniform bounds over the state set are supplied separately by the model's expansions.

For the canonical system, the low-competence expansions give

$$-\dot{L} = \eta_s \mathsf{S}_G(p)[1 + O(q)], \qquad \dot{q} = \eta_d \mathsf{D}_G(p)q^2[1 + O(q)]. \tag{75}$$

At fixed interior support these have orders $\alpha = 0$ and $\beta_{\mathrm{ph}} = 2$. To cover the rare-support limit with positive bounded coefficients, use the common exposure clock $du/dt = p(1-p)^{G-1}$. The expansions in Appendix C then give selection and development coefficients tending to $\eta_s\sqrt{G-1}$ and $2\eta_d\sqrt{G-1}$, respectively. Consequently,

$$\delta = 2, \qquad C(p, q) = \frac{\rho}{2}q^{-2}[1 + O(p) + O(q)], \tag{76}$$

recovering the canonical phase law from the actual drift orders.

G.2 Clock invariance and support classes

If $d\tau/dt = m_t > 0$, then

$$-\frac{dL/d\tau}{dq/d\tau} = -\frac{(dL/dt)/m_t}{(dq/dt)/m_t} = -\frac{dL}{dq}. \tag{77}$$

Therefore $C = -dL/dq$ is a phase property. If $C(q) = \kappa q^{-\delta}[1 + o(1)]$, direct integration gives

$$\int_{q_0}^{q_I} C(q)dq = \begin{cases} \dfrac{\kappa}{\delta - 1}q_0^{-(\delta-1)}[1 + o(1)], & \delta > 1, \\ \kappa\log(1/q_0) + o(\log(1/q_0)), & \delta = 1, \\ O(1), & \delta < 1. \end{cases} \tag{78}$$

Once support is rare, $L \sim \log p$, yielding Eq. (14). If $C(q) \ge cq^{-\delta}$ with $\delta > 1$ until a fixed support threshold is crossed, the fixed logit cost requires $q_e - q_0 = O(q_0^\delta) = o(q_0)$; a matching upper bound on C yields $\Theta(q_0^\delta)$.

G.3 Support-to-development coupling

In a shared network, updates sampled outside X can change q. Let $D(p, q, z) > 0$ denote a controlled useful-development speed, with other state summarized by z. The local elasticity

$$\gamma_{\text{eff}} = \left.\frac{\partial \log D}{\partial \log p}\right|_{q,z} \tag{79}$$

measures dependence on natural support and can be estimated by routing or sampling perturbations. For Proposition 2, assume $\dot{q} = p^{\gamma} d(q, z)$ with $\gamma > 0$ and uniform polynomial bounds on d and $1/d$. Put $F(q_0) = \kappa q_0^{-(\delta-1)}/(\delta-1)$ for $\delta > 1$. Monotone support loss gives $-\log p(q) \le F(q_0)[1+o(1)]$ throughout the segment; any polynomial competence prefactors contribute only $O(\log(1/q_0))$ to log time. This gives

$$\log \int_{q_0}^{q_I} \frac{dq}{p(q)^{\gamma} d(q, z)} \le \gamma F(q_0)[1 + o(1)]. \tag{80}$$

On any fixed interval $[a, b] \subset (0, q_I)$, the phase integral has already accumulated $F(q_0)[1 + o(1)]$. Its finite width and the polynomial bounds on d yield the matching lower bound, proving Eq. (15). For $\delta = 1$, inverse support is bounded by a power of $1/q_0$; for $\delta < 1$, support remains bounded below. The same integral is therefore at most polynomial. For $\gamma = 0$, $T = \int_{q_0}^{q_I} d(q, z)^{-1} dq$ is at most polynomial under the same uniform bounds, regardless of the depth of the support valley. The uniform bounds across the trajectory family give the stated asymptotic order.

G.4 Uniform power-law bounds

The sharp coefficient in Proposition 2 uses an asymptotic equivalent. The existence of an exponential-access barrier needs less.

Proposition 7 (Robust exponential-access order)**.** *Consider trajectories with decreasing support and increasing competence from $q_0 \to 0$ to fixed q_I, with initial support in a fixed compact subset of $(0, 1)$. Suppose that, uniformly for $q_0 \le q \le \bar{q} < q_I$,*

$$k_- q^{-\delta} \le C(q) \le k_+ q^{-\delta}, \qquad 0 < k_- \le k_+ < \infty, \quad \delta > 1, \tag{81}$$

with nonnegative, uniformly bounded integrated phase cost on $[\bar{q}, q_I]$. Suppose throughout the segment, for constants independent of q_0,

$$c q^m p^{\gamma_+} \le D(p, q, z) \le C_d q^{-m} p^{\gamma_-}, \quad m \ge 0, \quad 0 < \gamma_- \le \gamma_+ < \infty. \tag{82}$$

Then $-\log p(q_I) = \Theta(q_0^{-(\delta-1)})$ and

$$\log T(q_0; q_I) = \Theta(q_0^{-(\delta-1)}). \tag{83}$$

Proof. Integrating Eq. (81) bounds the total logit loss above and below by positive multiples of $q_0^{-(\delta-1)}$. The same lower bound has accumulated on any fixed interval $[a, b] \subset (0, \bar{q})$ for sufficiently small q_0. Monotonicity and the lower bound on D bound T above by a polynomial in $1/q_0$ times $\exp[O(q_0^{-(\delta-1)})]$. On $[a, b]$, the upper bound on D bounds T below by $\exp[\Omega(q_0^{-(\delta-1)})]$. Taking logarithms proves the claim. □

A controlled elasticity bound $0 < \gamma_- \le \partial \log D / \partial \log p \le \gamma_+$ implies the support powers in Eq. (82) if it holds over the full support range from a fixed reference level to the valley, at matched (q, z), and D at that reference level has uniform polynomial bounds. This follows by integrating the elasticity in $\log p$.

H Complete Update Fields

Section 4 applies the phase analysis to the learner's complete update field. The following decomposition specifies how clipping, reference regularization, and repeated optimization enter that field. For a batch sampled from π_{old}, treat its advantages as fixed and write

$$w_i(\theta) = \frac{\pi_\theta(z_i)}{\pi_{\text{old}}(z_i)}, \qquad \mathcal{J}_{\text{clip}}(\theta) = \frac{1}{G} \sum_i \min\Big\{ w_i \widehat{A}_i, \text{clip}(w_i, 1-\epsilon, 1+\epsilon) \widehat{A}_i \Big\}. \tag{84}$$

This is the one-step policy-gradient surrogate corresponding to the clipped objectives of Schulman et al. (2017) and Shao et al. (2024). At differentiable points, its gradient is

$$h_{\text{clip}} = \frac{1}{G}\sum_i \chi_i w_i \widehat{A}_i \nabla_\theta \log \pi_\theta(z_i), \qquad \chi_i = \begin{cases} \mathbf{1}\{w_i < 1+\epsilon\}, & \widehat{A}_i > 0, \\ \mathbf{1}\{w_i > 1-\epsilon\}, & \widehat{A}_i < 0. \end{cases} \tag{85}$$

Zero-advantage terms contribute zero; at clipping boundaries use the implementation's subgradient. At the first evaluation of a fresh on-policy batch, $w_i = 1$ and every nonzero-advantage term is active. The gradient is exactly the GRPO score update in Eq. (4).

Let $\mathcal{R}(\theta, z)$ be the chosen reference penalty, including its sampling convention, and λ its weight. Its gradient contributes to the update before the optimizer is applied:

$$h = h_{\text{clip}} - \lambda \nabla_\theta \mathcal{R}(\theta, z), \qquad v(\theta, z) = \mathbb{E}[\mathcal{U}(\theta, z; h) \mid \theta, z]. \tag{86}$$

Here $\mathcal{U}$ is the parameter update produced by the optimizer, and z includes its moments, the reference and old policies, retained samples, and schedule. The expectation averages the randomness remaining conditional on this state. Repeated optimization keeps the old-policy denominator and retained-batch state while θ changes, thereby changing both w_i and the active clipping terms. The auxiliary state evolves with the same training protocol.

The resulting selection and development drifts are $\nabla L^\top v$ and $\nabla q^\top v$. Their ratio determines C, and the support dependence of the latter determines the access exponent. This supplies an algorithm-independent route into the classification: each complete learner contributes its own drift orders and coefficients, which can be evaluated against Propositions 2, 6, and 7. The canonical calculation uses the fresh-batch score field to obtain explicit values; the complete-field formulation carries the analysis to other update protocols through their resulting dynamics.

I Canonical Simulation Details

I.1 Simulator and protocol

At state (p, q), a group is active with probability

$$P_{\text{active}} = 1 - (pq)^G - [p(1-q)]^G - (1-p)^G. \tag{87}$$

An event-driven simulator samples the geometric waiting time to the next active group and then samples its exact conditional multinomial composition. This is distributionally identical to naive group-by-group simulation, but skips zero-update groups. We use $\eta_s = \eta_d = 0.2$, 200 independent replicates per condition, NumPy PCG64DXSM, and 5,000-resample bootstrap intervals. Across group-size, initial-support, and stress suites, the design contains 48 conditions and 9,600 trajectories; every run reaches maturity. The grids and computational clocks are specified below.

I.2 Suites and clocks

Table 3 lists the main grids. Additional tests fix $G = 8$. They use $p_0 = 0.8$ with $q_0 \in \{0.020, 0.025, 0.030, 0.035, 0.040, 0.050\}$, and $p_0 = 0.2$ with $q_0 \in \{0.065, 0.080, 0.100, 0.150\}$.

Table 3: Canonical stochastic simulation grids used for primary fits.

suite	fixed parameters	main q_0 grid
baseline	$G = 8, p_0 = 0.2$	.040, .045, .050, .055, .060
group size	$G = 8, p_0 = .2$	.040, .045, .050, .055, .060
	$G = 16, p_0 = .2$	.035, .040, .045, .050, .055
	$G = 32, p_0 = .2$	.030, .035, .040, .045, .050
initial support	$G = 8, p_0 = .1$	.045, .050, .055, .060, .065, .070
	$G = 8, p_0 = .2$	.040, .045, .050, .055, .060
	$G = 8, p_0 = .5$	.030, .035, .040, .045, .050

The simulator records group attempts, rollout count $N_{\text{rollout}} = GN_{\text{group}}$, and active updates. Cross-G results use rollout count. Positive times and minima are analyzed in log space. Confidence intervals

use 5,000 nonparametric bootstrap draws over 200 independent replicates. Each replicate seed is derived from the suite name, condition identifier, and replicate index with SHA-256, using the first 128 bits to initialize NumPy PCG64DXSM. The formal run used 40 CPU workers and no GPU.

For the baseline $G = 8, p_0 = 0.2$, matched non-rare theory and stochastic median-scale group attempts are:

q_0	matched theory	stochastic median scale
0.050	8.4×10^4	92,531
0.055	3.4×10^4	38,848
0.060	1.59×10^4	15,836
0.065	8.38×10^3	9,529
0.080	1.98×10^3	2,204

I.3 Additional canonical results

Figure 5 shows the group-size and initial-support conditions beyond the compact main-text example. The leading slopes remain close to $\rho/2$ for log time and $-\rho/2$ for log depth. Dashed curves show theoretical values calculated from the finite-group dynamics. All displayed conditions reach maturity. Support-trajectory summaries use checkpoints with observations from at least half the seeds, up to each run's maturity time.

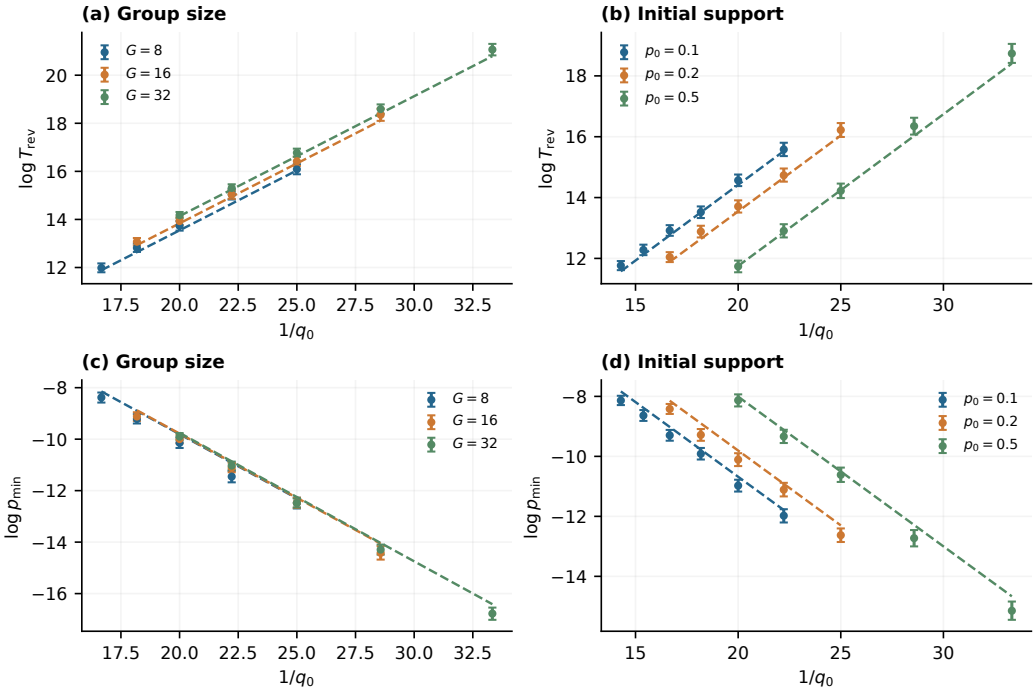


Figure 5: **Additional canonical results.** Top: reversal time. Bottom: support minimum. Left: group-size variation at $p_0 = 0.2$. Right: initial-support variation at $G = 8$. Markers summarize simulations and error bars are 95% bootstrap intervals. Dashed curves are the calculated theoretical values. Time counts rollouts.

J Shared-Neural Experimental Details

J.1 Data, initialization, and optimization

The official 60,000-example MNIST training set is split with seed 1729 into 50,000 RL-training and 10,000 calibration examples. The official 10,000-example test set is fixed for all trajectory

evaluation. The MLP uses two 256-unit ReLU layers and one flat 11-way softmax. A warm-start run with seed 31415 trains the ten digit logits while holding the fallback row fixed. Calibration checkpoints approach each target q_0; binary search then adds a global scalar fallback-logit offset to obtain $p_0 \approx 0.8$ while preserving the conditional distribution inside X.

The policy maps 784 image features through the two hidden layers to a single softmax over ten digit actions and the fallback. Each update samples 256 contexts and eight current-policy actions per context, applies the GRPO score update in Eq. (4) in one SGD step, and discards the samples. The main initial-competence grid is $\{0.10, 0.12, \ldots, 0.30\}$. Table 4 labels conditions by their prescribed initial competence q_0; plotting coordinates, regressions, and competence changes use each run's measured initial competence. Plot markers are means of the log quantities, with individual seed observations also shown.

Each condition uses RL seeds $\{0, \ldots, 9\}$, context batch size 256, group size 8, SGD with learning rate 0.01, momentum 0, and weight decay 0. Evaluation intervals are 5 updates for targets 0.24–0.30 and 50 otherwise. The reported horizon is 20,000 updates for targets 0.10–0.16 and targets at least 0.20; target 0.18 uses a 100,000-update run. The 0.10–0.16 conditions are horizon-censored and the $q_0 = 0.18$–0.30 conditions are measurement-resolved. A horizon-censored run has its observed minimum in the last 5% of a capped trajectory, before a completed valley and recovery can be established. For 0.10–0.16, the recorded minima occur between 19,850 and 20,000 updates. Figure 4 explains the interpretation of these observations.

J.2 Summary statistics

Table 4 reports arithmetic means and sample standard deviations across ten RL seeds, using $n - 1$ in the variance estimate. For each seed, Δq is the change from measured initial competence to competence at the recorded support minimum. Figure 4 averages support and time in log space. For HC conditions, both summaries use the observed 20,000-update window.

Table 4: Shared-neural results at initial support $p_0 = 0.8$: arithmetic mean $\pm$ sample standard deviation over ten seeds. Initial competence q_0 is in percent; Δq at the recorded support minimum is in percentage points (pp). Times count updates; $q_0 = 0.10$–0.16 are horizon-censored (HC) at 20,000.

q_0 (%)	$p_{\min}$	$T_{p_{\min}}$	Δq (pp)
10	$(2.20 \pm 0.04) \times 10^{-5}$	$19,990 \pm 21$	-0.20 ± 0.02
12	$(2.26 \pm 0.05) \times 10^{-5}$	$19,980 \pm 35$	$+0.27 \pm 0.02$
14	$(2.48 \pm 0.07) \times 10^{-5}$	$19,980 \pm 35$	$+0.91 \pm 0.03$
16	$(3.02 \pm 0.11) \times 10^{-5}$	$19,950 \pm 58$	$+1.38 \pm 0.07$
18	$(9.81 \pm 3.41) \times 10^{-5}$	$16,435 \pm 6,191$	-0.48 ± 0.48
20	$(1.37 \pm 0.23) \times 10^{-3}$	$1,515 \pm 284$	$+1.87 \pm 0.29$
22	$(6.77 \pm 1.11) \times 10^{-3}$	400 ± 47	$+3.54 \pm 0.20$
24	$(2.07 \pm 0.20) \times 10^{-2}$	208 ± 13	$+4.21 \pm 0.05$
26	$(6.23 \pm 0.51) \times 10^{-2}$	113 ± 7	$+3.88 \pm 0.13$
28	$(1.07 \pm 0.04) \times 10^{-1}$	94 ± 6	$+3.86 \pm 0.16$
30	$(1.43 \pm 0.03) \times 10^{-1}$	94 ± 10	$+3.84 \pm 0.23$

J.3 Scaling fits and uncertainty

For each initialization condition, let x be the mean of the ten realized $1/q_0$ values and let y be the mean log minimum support or mean log time. Ordinary least squares assigns equal weight to each condition. The displayed lines in Figure 4 and all numerical regression results below use the same seven uncensored conditions, $q_0 = 0.18$–0.30. Lines and bands span the mean inverse-competence coordinates of these conditions:

$$\widehat{\log p_{\min}} = 9.485 - 3.311x, \qquad R^2 = 0.976, \tag{88}$$

$$\widehat{\log T_{p_{\min}}} = -3.617 + 2.274x, \qquad R^2 = 0.917. \tag{89}$$

The 95% bootstrap slope intervals are $[-3.408, -3.221]$ for log depth and $[2.181, 2.375]$ for log time; the corresponding intercept intervals are $[9.159, 9.838]$ and $[-3.998, -3.267]$. These intervals use 5,000 resamples of the ten RL seed identities with replacement, retaining each selected seed across the initialization conditions, recomputing condition means, and refitting each line. The same procedure gives the pointwise confidence bands for the displayed seven-condition lines. The bands quantify seed variability in the fitted mean lines at the specified conditions. R^2 measures goodness of fit across condition means.

J.4 CONTROLS

Forced access starts from the target-0.20 answer parameters and removes the fallback from training-time sampling. Rewards are one for the correct digit and zero otherwise. All ten runs reach $q \geq 0.8$. With 2,048 actions sampled per update, mean threshold times of 300, 500, and 670 updates for $q = 0.4, 0.5, 0.6$ correspond to 614,400; 1,024,000; and 1,372,160 sampled actions.

The random-target control assigns each training example a fixed class independent of image content and uses fixed random held-out targets. It retains the natural answer–fallback policy and reward $r(Y) = 0.5$. Ten runs last 20,000 updates. Initial held-out conditional competence is 0.099872. The observed support minimum is 1.964×10^{-5} with 95% interval $[1.931, 1.999] \times 10^{-5}$, while competence at that minimum is 0.100012; mean change is 1.40×10^{-4} with interval $[1.15, 1.66] \times 10^{-4}$.

J.5 ABLATION DETAILS

The ablation study contains 20 fallback-reward runs ($b = 0.3, 0.7$), 20 group-size runs ($G = 4, 16$), and 30 initial-support runs ($p_0 = 0.2, 0.5, 0.8$). The ten $p_0 = 0.8$ runs are the shared baseline already reported in the main results. The same baseline is included in the reward and group-size comparisons. All use target $q_0 = 0.20$; unspecified settings are $p_0 = 0.8$, $b = 0.5$, and $G = 8$. Their horizons are 20,000 updates, with the same 50-update evaluation cadence and ten RL seeds as the baseline protocol.

Seed-bootstrap 95% intervals quantify variation over the ten RL seeds at the fixed warm start. For initial support 0.2, observed minima occur at 19,400–20,000 updates, so all ten runs meet the horizon-censoring rule. Their observed support minima and contraction ratios are upper bounds on the corresponding full-trajectory minima. Other ablation minima occur away from the horizon.

Table 2 in the main text reports support minima and their times as arithmetic means $\pm$ sample standard deviations across the ten seeds. The remaining table and Figure 6 retain the geometric-mean and bootstrap summaries for positive support quantities; competence is averaged arithmetically.

Table 5: Conditional competence and relative support contraction with seed-bootstrap 95% intervals.

condition	$q(T_{p_{\min}})$ [95% CI]	$p_{\min}/p_0$ [95% CI]
baseline	0.2207 $[0.2190, 0.2224]$	1.69×10^{-3} $[1.50\times10^{-3}, 1.86\times10^{-3}]$
$b = 0.3$	0.2445 $[0.2431, 0.2458]$	7.59×10^{-3} $[6.62\times10^{-3}, 8.62\times10^{-3}]$
$b = 0.7$	0.2064 $[0.2035, 0.2093]$	4.06×10^{-4} $[3.46\times10^{-4}, 4.75\times10^{-4}]$
$G = 4$	0.2196 $[0.2176, 0.2217]$	1.67×10^{-3} $[1.45\times10^{-3}, 1.88\times10^{-3}]$
$G = 16$	0.2260 $[0.2240, 0.2279]$	1.62×10^{-3} $[1.38\times10^{-3}, 1.89\times10^{-3}]$
$p_0 = 0.2$ (HC)	0.2190 $[0.2180, 0.2201]$	2.89×10^{-4} $[2.74\times10^{-4}, 3.02\times10^{-4}]$
$p_0 = 0.5$	0.2076 $[0.2058, 0.2094]$	3.27×10^{-4} $[2.85\times10^{-4}, 3.71\times10^{-4}]$

The maximum-competence statistic is $\max_{0 \le t \le T_{p_{\min}}} q(t)$ over evaluation checkpoints. Baseline maximum competence is 0.2348 [0.2344, 0.2353], with mean gain 0.0328 [0.0323, 0.0332]. For $b = 0.7$, it is 0.2269 [0.2264, 0.2273], with gain 0.0248 [0.0244, 0.0252]. These positive earlier gains coexist with later partial loss of competence.

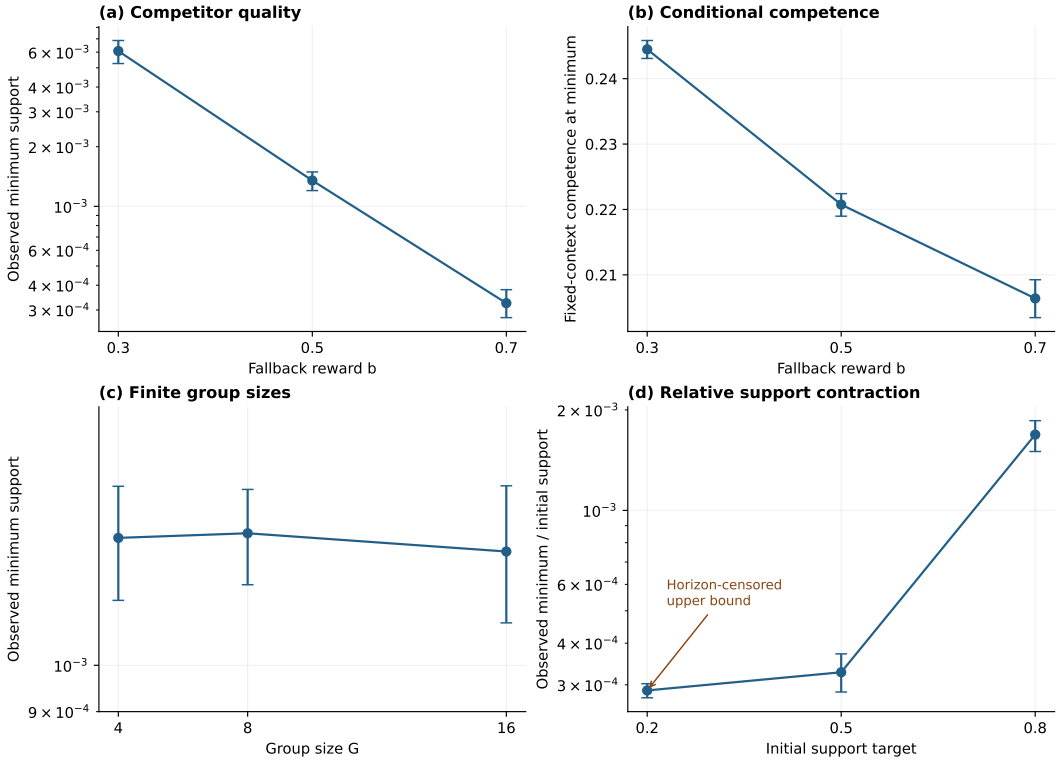


Figure 6: **Neural ablations.** The common baseline is included in every sweep. Points show geometric means for positive support quantities and arithmetic means for competence, with seed-bootstrap 95% intervals. The $p_0 = 0.2$ contraction is marked as horizon-censored. Connecting lines serve as visual guides.